Title

Uncovering Computations of Human Decision Making: Neurocognitive Modeling and Experimentation

Author

Sun, Qinhua

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,
IRVINE

Uncovering Computations of Human Decision Making: Neurocognitive Modeling and Experimentation

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Cognitive Sciences

by

Qinhua "Jenny" Sun

Dissertation Committee:
Professor Ramesh Srinivasan, Chair
Professor Joachim Vandekerckhove
Professor Jeffrey N. Rouder

2024

DEDICATION

In memory of my father Sun Yingguang.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank all the family and friends that have gone through the journey with me. I want to thank my parents, Sun Yingguang and An Shuang, for all the sacrifices they've made for me, and for always believing in me. I want to thank my husband Justin Moy, for supporting me on a daily basis in every possible way.

I would like to thank my advisor, Ramesh Srinivasan, for demonstrating what it means to be a great scientist and a supportive mentor throughout the years, and for teaching me countless invaluable skills that have helped me grow both academically and personally. I would also like to thank my thesis committee members, Joachim Vandekerckhove and Jeffrey N. Rouder, whom I greatly enjoyed learning from, as well as members of my previous milestones, Aaron Bornstein and Hung Cao.

I am fortunate to have been surrounded by supportive labmates and fellow graduate students, including Khuong Vo, whom I enjoyed collaborating closely with, Zhibin Zhou, Isaac Menchaca, Kathleen Medriano, Jeff Coon, and many others. I am also grateful to former lab members and Cognitive Science alumni, Michael Nunez and Javier Garcia, for all the professional support they provided me. I would like to thank my undergraduate advisor, Gil Einstein, a role model who I aspire to become.

I am grateful for the company and encouragement of my extended family, the Moy family, and for the close friends whom I've shared over ten years of friendship with: Aria Yuan Wang, my fellow neuroscientist who has empowered me on this journey since day one; Bella Mei and Zoey He, who have always supported me; and Christina Ye, Kyle Cai, and Yaxin He, who accompanied me at critical times despite distance.

A special note of appreciation goes to the friends I've met since 2023, a particularly challenging year in my life, including Richa Gadgil, for her uplifting spirit, and everyone in my routine badminton group, who continuously bring me positivity during tough times.

My research was supported by (1) grants #1850849 and #2051186 from the United States (US) National Science Foundation (NSF), (2) the Department of Cognitive Sciences, UCI, (3) ORAU and AEOP fellowships sponsored by DEVCOM Army Research Laboratory (ARL) with the Department of Defense (DoD).

VITA

Qinhua "Jenny" Sun

EDUCATION

Ph.D. in Cognitive Sciences with Concentration in Cognitive Neuroscience **2024**
University of California, Irvine

Master of Science in Cognitive Neuroscience **2022**
University of California, Irvine *Irvine, California*

Bachelor of Science in Psychology **2017**
Furman University *Irvine, CA*

RESEARCH EXPERIENCE

Graduate Research Assistant **2018–2024**
University of California, Irvine *Irvine, California*

TEACHING EXPERIENCE

Teaching Assistant **2018–2024**
University of California, Irvine *Irvine, California*

REFEREED JOURNAL PUBLICATIONS

Decision SincNet: Neurocognitive models of decision making that predict cognitive processes from neural signals. **2022**
Proc. of IJCNN

Deep latent variable joint cognitive modeling of neural signals and human behavior. **2024**
Neuroimage

REFEREED CONFERENCE PUBLICATIONS

Estimates of cognitive processes in decision making from neural signals by an interpretable neural network model. **2022**
Conference: 2022 Conference on Cognitive Computational Neuroscience

SOFTWARE

Decision SincNet `https://github.com/jennyqsun/EEG-Decision-SincNet`
Pytorch implementation of Decision SincNet.

ABSTRACT OF THE DISSERTATION

Uncovering Computations of Human Decision Making: Neurocognitive Modeling and Experimentation

By

Qinhua "Jenny" Sun

Doctor of Philosophy in Cognitive Sciences

University of California, Irvine, 2024

Professor Ramesh Srinivasan, Chair

The drift diffusion model (DDM) is a popular model of evidence accumulation that estimates parameters representing the underlying cognitive processes of decision making. DDM requires modeling the joint distribution of choice and response time (RT) with a Wiener first-passage time (WFPT) distribution to estimate a decision maker's speed, caution, and bias, parameterized as *drift rate*, *boundary separation* and evidence *starting point*, respectively. Recent research demonstrates promising modeling results when parameters are allowed to vary across trials while being constrained by brain signals. We have developed Decision SincNet, a novel, interpretable neurocognitive model that allows trial-level estimates of DDM by mapping EEG brain signals to the model parameters given behavioral data, through optimization of Wiener likelihood using gradient descent. Single-trial EEG data were used to represent the most likely cognitive parameters that gave rise to the observed choice response time. Critically, the lightweight neural network model was designed to automatically identify the neural correlates of different cognitive parameters in time and frequency domains without feature engineering. We showed that single-trial estimates of drift and boundary performed better at predicting RTs than median estimates did in both training and test datasets. This suggests that the model can successfully learn to extract meaningful trial-level EEG features to estimate Diffusion model parameters, and it generalizes well to out-of-sample brain

data. We improved the model's scientific interpretability by introducing an attention block to rank the learned frequency bands by importance. Interpretability methods were used to visualize how neural signals within the important frequencies were processed to estimate each parameter. The model was tested on two datasets. Results and architecture of the Decision SincNet model provided neural evidence for the DDM and demonstrated the possibility of an end-to-end neurocognitive modeling framework on a trial-level. To directly test the accumulation-to-threshold assumption in evidence accumulation, we introduced a probabilistic perceptual decision-making (pPDM) task to investigate computations of decision making processes through experimental manipulation. The task involves presenting stimuli as a succession of random samples with a known probability biased towards one of two alternatives. We quantitatively tested fundamental theories in sequential sampling during evidence accumulation. Hypothesis-driven models were used to evaluate which model best captures human behavior, under a Machine Learning framework for model comparison. Our results suggest that humans continuously integrate evidence, but multiple mechanisms are needed to explain the complex behavioral patterns observed. We developed two new models that best fit the data collected. One model is the integration of evidence weighted overtime. This model is similar to a discrete OU process with a recency effect, such that subjects focus on recent samples and suppress prior ones. The other model is a non-linear gradient boosted tree-based model that utilizes a combination of features: integration of evidence, number of samples seen, the order in which samples are presented, and pattern of recent samples. Our results suggest that computations during decision making consider all the factors as well as the interactions between them. Level of evidence and number of samples are the most dominant criterion. Subjects showed a higher probability of making a decision as evidence and the number of samples increased, but the relationships were both non-linear. Together, Decision SincNet and behavioral results from pPDM contribute to the theoretical understanding of human decision making through neurocognitive modeling and experimentation.

Chapter 1

Background

1.1 Perceptual Decision Making

Perceptual decision making is a ubiquitous cognitive process in everyday life. As humans, we constantly process sensory information and convert them into discrete categorical responses. For example, we need to decide whether we should release the brake or not given the movement of a car in the front. Contrary to choice behavior such as deciding which house to purchase, these rapid decisions made at a lower level of cognition are considered perceptual decision-making. In the field of mathematical psychology and computational neuroscience, research on perceptual decision making aims to develop computational models that describe the mechanism of how sensory input produces motor response during perceptual decision-making tasks.

Certain behavioral patterns can be consistently observed. In laboratory settings, perceptual decision making has been investigated using two-alternative forced-choice (2AFC) tasks, in which subjects are asked to use sensory information such as visual cues to make a binary response. For example, if one is asked to decide whether there are more red or blue pixels in

Fig 1.1A, one would quickly be able to provide the obvious answer with more red pixels due to the substantial supporting evidence. The task becomes more challenging if one is asked to make a decision using Fig 1.1B, as evidence between the two answers is closer in ratio, but after some deliberation one will likely conclude that there are more blue pixels. Therefore, if the choices are easy, people tend to respond faster, and if choices are hard, people tend to respond slower. If these trials are repeated many times, the faster response times (RTs) from easy condition will be narrowly distributed (as illustrated in 1.1C), and the slower RTs from hard condition will be widely distributed with a longer tail (as illustrated in1.1D). To explain this phenomenon, one of the most important theoretical views suggests that decisions are made through gradual accumulation of evidence coming from sensory information followed by an overt response [53, 49, 52].

Over the last several decades, mathematical psychologists have tried to model the behavioral data consisting of response time (RT) and accuracy from these 2AFC tasks. Such efforts have given rise to the development of a class of formal mathematical models called the sequential sampling models, suggesting that we continuously sample and integrate pieces of evidence like until a threshold is reached. These models includes the standard drift-diffusion model which has gained much popularity [38]. These models assume that a decision variable (DV) is represented by a gradual accumulation of samples drawn from noisy sensory evidence until the cumulant reaches either the upper or the lower boundary representing the two choices. Besides, the models use non-decision components to capture sensory encoding before evidence accumulation, and motor execution after it. These evidence accumulation models have gained more popularity as they have been supported by a strikingly similar diffusion process with variability over time in neural recordings in macaques [44] and humans [18].

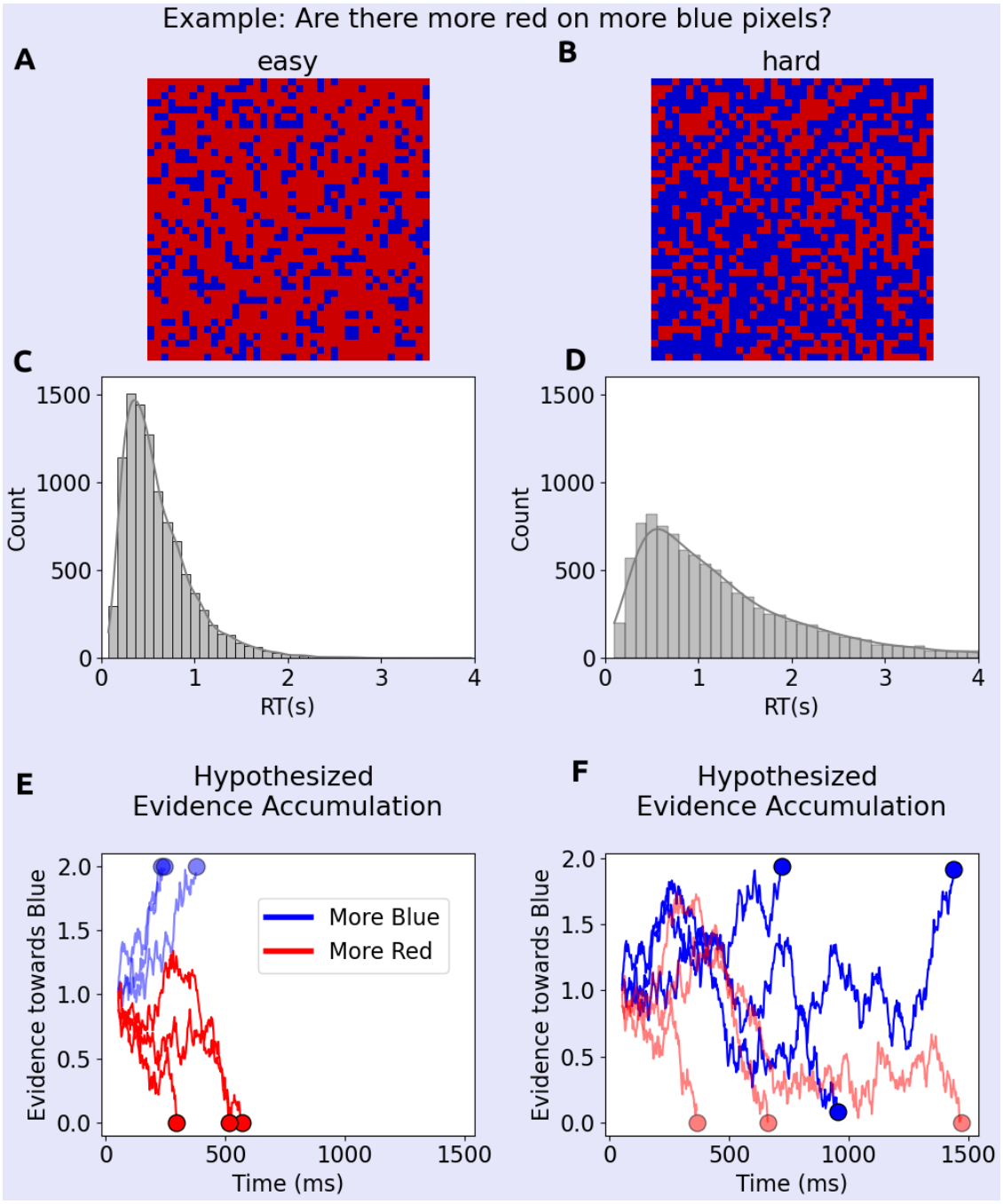


Figure 1.1: Demonstration of Evidence Accumulation during decision making. A&B: Examples of easy and hard 2AFC perceptual decision making tasks. C&D: Examples of the resulting RT distributions from easy and hard trials. E&F: The hypothesized random walks during evidence accumulation.

1.1.1 The development of classic Perceptual Decision Making frameworks

Signal Detection Theory (SDT) was one of the first and most influential frameworks that uses formal mathematics to infer the underlying noisy process purely from categorical behavioral response [16]. In SDT, the absence of a target (noise) and the presence of a target (signal plus noise) are assumed to be from two distributions. These two distributions are the underlying sensory representation learned by the decision maker, based on the physical properties of the stimuli observed. In other words, two internal states are learned: the distribution of sensory evidence when target is absent (noise), and the distribution of sensory evidence when target is present (noise + signal). Subsequently, a DV can be constructed given current sensory evidence, responding to the question "which of the two previously-learned sensory states gave rise to the current evidence?" Typically, the DV is the relative likelihood of of evidence given one state over another:

$$LLR = \log \frac{p\left(e \mid \text{target is present}\right)}{p\left(e \mid \text{target is absent}\right)} \tag{1.1}$$

Sequential Sampling Models

Thus, if the target is more likely to be present given the evidence (LLR > 0), a response will be made indicating that the target is present. If the target is more likely to be absent (LLR < 0), the alternative response will be made. Although SDT continues to be a powerful framework, it doesn't capture the time it took to make a decision into an account, and it has an underlying assumption that the DV is only guided by sensory evidence. Sequential Sampling Models are different from SDT by assuming that evidence is continually sampled overtime until a decision is made. Therefore, DV is now constructed by the accumulation

of each new sample of evidence [15]. We can start formalizing this process by using the Sequential Probability Ratio Test (SPRT). The DV now is the sum of relative likelihood between the two hypothesis test among hypothesis h_1 and h_2 at each time point (t = 1, 2,...,n) with each piece of evidence:

$$\begin{aligned} \log LR_{12} &\equiv \log \frac{P\left(e_1, e_2, \ldots, e_n \mid h_1\right)}{P\left(e_1, e_2, \ldots, e_n \mid h_2\right)} \\ &= \sum_{i=1}^{n} \log \frac{P\left(e_i \mid h_1\right)}{P\left(e_i \mid h_2\right)} \end{aligned} \tag{1.2}$$

The DV keeps being updated and it's calculated by the running sum of weights obtained by the LLR of each sample of evidence, until it reaches a positive or negative criterion (the decision bound). When the evidence asmples are Gaussian random variables, the LLR ratio is also Gaussian. Therefore, the running sum could be a process mathematically described as a random walk. Therefore, it was proposed that we can use a diffusion process, approximated by a Wiener diffusion process or Brownian motion process, to describe the brain during decision making [38]. In a general diffusion model, drift and diffusion coefficient are the mean rate of change and the variance in the rate of change of sensory evidence in a unit interval. They depend on both the time variable and the state variable, namely, the evidence. One way to define X_t, the general diffusion, is by satisfying the stochastic differential equation, as the solution is a stochastic process:

$$dX_t = \mu\left(X_t, t\right) dt + \sigma\left(X_t, t\right) dW_t \tag{1.3}$$

Where W_t is the Browanian motion process (Wiener process).

X_t is the diffusion state, dX_t denotes the the change in X over a small time interval dt. For simplicity, we define the basic, unbiased DDM assuming the drift and diffusion coefficient are constant [38]:

$$dX_t = \mu dt + \sigma dW_t, X_0 = 0 \tag{1.4}$$

Change in the amount of evidence includes two parts: The constant drift μ that represents the average increase in evidence supporting the one choice per time unit, and the second term, σdW_t, which represents white noise and has a Gaussian distribution with mean 0 and variance σ^2. Hence, X_t grows at rate μ on average, but solutions also diffuse due to the accumulation of noise.

In summary, accumulation-to-bound models such as the DDM suggest that a DV is updated by gradual sampling of sensory evidence in favor of one of the two outcomes. This framework is significant because it distinguishes sensory evidence from the DV. A DV is a variable that gets updated with time and is indicative of the final decision, whereas evidence is momentary [49]. Once a threshold is reached with enough evidence, a motor movement is triggered to make a decision. In the next section, we will review the empirical support for this sequential sampling framework.

1.1.2 Neurophysiological evidence of Sequential Sampling

One of the most studied tasks developed to study the relationship between sensory encoding and perceptual decision making is the Random-dot Motion Discrimination Task (RDM). The task uses a single stimulus containing randomly moving dots in which a certain percentage of the dots are moving coherently to one direction, and the response needs to be one of the

two directions provided. Direct neural recordings in monkeys during a RDM and revealed neural evidence of gradual decision formation in the lateral intraparietal cortex (LIP), an area that receives inputs from MT and MST and sends output to the FEF and SC [44]. They introduced a novel RT condition where the random dots extinguished once break in fixation detected, such that any modulation of LIP activity that accompanies motion viewing is not a consequence of decision but the formation of decision. They found a discharge of neurons 220ms after onset until a threshold is reached $\tilde{5}$0ms before the saccade. After 220ms, neural responses reveal the differences that lead to different decision outcomes. For the trials where the final saccades are within the LIP neuron's receptive fields, firing rates increase like a ramp, and the slope of firing rate change is dependent on motion strength and direction of stimuli. When the final saccades are outside of the receptive fields, firing rate tends to decline. These ramps end before initiation of a saccade when the neurons reach a critical firing rate. These results reflect the cognitive processes containing evidence accumulation and non-decision time when fitting the behavioral data to the diffusion model.

Followed by this line of research, other studies using non-invasive approaches such as EEG recordings have identified other signals that are associated with sensory evidence accumulation process. A late component was identified uring discrimination tasks that correaltes with mean drift rate as derived from a diffusion model simulation using behavioral responses [36]. This negative signal occurring 300ms after the stimulus is generally considered as a P300 signal. During a continuous-monitoring version of the RDM designed top avoid sensory-evoked potentials, centroparietal positive potential (CPP) has a buildup rate systematically correlated with sensory evidence strength around 170ms after onset of the stimulus [21]. In addition, CPP reached a potential when committing to a response, reflecting the accumulation-to-threshold dynamics. The slope indicating the buildup rate of CPP is associated with RT within each coherence level. Interestingly, they found that the lateralized readiness potential (LRP) in the frontocentral electrodes also showed this build-to-threshold pattern.

1.2 Estimating parameters of the Diffusion Model

One of the major challenges of modeling the DDM is the need to account for variability across trials. For each trial, cognitive components could vary due to internal and/or external noise, thus giving rise to different behaviors reflected in choice response time data. For each trial, subjects might possibly have different information processing speed, level of caution and starting point. A notable success in advancing DDM models is by allowing drift rate and starting point to vary [40]. Variability across trials can uniquely model the distributions of fast errors versus slower correct responses. Bayesian Hierarchical Drift Diffusion Model (HDDM) gained popularity as it produces full posterior distribution of each parameter to quantify uncertainty, accounts for differences across individuals and experimental conditions, and can be implemented using open-source software packages available[60, 66, 29].

Given the existing neural evidence of evidence accumulation, previous work has successfully shown that incorporating single-trial EEG measures of attention into the HDDM yields better out-of-sample predictions on accuracy and reaction time distributions relative to using behavioral data alone [31]. However, these models are bespoke and strongly confirmatory, and do not offer the flexibility needed to detect predictive patterns in an exploratory fashion. Therefore, a natural subsequent question is whether there exists a modeling approach that can simultaneously discover trial-level EEG data to constrain the estimate on single trials.

1.3 Structure of the Dissertation

In Chapter 2, we will present a method to estimate cognitive processes from neural signals by using an interpretable Neural Network model Decision SincNet. We showed the results using our original model architecture on Dataset 1, followed by an improved modeling framework demonstrated on both Dataset 1 and Dataset 2. In Chapter 3, we will present interpretabil-

ity methods that can be used on Decision SincNet, and show how to utilize the design of Decision SincNet to gain scientific insights regarding the relationship between brain and latent cognitive processes. In Chapter 4, we will introduce a new task to test fundamental theory in decision making, and showed behavioral results along with new models that we developed.

Chapter 2

Decision SincNet: Neurocognitive models of decision making that predict cognitive processes from neural signals

2.1 Introduction

Perceptual decision making is a crucial part of cognition. Humans must rapidly translate sensory information into behavioral responses in order to achieve their goals, e.g., stop at a red traffic light. In the field of mathematical psychology and computational neuroscience, much effort has been dedicated to develop computational models that describe the mechanism of perceptual decision-making. The Drift-Diffusion Model (DDM) is the most widely used model [39] to explain choice and response time data in perceptual decision making tasks, by assuming that decisions are made through sequential sampling and integration of

sensory information[31].

Fig. 2.1 depicts the Drift-Diffusion Model (DDM), where the x-axis is the time from viewing of a stimulus. After some processing time for neural encoding of the sensory stimulus, a Decision Variable (DV) begins a random walk process from starting point z between two decision boundaries, which represent the two choice options. The DV is updated by gradually accumulating samples drawn from noisy sensory evidence in favor of one of the two outcomes, until it reaches either the upper or the lower bound. Subsequently, the motor system then executes a response after the DV reaches one of two boundaries.

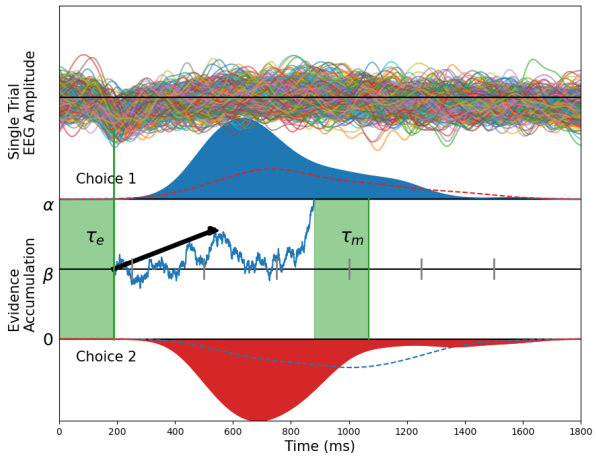


Figure 2.1: Demonstration of Drift Diffusion Model during a two-choice decision making task with Non-deicison time indicated in green. For a given trial, after visual encoding time (τ_e), a DV begins evidence accumulation process and will hit either the upper of the lower bound. Mean rate of evidence accumulation of is indicated by the vector in black. Blue curve indicates the probability density function (pdf) of RTs when Choice 1 is correctly executed, and red curve indicates the pdf of RTs when Choice 2 is correctly. An error is made when the DV drifts to the wrong bound due to noise. Dotted curve indicates the pdf of RTs for incorrect trials. The axis above shows the EEG data for each trial after using SVD to maximize N200 signals, which can be used to track onset of evidence accumulation.

The average increase of change in a unit interval during evidence accumulation is parameterized as the drift rate δ, and the instantaneous variance in the rate of change is parameterized as diffusion coefficient ς. This variance simply scales the model and is fixed to 1. Bound-

ary separation α describes the distance between two choices. The parameter β encodes the starting position of evidence accumulation, which reflects the bias towards one of the two choices. When β is 0.5, the starting position is in the middle between two boundaries and thus the start of evidence accumulation is unbiased between the two choices. Visual encoding time before evidence accumulation and motor execution after evidence accumulation could be expressed as τ_e and τ_m [30] respectively, but only the sum of both processes, non-decision time τ, can be observed with behavior alone. The model is also referred to as Wiener Diffusion Model, because an unbiased, continuous evidence accumulation can be expressed as follow[5]:

$$dX_t = \delta dt + \varsigma dW_t,\ x(0). \tag{2.1}$$

X_t denotes the diffusion state, dX_t denotes the change in X over a small time interval dt, and W_t is the Wiener Process. Evidence accumulated continuously will result in the distribution of boundary cross times, described as the Wiener first-passage time (WFPT) distribution: $\mathcal{T} \sim Wiener(\alpha, \beta, \tau, \delta)$.

One of the most critical aspects of the DDM is that different parameters can empirically represent underlying cognitive components of the decision making process. During decision making experiments, separate manipulation of processing speed, caution, bias, and motor execution could be reflected in changes in the corresponding parameters δ, α, β, and τ [64]. Thus, there has been considerable interest in linking the DDM parameters to brain activity, and more recently, interest in developing neurocognitive models that incorporate brain activity directly into the diffusion model framework to obtain novel insights into decision making, in particular, to the variability in decision making across observations.

Simulation studies with the DDM have demonstrated the importance of trial-to-trial variability in drift rate, starting point, and boundary to empirically model two-choice response

time data [40]. Allowing drift rate and starting point to vary from trial to trial can uniquely capture properties of response time distribution for correct and incorrect choices. Variability in caution reflect adjustments in speed-accuracy trade-off.

Although variability in cognitive processes in decision is well-known, choice-response time does not provide enough data to estimate model parameters at a trial level. Often models with model-intrinsic trial-to-trial variability, such as assumptions of a normal distribution of drift rates are fit to data. Fitting these models provide estimates of summary parameters across trials (e.g. mean and standard deviation of the drift rates across trials). One method to estimate DDMs while allowing intrinsic trial-to-trial variability of parameters is to use Bayesian Hierarchical Diffusion Models [60]. The Bayesian modeling approach is used to compute posterior distribution of model parameters, and employs Monte Carlo Markov Chain (MCMC) algorithm to generate samples from the posterior distribution until they converge.

Accumulation-to-bound patterns have also been found in different areas of the brain[44, 34]. There has been a growing interest to incorporate neural data along with behavioral data to build *neurocognitive* models of decision making [31, 30, 58, 57]. Previous work has shown that incorporating single-trial EEG measures of attention into the HDDM yields better out-of-sample predictions on accuracy and reaction time distributions relative to using behavioral data alone[31]. Model parameters (δ, τ, ς) on each trial were assumed to be a linear combination of single-trial EEG. The EEG measures were derived using known stimulus-locked EEG signals, i.e., event-related potentials (ERPs) estimated by averaging the trials. More specifically, to augment the signal-to-noise (SNR) ratio of single-trial EEG, singular value decomposition (SVD) was used to find channel weights that maximally explain the variance of specific evoked potentials (N200, P200). These weights were applied to single-trial EEG to obtain latency and amplitude measures of the N200/P200 per trial as regressors onto HDDM parameters. This successful line of work has suggested that trial-level neural dynamics account for some of the trial-to-trial variability in the HDDM model

parameters. However, while using MCMC alone can robustly estimate posterior distributions from which the trial parameters are drawn, it does not have the resolution to directly link trial-level neural representation to trial-level parameters estimated. Moreover, this approach was limited to using trial averaged ERP signals to provide a template of hypothetical signals linked to DDM parameters.

The current research aims to use a neural network to estimate drift rate δ and boundary α parameters of the DDM on single trials directly from the raw EEG data. Machine learning approaches such as filter bank common spatial patterns (FBCSP) have been widely used to extract EEG features [2], but it has the disadvantage of requiring artificially-selected frequency bands. Convolutional Neural Networks (CNNs) have shown promising results on decoding brain activities [1, 68] using raw EEG data, but filters and feature maps obtained can be hard to interpret. SincNet is a recently proposed deep learning neural network used to process raw time series data such as speech and EEG data[42]. The key feature of SincNet is that each kernel from the first layer of CNN is parameterized as a sinc function and acts as a band-pass filter that could be applied to the time series. Two cut-off frequencies are the only trainable parameters needed. Therefore, SincNet is advantageous because fewer parameters are needed and the filter parameters are themselves optimized by the training data. The model structure has successfully been used on different EEG decoding tasks since its inception[6, 27].

We developed a Decision SincNet model that can learn the windows of time and frequency bands in the EEG that are useful to predict trial level DDM parameters. This Decision SincNet model is scientifically significant for the following reasons.

1. By using the WFPT likelihood as a loss function, we can use gradient descent to learn an end-to-end model to fit two DDM parameters (drift rate and boundary) from raw EEG data, and apply it to predict cognitive parameters in new unseen brain data.

2. By using neural network models, we can get trial-level predictions of parameters

3. By applying interpretation techniques to the SincNet layer and depthwise layer, we can automatically learn relevant neural dynamics (e.g., time windows and frequency bands) that are critical to modeling the evidence accumulation process.

In this chapter, we will focus on the architectural design, training, and evaluation framework of Decision SincNet. The interpretability techniques and results used on Decision SincNet will be discussed in Chapter 3.

2.2 Methods

2.2.1 Two Behavioral and EEG datasets

Two dataset were used to evaluate the Decision SincNet model. Both dataset were collected at Univeristy of California, Irvine at the Human Neuroscience Lab.

Dataset 1

Dataset 1 contains behavioral and EEG data collected while participants ($n = 45$) performed a two-alternative forced-choice task where they discriminated whether a Gabor patch presented with added dynamic noise is higher or lower spatial frequency. Task difficulty was manipulated by the difference in spatial frequency (perceptual evidence) between the two choices in order to manipulate the evidence available to make the discrimination. Participants performed the task in blocks of trials at 3 levels of difficulty (spatial frequency difference). Each subject performed 360 trials, while 128 channels of EEG and behavioral data were recorded. Independent Component Analysis (ICA) based artifact rejection method was

used on EEG data to remove eyeblinks, electrical noise, and muscle artifact. EEG data were bandpass filtered to 1 to 50 Hz, and then downsampled from 1000 Hz to 500 Hz prior to data analysis. The data for each subject were divided into 80% for training and the remaining 20% for testing.

Dataset 2

Dataset 2 contains behavioral and EEG data collected while participants performed a two-alternative forced-choice task where they had to decide whether a Gabor patch presented with added dynamic noise is higher or lower spatial frequency [for details, see Experiment 2 by 30]. Task difficulty was manipulated by adding spatial white noise to manipulate the quality of the perceptual evidence available to make the discrimination. The signal and the noise flickered at 40 and 30 Hz frequencies, respectively. Figure 2.2 is an example of a high sptial frequency stimulus. 4 participants performed the task in blocks of trials at 3 added noise levels (low, medium, and high). Each subject performed approximately 3000 trials over 7 experimental sessions, while 128 channels of EEG and behavioral data were recorded. The independent component analysis (ICA)-based artifact rejection method was used on EEG data to remove eyeblinks, electrical noise, and muscle artifacts. A subset of 98 EEG channels were selected, excluding channels located in the outer ring. EEG data were bandpass filtered to 1 to 45 Hz in the frequency domain and then downsampled from 1000 Hz to 500 Hz in the time domain prior to data analysis. The data for each subject were divided into 80% for training and the remaining 20% for testing.

2.2.2 Wiener First-Passage Time Model of Response Time

For each trial, the likelihood of Wiener First-Passage Time (WFPT) was calculated with RT (t) by using a small-time approximation

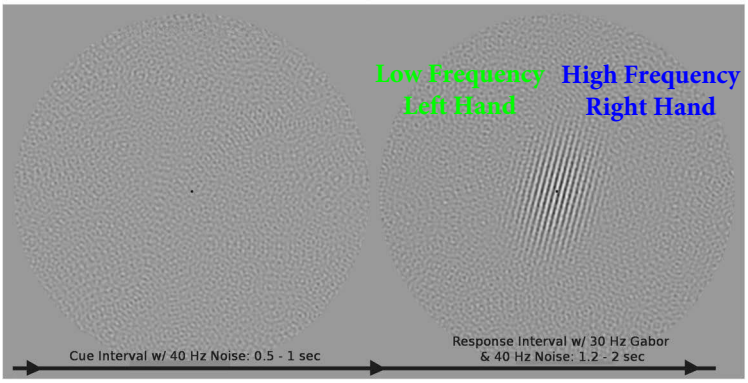


Figure 2.2: Demonstration of a high spatial frequency stimulus [30].

$$\begin{aligned}\text{Wiener}(t \mid \delta, \alpha, \beta, \tau) = \frac{1}{\alpha^2} \exp\left(-\delta\alpha\beta - \frac{\delta^2 t}{2}\right) \\ \times \frac{1}{\sqrt{2\pi(t-\tau)^3}} \sum_{[(m-1)/2]}^{[(m-1)/2]} (\alpha + 2m) \\ \times \exp\left(-\frac{(\beta+2m)^2}{2(t-\tau)}\right)\end{aligned} \tag{2.2}$$

where m is the number of expansion terms required for approximation of the likelihood function. We fix m to be 10, as it is sufficient for the approximation for modeling data where $t < 2$[28]. β is set to be 0.5, so that the starting point is always unbiased at $z = \beta\alpha$. τ is non-decision time, which is set to be $0.93 \cdot RT_{min}$ for each subject, approximating Bayesian MCMC modeling results [30].

Given a dataset $\mathcal{D} := \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ consisting of N trials $\boldsymbol{x}_n \in \mathbb{R}^{C \times D}$, i.e., EEG signals of C channels by N time samples, and corresponding observed response times $t_n \in \mathbb{R}, n = 1, \ldots, N$. The likelihood factorizes according to

$$
\begin{aligned}
p(\mathcal{T} \mid \mathcal{X}, \boldsymbol{\theta}) &= p\left(t_1, \ldots, t_N \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{\theta}\right) \\
&= \prod_{n=1}^{N} p\left(t_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}\right) \\
&= \prod_{n=1}^{N} \text{Wiener}\left(t_n \mid \delta(\boldsymbol{x}_n, \boldsymbol{\theta}), \alpha(\boldsymbol{x}_n, \boldsymbol{\theta}), \beta, \tau\right)
\end{aligned}
\tag{2.3}
$$

where we defined $\mathcal{X} := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ and $\mathcal{T} := \{t_1, \ldots, t_N\}$ as the sets of inputs and corresponding targets, respectively. Both the drift rate δ and the boundary α are the functions of $\boldsymbol{x}$ parameterized by the a deep neural network $\boldsymbol{\theta}$.

2.2.3 Optimization and Loss Function

To find the optimal parameters $\boldsymbol{\theta}^\star$ of the non-linear regression problem, we minimize the negative log-likelihood

$$
\begin{aligned}
\boldsymbol{\theta}_\star &= \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\theta) := \arg\min_{\boldsymbol{\theta}} -\log p(\mathcal{T} \mid \mathcal{X}, \boldsymbol{\theta}) \\
&= \arg\min_{\boldsymbol{\theta}} -\sum_{n=1}^{N} \log \text{Wiener}\left(t_n \mid \delta(\boldsymbol{x}_n, \boldsymbol{\theta}), \alpha(\boldsymbol{x}_n, \boldsymbol{\theta}), \beta, \tau\right).
\end{aligned}
\tag{2.4}
$$

In the initial modeling attempt [54], we discovered that the boundary parameter α tends to overfit first as the gain of varying the boundary parameter outweighs the gain of varying the drift parameter given the nature of the WFPT likelihood function. Therefore, we first used the average boundary within a training batch $\bar{\alpha}$ instead of α in Equation 2.4 to update trial-level boundary while considering the fit within a batch. The solution to this issue is addressed in the models in follow-up experiments.

2.2.4 Model Architecture

Table 2.1: Architecture of the Decision SincNet Model

Block	Layer	# filters	size	# params	Output	Activation
1	Input				(98, 500)	
	BatchNorm			2	(98, 500)	
	SincConv2D	32	(1, 131)	64	(32, 98, 370)	
2	BatchNorm			64	(32, 98, 370)	
	SeparableConv2D	64	(98, 1)	6272	(64, 1, 370)	ReLU
3	BatchNorm			128	(64, 1, 370)	
	AvgPool2D		(1, 45)		(64, 1, 8)	
	Dropout				(64, 1, 8)	
4	Flatten				512	
	Dense (drift rate)			513	1	
	Dense (boundary)			513	1	Sigmoid

The Decision SincNet model is built to use trial-level EEG data to simultaneously predict drift rate and boundary separation in an individual participant. The model architecture is similar to Sinc-ShallowNet[42] and consists of four blocks, as shown in Table 1 and visualized in Fig. 2.3. The model applies band-pass filters and spatial filters on the EEG data, pools the filtered data, and finally predicts the two parameters of the diffusion model in-parallel. There are 7556 trainable parameters in total. All the methods mentioned below may be reproduced by using the code on Github (https://github.com/jennyqsun/EEG-Decision-SincNet). The Deep Learning Models were developed in PyTorch [35], and trained from scratch using a workstation equipped with NVIDIA GeForce RTX 2080 Ti and 64 GB of RAM. On average, a model per subject takes five minutes to train.

The first block is a Sinc-convolutional layer[6]. Thirty-two kernels are parameterized by two cutoff frequencies of a Sinc Function. The low and high frequencies (f_1 and f_2) are trainable parameters of the model during learning, so the kernel can be expressed in the time domain as:

$$\boldsymbol{k}_j[n] = 2f_2 sinc(2\pi f_{2,j} n) - 2f_1 sinc(2\pi f_{1,j} n) \tag{2.5}$$

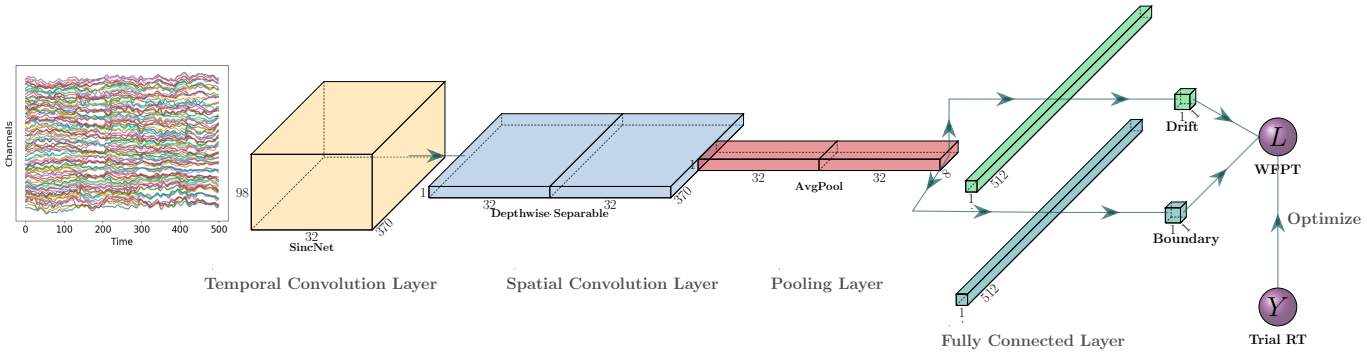


Figure 2.3: Visualization of Decision SincNet model architecture.

where Sinc Function is defined for $x \neq 0 \quad \text{bysinc}(x) = \frac{\sin(\pi x)}{\pi x}$. To avoid ripples in the signal, a Hamming window is applied to the bandpass filters[42]. Thus, the output from the SincNet layer is from a 2D Convolution between i_{th} trial EEG data and the j_{th} kernel $\boldsymbol{k}_j$. Since the kernel size has only one dimension, it is equivalent to 1D Convolution between signals from each channel $\boldsymbol{x}_c$ and $\boldsymbol{k}_j$:

$$\boldsymbol{y}_{c,j}[n] = \boldsymbol{x}_c[n] * \boldsymbol{k}_j[n] = \sum_{l=0}^{L-1} \boldsymbol{x}_c[n-l] \cdot \boldsymbol{k}_j[l] \tag{2.6}$$

where $c \in [0, C-1]$, $j \in [0, K-1]$ with K representing the total number of kernels, $\boldsymbol{y}_{c,j}[n]$ is the filtered signals as the output from the Sinc layer, and L is the length of the kernel. We set $L = 131$ such that lowest central frequency possible is around 4 Hz when sampling rate is 500 Hz[6].

The second block is a depthwise convolution layer where spatial kernels are applied to time series data with a depth of 2. Each temporally-filtered version of the input convolves with two spatial filters followed by the ReLU activation function. The third block consists of pooling operation. The window size and stride size are always set in a way such that they are equivalent to 250ms and 100ms, respectively, when scaling back to the original input.

Following the Multi-task learning (MTL) paradigm [45], the last block has two "task-specific"

layers that share all the previous hidden layers. MTL improves model performance by acting as a form of regularization, causing the model to prefer hypotheses that explain more than one task. The shared temporal- and spatial-filtered features are flattened and fed to two separate fully-connected (FC) layers to simultaneously produce two outputs, trial drift rate (δ) and trial boundary (α). The two parameters are fit using the same loss function during training. Predicted drift rates are clamped such that $\hat{\delta} \in [-6, 6]$, in order to avoid extreme values of the negative log likelihood. A modified Sigmoid activation function,

$$\Phi(z) = \exp\left(\frac{1}{1+\exp(-z)}\right) \tag{2.7}$$

constrains the predicted alpha to a realistic range, $\hat{\alpha} \in (1, 2.72)$.

To constrain the predicted value of boundary, output from its corresponding FC layer is followed by a modified Sigmoid activation, as shown in Fig. 2.4: The derivative of the

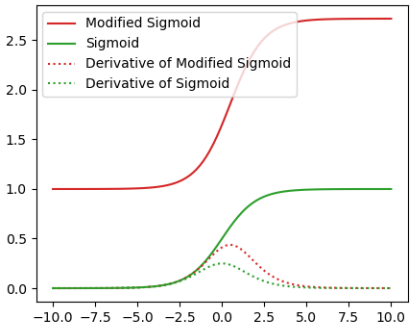


Figure 2.4: Demonstration of the modified Sigmoid activation function used for the FC layer to predict boundary.

modified Sigmoid is:

$$\Phi'(z) = \frac{\exp(\frac{1}{\exp(-z)+1} + z)}{(\exp(z)+1)^2} \tag{2.8}$$

Given a drift rate δ, the DV could hit either of the boundaries due to noise. Variability across time in the diffusion process can drive the process to terminate at the wrong boundary by mistake. A diffusion model that views correct and incorrect choices as the two boundaries will produce two different shapes of probability distributions. Decision SincNet is designed to make a point estimate of diffusion model parameters using one EEG trial. We modeled the two boundaries as the two choices, where drifting to the upper bound represents one choice, and drifting to the lower bound represents the process towards the other choice, and assumed there was no starting bias between the choices, which is a reasonable assumption for the data sets tested here. The response times for two choices are assumed to have identical probability density function and could be estimated using the same likelihood function. The main purpose of our approach is to directly predict model parameters from the EEG data on each trial.

2.2.5 Training

Before training, each subject's data were split into 80% for training and validation, and the other 20% for testing. For each subject, 1 second of EEG trials after stimulus onset were used as inputs, and the corresponding RTs were used as training labels. Separate models are trained for each subject. Initial bandpass filters were randomly selected from a uniform distribution, $U \sim (1, 32)$. Weights are updated by the gradient descent as

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_i - \eta \frac{\partial}{\partial \boldsymbol{\theta}_i} \mathcal{L}(\boldsymbol{\theta}) \tag{2.9}$$

η is the learning rate and it was set to be 10^{-3}. The model was trained using batch size of 64, and Adaptive moment estimation (Adam) optimizer was used. To avoid overfitting, early stopping technique was applied with a patience score of 20.

2.2.6 Model Evaluation

Because Decision SincNet predicts two parameters that can not be observed directly, model performance was evaluated by focusing on the following issues: model comparison, trial-to-trial variability of the parameters within each subjects, Log Likelihood Test, model generalization to unseen data, and uncertainty of model prediction.

Comparison Between Decision SincNet and Bayesian MCMC Models

The distribution of trial estimates for a subject obtained from Decision SincNet should fall within reasonable ranges. We compared our results to Bayesian MCMC model fits using JAGS sampler to estimate posterior distributions of the model parameters using only behavioral data. The fit of a basic model use the following prior structure[31]:

$$\begin{aligned} (\delta|\mu,\sigma) &\sim \mathcal{N}(0,\, \sigma^2), \sigma^2 \sim \Gamma(1,1) \\ (\tau|\mu,\sigma) &\sim \mathcal{N}(0.5,\, \sigma^2), \sigma^2 \sim \Gamma(0.3,1) \\ (\alpha|\mu,\sigma) &\sim \mathcal{N}(1,\, \sigma^2), \sigma^2 \sim \Gamma(1,1) \end{aligned} \tag{2.10}$$

After obtaining the posterior distributions of the parameters, we compare the distribution of the trial estimates and check if they were within sensible ranges.

Trial-to-trial Variability of the Estimates

Spearman's Rank-Order Correlation was used to examine the direction and strength between estimated parameters and observed RTs. Strong correlations would suggest that the model captures trial-by-trial neural variability that are reflected in RTs and can be mapped onto drift rates and boundaries.

Model Comparison by Log Likelihood of WFPT

To check whether EEG data have the resolution to give rise to meaningful trial-level estimates, we can compare the sum of negative log-likelihood $-\sum \log$ Wiener $(t_i \mid \delta, \alpha, \beta, \tau) \triangleq \ell(\delta, \alpha \mid t)$ of the training RT between trial level estimates of parameters and median estimates of parameters. We compare the possible combinations of likelihood of single trial estimates with median estimates, namely, $\ell\left(\delta, \alpha \mid t_i^{train}\right)$, $\ell\left(\delta, \overline{\alpha} \mid t_i^{train}\right)$, $\ell\left(\delta, \overline{\alpha} \mid t_i^{train}\right)$, $\ell\left(\overline{\delta}, \overline{\alpha} \mid t_i^{train}\right)$. If the sum of the negative likelihood of either trial drift and trial alpha is smaller than that of median drift and alpha, we can conclude that the brain data is meaningfully linked to the trial to trial variability in behavior.

2.2.7 Generalization

To check whether the trained model could be generalized to unseen data, we compared sum of negative log-likelihood of test RT data using the median parameter estimates obtained from training data $\ell\left(\overline{\delta^{train}}, \overline{\alpha^{train}} \mid t_i^{test}\right)$ with all combinations of trial parameter estimates from trial EEG data $\ell\left(\delta^{test}, \alpha^{test} \mid t_i^{test}\right)$. Higher likelihood indicates that single trial EEG produced meaningful trial level predictions of DDM parameters.

Predictive Uncertainty

Monte Carlo dropout [14] is applied to estimate the uncertainty of predictive drift rates and boundaries. Mathematically, the use of dropout in a neural network can be viewed as Bayesian approximation of a Gaussian process [10]. Dropout is enabled during predictions in which different neurons are randomly deactivated with a constant dropout rate. After the repeated random sampling process, we can approximate the following predictive distribution

$$p\left(\delta, \alpha \mid \boldsymbol{x}\right) = \int p\left(\delta, \alpha \mid \boldsymbol{x}, \boldsymbol{\theta}\right) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \tag{2.11}$$

2.2.8 Follow-up: Customization of loss function and variants of model architecture

Customization of loss function

To solve the issue on trial-level boundary parameter, we then find the optimal parameters $\boldsymbol{\theta}^{\star}$ of the model, we minimize the sum of negative log-likelihood while maximizing the correlation between drift parameter and RT:

$\boldsymbol{\theta}^{\star} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\theta)$ with

$$\mathcal{L}(\theta) = -\sum_{n=1}^{N} \log\left(t_n \mid \delta(\boldsymbol{x}_n, \boldsymbol{\theta}), \alpha(\boldsymbol{x}_n, \boldsymbol{\theta}), \beta, \tau\right) - \mathrm{corr}(\boldsymbol{\delta}, \boldsymbol{t_n}). \tag{2.12}$$

in order to better guide the model to find solutions for both parameters.

We also experimented with another method to constrain the boundary parameter by adding a regularization penalty term on α parameter the in the loss function during optimization:

$$\mathcal{L}(\theta) = -\sum_{n=1}^{N} \log\left(t_n \mid \delta(\boldsymbol{x}_n, \boldsymbol{\theta}), \alpha(\boldsymbol{x}_n, \boldsymbol{\theta}), \beta, \tau\right) - \mathrm{corr}(\boldsymbol{\delta}, \boldsymbol{t_n}) - \sum_{n=1}^{N} \log(\boldsymbol{\alpha}). \tag{2.13}$$

This is because suppose to estimate α, with Bayes Rule, we have:

$$\begin{aligned}
\hat{\boldsymbol{\alpha}} &= \arg\max_{\boldsymbol{\alpha}} L(\alpha \mid t) \\
&= \arg\max_{\boldsymbol{\alpha}} p(t \mid \alpha) p(\alpha) \\
&= \arg\max_{\boldsymbol{\alpha}} \log(p(t \mid \alpha) p(\alpha)) \\
&= \arg\max_{\boldsymbol{\alpha}} \log p(t \mid \alpha) + \log p(\alpha)) \\
&= \arg\min_{\boldsymbol{\alpha}} -\log p(t \mid \alpha) - \log p(\alpha)
\end{aligned} \tag{2.14}$$

If $p(\alpha)$ is a Gaussian distribution, the prior equivalent to L2 regularization on parameter α [43].

Lastly, we experimented with adding Ridge and Lasso regularization on the Neural Network parameters $\boldsymbol{\theta}^{\star}$, in order to prevent the model from overfitting and producing in extreme values that are hard to interpret.

Variants of Model Architecture

For Dataset 2, we also experimented with different aspects of the model architecture. The original model utilizes Multi-Task Learning (MTL) and split the layers after the spatial convolution with a depth dimension > 1, resulting in three sets of shared spatial weights for each frequency bands passed to separate linear layers to make the final predictions. We compared this architecture with models that reduce the depth dimension to 1 so that only one set of spatial filters are learned within each frequency for each parameter. Additionally, we experimented to determine whether the Sinc layer should be shared or not. When the Sinc layers are trained separately to predict the two parameters, we always set the spatial depth to 1, so that each parameter has only one set of Sinc kernels and one corresponding set of spatial weights. We also experimented with whether the cutoff frequencies of the

Sinc filters should be clamped to avoid picking up high frequency noise. Table 2.2 shows a summary of the final models that were selected based on model performance and stability.

We also used the same architecture to simultaneously predict choice for each trial. Note that, since choice is not optimized in the WFPT loss function and instead just adding a binary Cross Entropy loss without sharing any of the layers, this approach is equivalent to training a separate network for classficication. However, we stack them together so that both DDM parameters and choice can be predicted using a single module. The intent for combing choice and DDM parameters was to introduce choice within the Wiener loss function to capture correct and incorrect trials.

Model	**Share Sinc Layer**	**Spatial Depth**	**Clamp**	**Prior on α**	**Regularization**
Model 1	yes	yes	no	no	no
Model 2	no	no	no	no	no
Model 3	no	no	no	no	L2
Model 4	no	no	no	yes	no
Model 5	no	no	yes	yes	no
Model 6	yes	no	yes	yes	no
Model 7	yes	no	no	no	L1

Table 2.2: Summary of final models selected for Model Comparison using Dataset 2.

2.3 Results

2.3.1 Dataset 1: Original Model

Model Performance

Fig. 2.5 presents the comparison of model performance of one subject using the trained Decision SincNet model. Fig. 2.5A and Fig. 2.5B show the trial-by-trial correlations between estimated parameters and observed RTs from the training data. The Spearman correlation between fitted drift rates (δ) and 1/RT are strong ($\rho = 0.937, p < 0.001$). 1/RT is used as

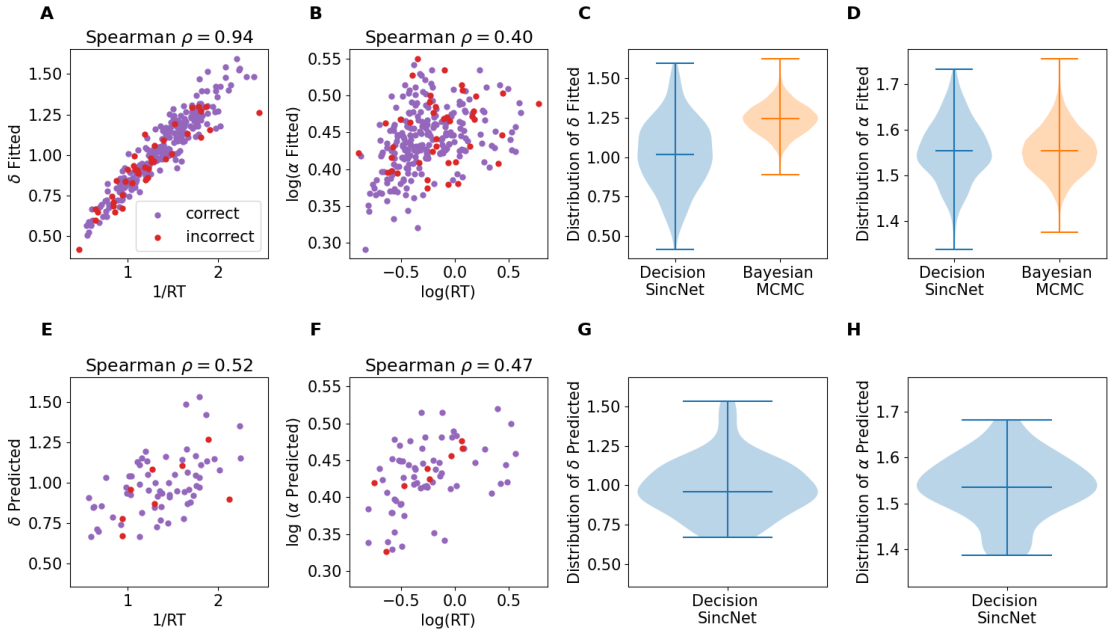


Figure 2.5: Model performance of one subject. First row are results of training data using the Decision SincNet model for this subject. Second row are results on test data. Panel A and B show model performance on fitted results. Scatter plots depict parameters predicted from trained model against observed RTs for each trial. Trials with correct responses are labeled in purple, and trials with incorrect responses are labeled in red. Panel C and D, comparisons of distributions based on drift rates and boundaries estimated from Decision SincNet with Bayesian MCMC (without EEG data). Distributions of Decision SincNet (blue) are plotted using all of the trial estimates, and distributions of Bayesian MCMC (orange) are plotted using the 30,000 samples obtained from MCMC when the Markov chains have converged. The three bars indicate the top, median, and bottom of the violin's distribution, respectively. Panel E and F show model performance on test data. Panels G and H show distributions based on drift rates and boundaries predicted from test EEG data using Decision SincNet. 1/RT are used as a proxy of drift rates on x-axis in Panels A and E. Log(RT) and Log(α) is used in Panels B and F for visualization purposes.

a proxy of speed of information processing. Spearman Correlation between fitted alpha and RT are moderately strong ($\rho = 0.401, p < 0.001$). Logarithm of RT and boundary is used to uncluster the data for visualization.

Fig. 2.5C and Fig. 2.5D compare the distributions based on parameters estimated by different models from the training EEG data. Random samples from the posterior distributions based on Bayesian MCMC model are drawn to match to the samples size of Decision SincNet during

training, and Kullback–Leibler (KL) Divergence was calculated for each of the parameters. Divergence of distribution of trial drift rates based on Decision SincNet from the posterior distribution of drift based on Bayesian MCMC is 77.714. Divergence of distribution of trial boundary based on Decision SincNet from the posterior distribution of alpha based on Bayesian MCMC is 0.347. These results suggest that single trial drift rates estimated from Decision SincNet contains more trial specific information, as it is more distinct from the posterior distribution of drift from the Bayesian MCMC.

Fig. 2.5E and Fig. 2.5F show the trial-by-trial correlations between estimated parameters and observed RTs from test data, which the model has never been seen. The correlation between fitted drift rates and 1/RT are moderately strong ($\rho = 0.517, p < 0.001$), and the correlation between fitted alpha and RT are moderately strong ($\rho = 0.468, p < 0.001$). More examples examples are shown here in the Fig. 2.6. Each column represents the performance of a subject, and each row represents the relationship between response times (x-axis) and parameters estimated from the model (y-axis). Spearman correlations are also reported in caption.

We also compared the sum of negative likelihood as described above in Method Section (E.3). Our results indicate that using single trial estimates of both drift and boundary performs the best, as $\ell\left(\delta, \alpha \mid t_i^{train}\right)$ has the lowest value. This suggests that the model was trained to extract sufficient brain data to distinguish RT variability on each trial by mapping brain signals to model parameters. When we performed out-of-sample prediction by using the trained model to predict unseen EEG data, all the likelihoods of the single trial estimates $\ell\left(\delta, \alpha \mid t_i^{test}\right)$, $\ell\left(\delta, \overline{\alpha} \mid t_i^{test}\right)$, $\ell\left(\overline{\delta}, \alpha \mid t_i^{test}\right)$, outperform the likelihood of the median estimates obtained from training data, $\ell\left(\overline{\delta^{train}}, \overline{\alpha^{train}} \mid t_i^{test}\right)$, implying a successful predictive model with ability to generalize to unseen data. The sum of negative likelihood results of the subject in Fig. 2.6is shown as an example in Table 2.3.

$\bar{\delta}, \bar{\alpha} \mid RT_{\text{train}}$	$\delta_i, \alpha_i \mid RT_{\text{train}}$	$\delta_i, \bar{\alpha} \mid RT_{\text{train}}$	$\bar{\delta}, \alpha_i \mid RT_{\text{train}}$
506.811	**454.673**	498.069	467.614

$\bar{\delta}, \bar{\alpha} \mid RT_{\text{test}}$	$\bar{\hat{\delta}}, \bar{\hat{\alpha}} \mid RT_{\text{test}}$	$\hat{\delta}_i, \hat{\alpha}_i \mid RT_{\text{test}}$	delta $_i, \bar{\hat{\alpha}} \mid RT_{\text{test}}$	$\bar{\hat{\delta}}, \hat{\alpha}_i \mid RT_{\text{test}}$
129.03	131.69	**118.397**	131.508	119.047

Table 2.3: Negative Log Likelihood for the subject from Fig. 2.5.

Monte Carlo Dropout

Fig. 2.7 is the results from Monte Carlo Dropout that captures the uncertainty of the trained model. We used the model trained for the subject and apply different dropout of neurons at test time for 5000 times. Blue dots in the top figure are median drift rate from the predictions sorted from small to large, and blue dots in the bottom figure are the median boundary estimates sorted. Red dots are the corresponding 1/RT for that trial as proxy of speed. Dark and light blue area represents the Standard Deviation, and the 90% Credible Interval respectively. The posterior distributions of model predictions of both parameters fall within reasonable ranges.

The model was able to fit and predict drift and boundary for all subjects (n=45) within reasonable ranges. During training, 41 out of 45 subjects showed the likelihood test of the single trial estimates of both parameters outperform the likelihood test of the median estimates, suggesting the benefit of single trial estimates for model fitting. When we test the trained model on unseen data, 14 out of 45 subjects have demonstrated that the likelihood of having either or both single trial estimates outperforms the likelihood of the median estimates, indicating that meaningful single-trial parameter estimates could be generalized to out-of-sample brain data for a subset of the subjects. We believe the performance in out-of-sample prediction would be improved with more data in each participant.

$\bar{\delta}, \bar{\alpha} \mid RT_{\text{train}}$	$\delta_i, \alpha_i \mid RT_{\text{train}}$	$\delta_i, \bar{\alpha} \mid RT_{\text{train}}$	$\bar{\delta}, \alpha_i \mid RT_{\text{train}}$
511.452	**441.696**	492.361	463.776

$\bar{\delta}, \bar{\alpha} \mid RT_{\text{test}}$	$\bar{\hat{\delta}}, \bar{\hat{\alpha}} \mid RT_{\text{test}}$	$\hat{\delta}_i, \hat{\alpha}_i \mid RT_{\text{test}}$	delta $_i, \bar{\hat{\alpha}} \mid RT_{\text{test}}$	$\bar{\hat{\delta}}, \hat{\alpha}_i \mid RT_{\text{test}}$
130.291	123.621	**113.579**	123.812	113.643

Table 2.4: Negative Log Likelihood from the same subject from Dataset 1 shown above with the improved loss function.

Results using revised loss function with correlation term

We tested the improved loss function using Equation 2.12 on Dataset 1, and were able to use trial-level boundary to fit the model with better results. Figure 2.8 shows the same subject as in Figure 2.5 fitted with the same model architecture (as in Table 2.1) and the new loss function. Boundary Parameter can fit better both in training and test data. Drift is more consistent with posterior distribution using MCMC results, and boundary has the same posterior median from MCMC. Table 2.4 demonstrates that single-trial drift and boundary given the test RT has the lowest NLL among other estimates. The NLL test also outperformed the previous model. These results indicate that the model has improved its ability to make out-of-sample predictions when trial-level boundary parameter is optimized.

2.3.2 Dataset 2: Model Comparison

Figure 2.9 shows the out-of-sample performance on test data for each subject in Dataset 2 with the models in Table 2.2. Two out of the four subjects can fit the model well. Model 7 demonstrated the best performance, utilizing shared Sinc filters, separate spatial weights, and and L1 regularization on the neural network parameters θ_*. Model 4 also performs well, employing separate Sinc Layers and regularization on the boundary parameter α. These results suggest that the previous architecture with multiple spatial depths was not essential.

Results of the best-performing subject (highlighted in green in Figure 2.9) using Model 7 is

demonstrated in Figure 2.10. The model demonstrates a strong fit on the training data and it generalizes well to the test data with high Pearson Correlation. Confusion matrices are provided to evaluate the performance of choice classification, showing that the probabilities of correctly predicting the actual classes are both above chance for the test data. and above 0.9 for the training data.

Critically, we visualized the distributions of RT data and the distributions of estimated parameters for training and test data across the three experimental conditions of difficulty: low SNR, medium SNR, and high SNR. Low SNR is the hardest condition and is expected to have slow RT, lower drift and higher boundary. Figure 2.11 shows that the distributions of both drift and boundary reflect systematic difference across conditions. The median of distribution of trial-level parameters can better separate out the three conditions compared to RT. This pattern is observed across all subjects in Dataset 2, sometimes even in the cases where RT alone fails to show differences among conditions (see Figures A.2 to A.6). Results for the other three subjects are presented in appendix.

2.4 Discussion

2.5 Conclusion

We proposed and validated the Decision SincNet, a neurocognitive model of decision making that integrates EEG data and the drift diffusion model estimated by a neural network. The end-to-end model uses the likelihood function of Wiener first-passage time to train the model on one second of EEG data given the RT. We have shown that the model can simultaneously predict two cognitive parameters of Drift-Diffusion model that reflect speed and caution during decision making. The model uses interpretable convolutional layers, that

automatically learned bandpass filters, channel weights, and critical time windows which can be inspected to understand EEG relationships to human cognition. Various methods of model evaluation and post-doc analysis of the model have also been discussed. Examples of the model results are reported. To bypass the issue of overfitting boundary parameter, we have developed a follow-up approach to address this issue by additionally forcing the drift parameter to find solutions that correlate with the RT during optimization.

We believe the model performance could be improved by manipulating the Decision SincNet architecture, in order to fit the entire set of DDM parameters, including bias and non-decision time. Future work should also extend the model to incorporate likelihood functions for incorrect trials with a joint behavioral and EEG dataset that has lower accuracy. Finally, other explainable deep learning techniques for EEG signals such as occlusion sensitivity analysis for estimating source signals [20] should be explored.

2.5.1 Application on Generative Modeling

The development of Decision SincNet with WFPT optmized using gradient descent motivated a further collaborative line of work on generative modeling framework with Variational Autoencoders [62]. NCVA was introduced to map EEG and behavioral data into a joint latent space, such that the model can generate task-relevant high-dimensional EEG signals containing corresponding neural features on a trial level from low-dimensional behavioral data, as well as estimating cognitive model parameters from EEG signals. By including Wiener Likelihood in the loss function, the model learns cognitive parameters conditioned on behavioral data, and learns neural latent variables conditioned on cognitive parameters. This joint modeling framework is an innovative line of work in computational cognitive neuroscience, as it allows for trial-level inference and hypothesis testing.

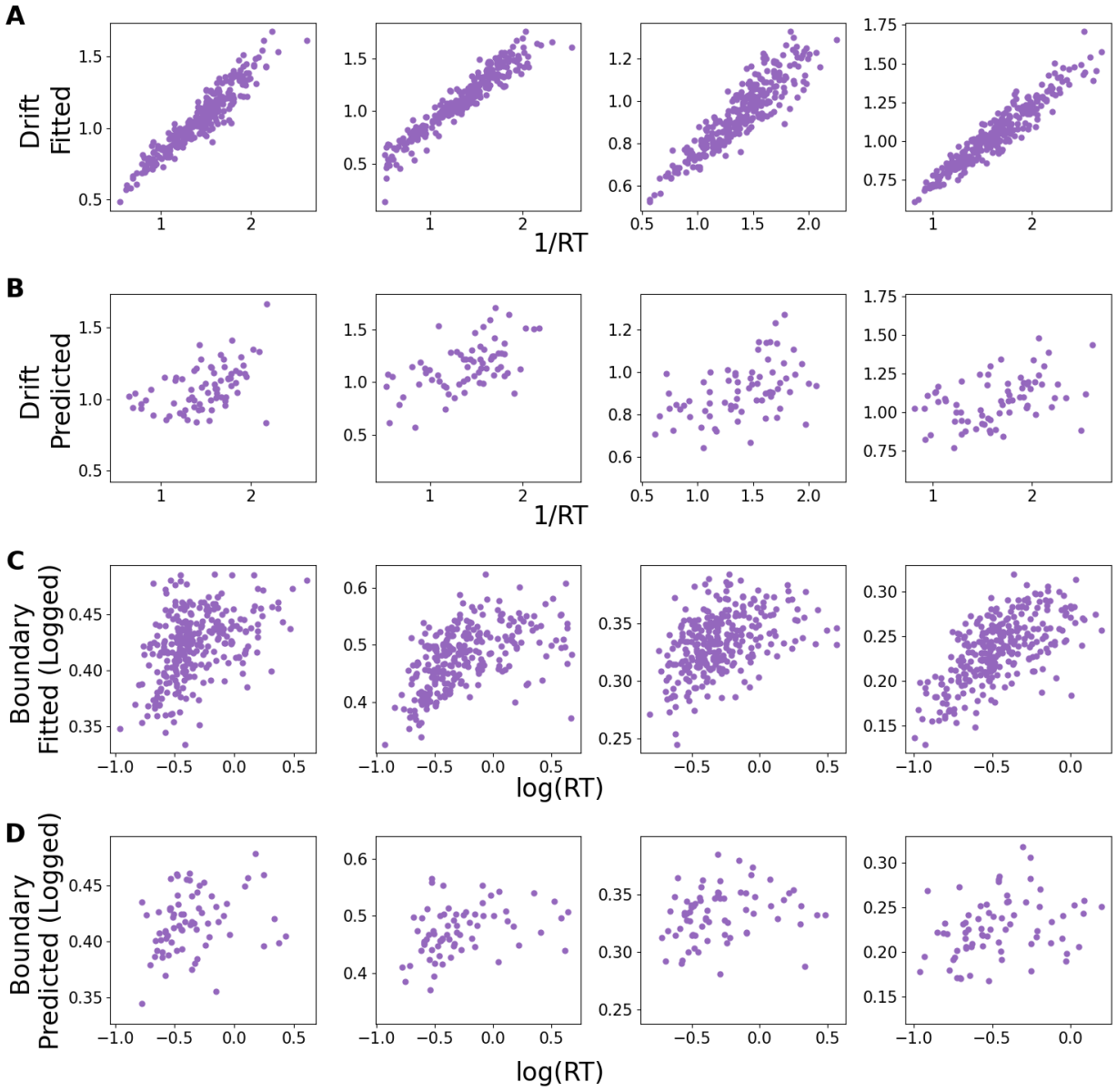


Figure 2.6: Scatterplots results of 4 example subjects. Each column represents the results of one subjects. Panel A shows the scatter plots of predicted drift against 1/RT with the training data. Spearman correlations from left to right are $\rho = 0.931^{***}, 0.963^{***}, 0.900^{***}, 0.940^{***}$. Panel B shows the scatter plots of predicted drift (δ) against 1/RT with the test data. Spearman correlations are $\rho = 0.506^{***}, 0.540^{***}, 0.472^{***}, 0.429^{***}$. Panel C shows the scatter plots of predicted boundary $\log(\alpha)$ against log(RT) with the training data. Spearman correlations are $\rho = 0.506^{***}, 0.426^{***}, 0.469^{***}, 0.631^{***}$. panel D shows the scatter plots of predicted $\log(\alpha)$ against log(RT) with the test data. Spearman correlations are $\rho = 0.285^{*}, 0.391^{***}, 0.272^{*}, 0.336^{**}$. Correct and incorrect trials were both included. * Correlation is significance at the 0.05 level (2-tailed) ** Correlation is significance at the 0.01 level (2-tailed) ***Correlation is significance at the 0.0001 level (2-tailed)

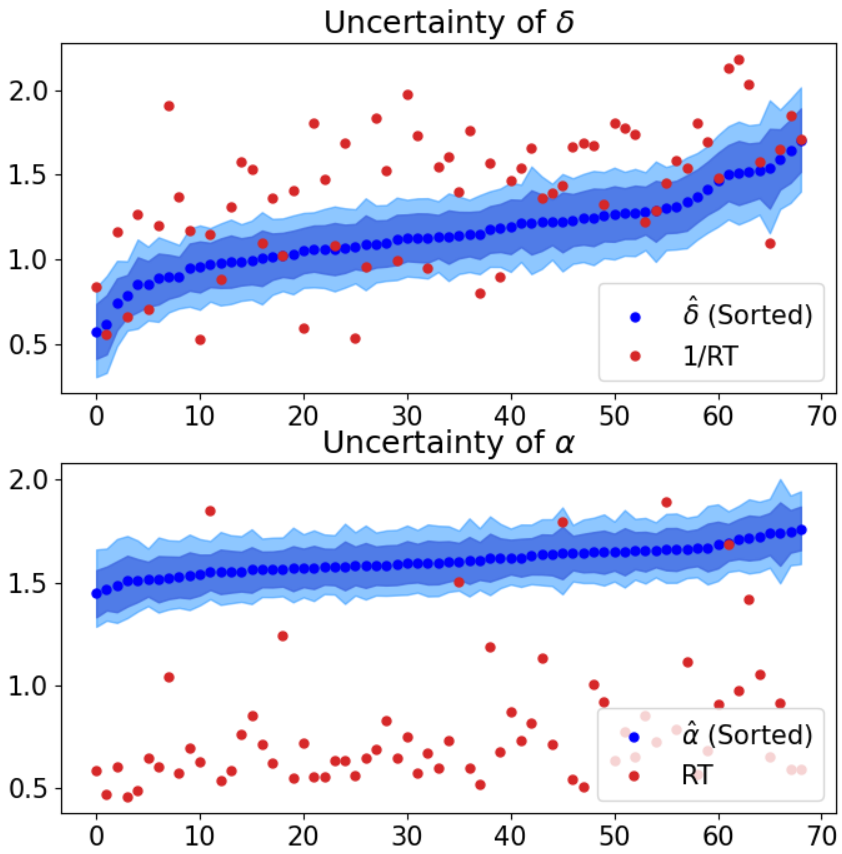


Figure 2.7: Uncertainty of the model prediction obtained from Monte Carlo Dropout. Blue dots in the top figure are median drift rate estimates sorted from small to large. Red dots are the corresponding 1/RT for that trial as proxy of speed. Dark blue area represents the Standard Deviation, and light blue area represents the 90% Credible Interval. Blue dots in the bottom figure are median boundary estimates sorted from small to large. Red dots are the corresponding RT. Prediction were repeated 5000 times with different neurons deactivated.

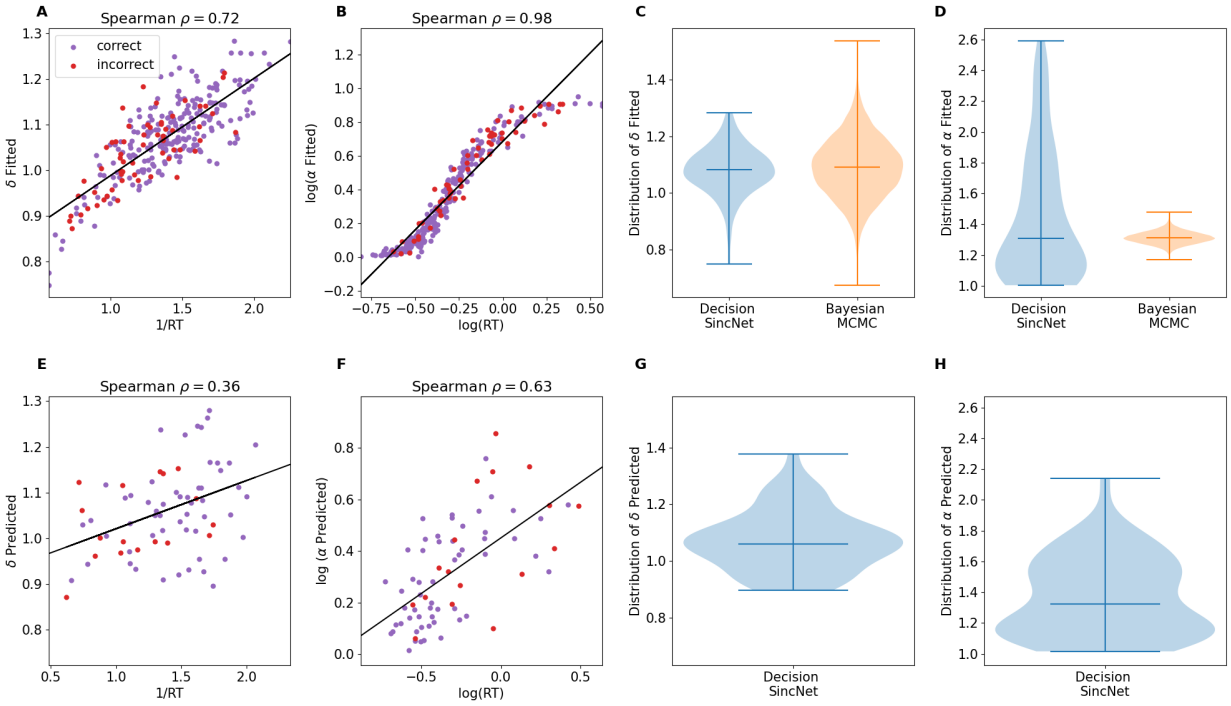


Figure 2.8: Model performance for the same subject, as presented earlier, utilizing the improved loss function.

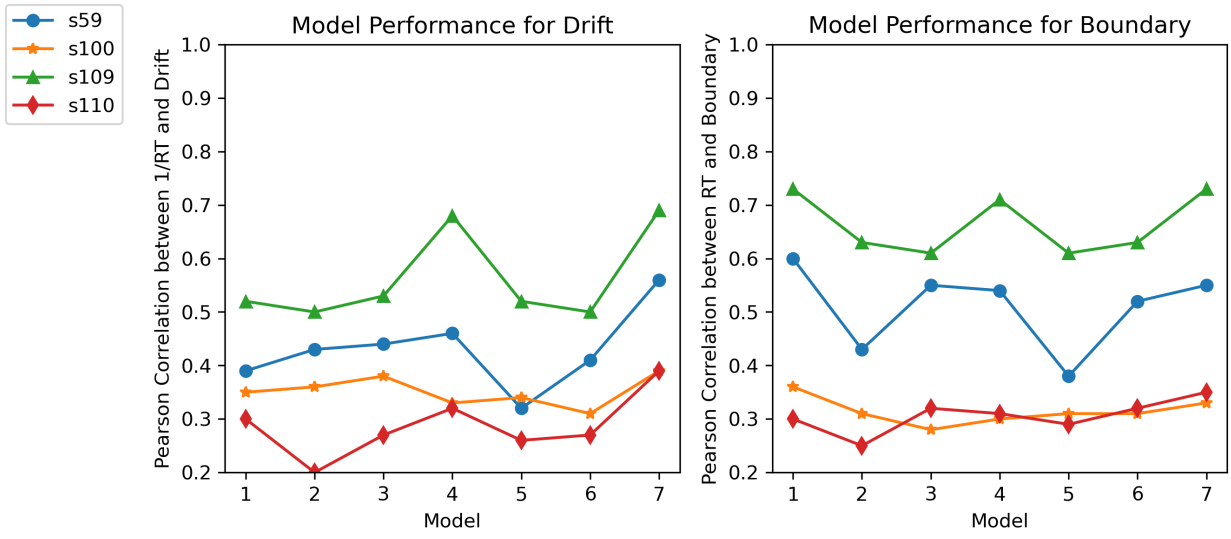


Figure 2.9: Model performance on test data by subjects in Dataset 2 for each model in Table 2.2. X-axis represents choice of model, and y-axis are the Pearson Correlation Coefficients between 1/RT and drift (left) and RT and boundary (right).

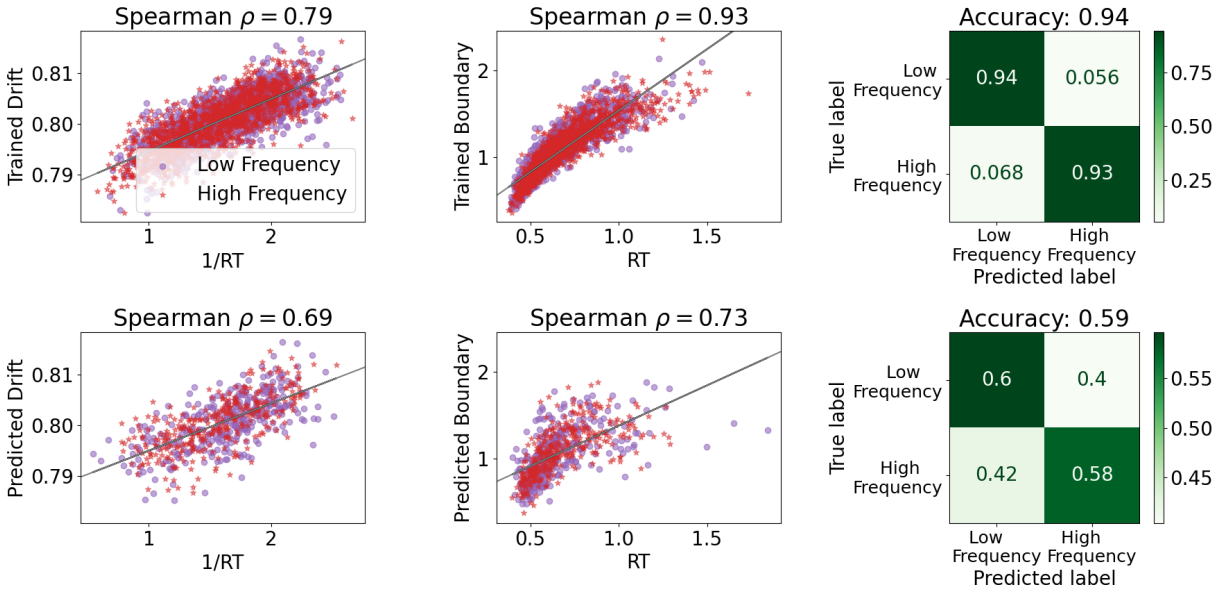


Figure 2.10: Model 7 performance for subject s109 in Dataset 2.

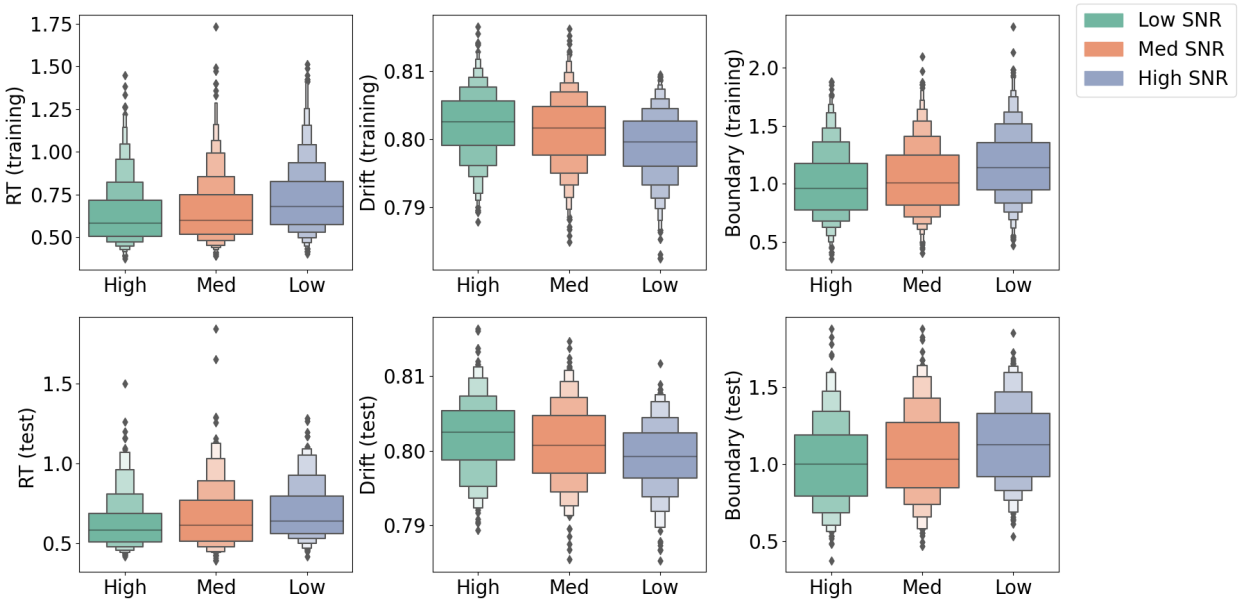


Figure 2.11: Distributions of RT, Drift and Boundary by experimental condition for subject s109 in Dataset 2.

Chapter 3

Interpetability of Decision SincNet

3.1 Introduction

Deep neural networks have become powerful tools in many domains. Increased applications of AI in healthcare and scientific fields have called for special attention to explainable AI (XAI) and the interpretability of deep learning models. The taxonomy and categorization of interpretability have been discussed in detail in numerous papers [70, 37, 47], yet a consensus hasn't been reached, partially due to differences in the end goal of domain applications and what is needed be explained for audiences in that domain. Generally speaking, the approaches can be categorized into intrinsic methods and post-hoc methods. Intrinsic methods focus on designing model structure to satisfy certain interpretability objective(s) in mind. Intrinsic methods, such as attention mechanisms, actively modify the model and/or training process during optimization. Post hoc methods, on the other hand, are used after a model is trained. Common post hoc methods to visualize CNN often involves gradient-based methods [51] or Activation Map such as CAM or Grad-Cam[51, 12, 71, 48]. These methods are often use in interpretable deep learning research in neuroimaging [69, 55]. The evaluation

criterion for interpreting neural network models can be qualitative or quantitative. Qualitative methods focus on providing insights for human understanding and reasonability, and quantitative methods usually involves to what extend can one uses known information to mimic the outcome of blackbox model.

We developed Decision SincNet [54], a specific Neural Network model that can simultaneously predict single trial cognitive parameters from neural signals given behavioral data and extract relevant EEG features without feature engineering. The architecture of Decision SincNet is inspired by Shallow-SincNet desgiend with the domain knowledge of Signal Processing [42]. When we incorporate a well-tested model of behavioral data as the likelihood function for the loss function, we gain the power of single-trial inference on evidence accumulation rate and boundary, as well as the ability to analyze EEG data on a single-trial basis. In Chapter 2, we demonstrated the model's predictive capability using various methods. In this chapter, we aim to understand how the model extracts high-dimensional, noisy single-trial EEG signals to make predictions. Our goal is to gain insight into the relationship between brain and cognitive processes by fully leveraging the intrinsically interpretable structure of SincNet.

In this chapter, we will present methods and qualitative results for interpreting each layer of the Decision SincNet, demonstrating the approaches to discover neural correlates of behavioral patterns using the current modeling framework. We will provide a comprehensive summary of possible methods that can be used through visualization. Interpretability methods include both intrinsic approaches, where we inspect the sinc kernels and attention vectors that were updated in the neural network during the training process, and post-hoc analyses, such as analysis on saliency and activation maps.

3.2 Methods

We will use one subject from Dataset 1 and one subject from Dataset 2 throughout this chapter. We will present results on temporal filters and spatial weights from the two subjects first using a model architecture with shared spatial layer. The spatial layer has a depth dimension of 3, resulting in 3 sets of weights followed by separate linear layers to predict each parameters. To improve the scientific interpretability of the Decision SincNet, we will then use a subject from Dataset 2 to demonstrate another model architecture where neither Sinc layer nor spatial layers was shared. Each parameter is predicted from its own set of frequency bands and 1 set of corresponding weights by removing the depth-wise dimension.

3.2.1 Model Weights as Temporal and Spatial Filters

After training the model, the weights of Decision SincNet at each layer can be inspected to understand how each layer processes neural signals to make the final estimates, a model-specific method given the model design. Figure 3.1 illustrates the process: a channel-by-time EEG sample $\boldsymbol{x}_n \in \mathbb{R}^{C \times N}$ is temporally filtered by the sinc kernels, summed using spatial weights, passed through a ReLU layer, pooled and- flattened into a linear layer to make the final prediction on choice and latent cognitive processes.

Weights as Sinc Functions

Sinc kernels as clear bandpass filters are parameterized by trainable weights representing the cutoff frequencies $f_{1,j}$ and $f_{2,j}$. Figure 3.1 demonstrates the sinc filters in the time domain. Filters in blue are examples of the kernels initialized with central frequency from 1 to 30 with a a width of 2. Each sinc filter has a well-defined response spectra. Filters in orange are examples of sinc functions are training. We eaxmined whether filters learn to move from

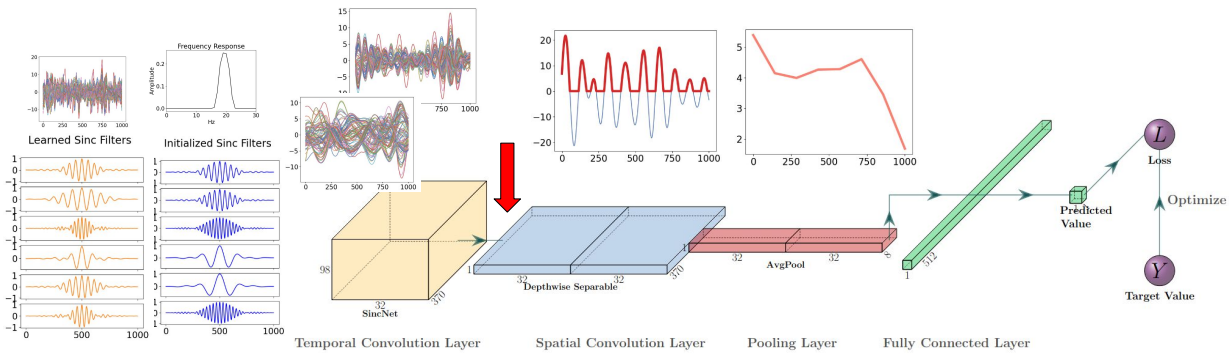


Figure 3.1: Visualization of the input and output of each layer, along with Sinc kernels in a simplified Decision SincNet. The red arrow indicates the insertion of attention module.

their initial random frequency bands, and whether bandwidth ($|f_{1,j} - f_{2,j}$) changed.

Weights from Spatial and Linear Layers

Spatial weights are learnable parameters used to compute weighted sums to obtain different linear combinations of multi-channel time series within the same frequency band into a univariate time series. When the depth parameter of the spatial convolution layer is set to 3, three different sets of weights are learned and applied on the same multivariate time series within a given bandwidth.

When the spatial layer is shared for both drift rate δ and boundary α parameters, the difference that affect the final predictions are the weights learned in the subsequent Fully Connected Layer, representing how the univariate times series pooled into overlapping time windows are weighted across frequency bands. In the Fully Connected Layer, the intercept (bias) is set to zero during training.

The linear weights from the last layer can be obtained as follow. Let $X \in \mathbb{R}^{k \times c \times n}$ represents the time series after the Sinc Convolutional Layer with k trainable sinc filters. The spatial weight can be represented as $W_s \in \mathbb{R}^{k \times d \times c}$ with d as the depth parameter. The output is $X' \in \mathbb{R}^{k \times d \times n}$ for each frequency band. The final output $Y = \mathbf{w}_l \cdot \mathbf{x}''$ where $\mathbf{x}'' \in \mathbb{R}^{n'}$ is a

one-dimensional vector flatten from Pool(Relu(X')). Given the weight vector $\mathbf{w}_l$, we can project back the weights onto the j^{th} frequency band of interest, and average the weights at each time window and over the depth dimension, similar to the idea of Class Activation Mapping [71]. Since the weights are relative to each other, we can interpret the normalized weights as how the univariate time series are weighted over time to each predict drift and boundary.

3.2.2 Gradient-Based Saliency Map

One of the most widely used interpretability methods developed for Convolutional Neural Networks is Saliency Map [51], also referred to as Vanilla Gradient. This approach originated from image classification and shows that one can infer which pixels affect the final output the most by computing the gradient of the output score for the class of interest with respect to the input pixels. This method gives us a map that is the same size as the input features with negative to positive values.

Given a trained model, a forward pass is performed on an input i of interest followed by the calculation of the gradient of an output Y with respect to a feature of interest F as $w = \frac{\delta Y}{\delta F}\big|_{F_i}$. Important features are identified as the ones with the largest normalized gradient. The use of the gradient relies on the approximation of the output by applying a first-order Taylor expansion $Y(F) \approx w^T F + b$.

We use this gradient-based technique to interpret the learned bandpass filters and spatial filters for each subject [6, 51], and extended the above method to further interpret pooled time windows. We obtained the importance of sinc filter by averaging the normalized gradients over channel and time dimension given the Saliecy Map with the size of $k \times c \times n$. For each filter k, we obtained the important channels by averaging over the third dimension. Finally, we averaged the results from all test data to visualize the results. Therefore, for

each subject, we can obtain critical temporal filters and the corresponding, spatial weights and time periods.

3.2.3 Feature Map Visualization

A feature map (also used interchangeably as activation map), is the result of convolving the filters with either the input data or the feature map from the previous layer [33, 71]. Given the design of the Decision SincNet, we will mainly focus on the feature maps given the spatial weights, because the maps following the sinc layer are self-explanatory as the weights are band-pass filers. Within each frequency band, we will visualize on topographic maps to compare signals with or without spatial weights. In addition, we will further visualize the spatial weights as ERPs, calculated by the weighted average of channel-by-time signals. When the model architecture has a depth more than one in the spatial layer, visualizing feature map for different map helps us discover different linear combinations of the signal that were used for predictions.

3.2.4 Attention Block to Rank Sinc Filter Importance

To further improve the scientific interpretability of SincNet, we introduced an *attention block* [19]. The architectrual unit was originally designed to increase channel-wise feature dependency. In Decision SincNet, since the channels are different interpretable sinc filters by design (as opposed to RGB channels or abstract features), the attention block thus allows us to rank the learned frequency bands by importance. Given the weights of each filter, The sinc filters can also be visualized as power spectrum by applying FFT on the weighted sum [50]. Note that because in practice all the filters are normalized as the sum of all kernel samples need to be one, we normalize the spectra by the width of each filter purely for visualization. Given the important frequency bands, activation maps can be used to visualize changes in

predictive EEG patterns over time in the following layers.

The self-attention mechanism was introduced after the temporal convolutional layers as shown in Figure **??** indicated by the red arrow. We inserted a "Squeeze-and-Excitation" computational unit for the sinc filters, such that the model can use global information to recalibrate sinc filters by scaling up the important filters and scaling down the import filters using the input data itself. After the sinc convoluational layer, output is aggregated across time and spatial dimension to a single value by using adaptive pooling. The aggregation is followed by two fully-connected layers followed by a ReLU and a Sigmoid activation function. The weights were reshaped and applied back to the feature maps that would get fed forward to the next block. In the context of Decision SincNet, this unit can be directly thought of increasing the magnitude of signals using a non-linear model of the signal itself. Such computational unit has proven to improve model performance, and importantly we can pass EEG data to the trained model and inspect the weights as a measure to rank the relative importance of temporal filters. Because the attention module learns how to model the input data, the rank of importance can be different depending on the input data.

3.3 Results

3.3.1 Visualization of Gradient-Based Saliency Map at Each Layer

Fig. 3.2 shows an example from the same subject using the method. Three different color demonstrates three different frequency bands that are most critical to the prediction based on their normalized gradient. Absolute values of the spatial filters associated with a specific band-pass filter are then analyzed for the significance of each electrode. Panel A demonstrates the important weights identified. Each subplot shows two sets of important weights for a specific frequency band. This is because two separate spatial filters are applied to each

filtered output from SincNet layer. Panel B demonstrates the important time windows obtained from the saliency map of the pooling layer, labeled by the center frequency of the corresponding critical frequency bands. In the example subject shown here, gamma band (31 Hz, 41 Hz) throughout the decision interval, and beta band (24 Hz) activity around the response were the most important bands. A notable caveat of Saliency Maps as observed in our results is the unstability of feature importance, as the gradients are very sensitive to small changes in model.

3.3.2 Sinc Filters

The weights can be inspected to visualize the frequency bands that were learned. Two examples of interpretable bandpass filters, one from Dataset 1 and one from Dataset 2, are shown in Figures 3.3 and 3.4 respectively. These filters are ranked and sorted based on the attention weights. We can see that the filters change their cutoff frequencies ($f1$, $f2$) and bandwidth $|f2 - f1|$ from their initial values during optimization, so that better sinc filters can be used on the input signals to extract relevant signals. Both results come from models that share spatial layers between two parameters, using more than one depth filter, and split into two heads for predictions before the Fully Connected Layers.

3.3.3 Shared spatial weights

Dataset 1: shared spatial weights and separate linear weights

After identifying the most to least important frequency band, we can obtain the spatial weights and activation map following the sinc convoluational layer over time.

Figure 3.5 shows the four sets of spatial weights corresponding to the four most important

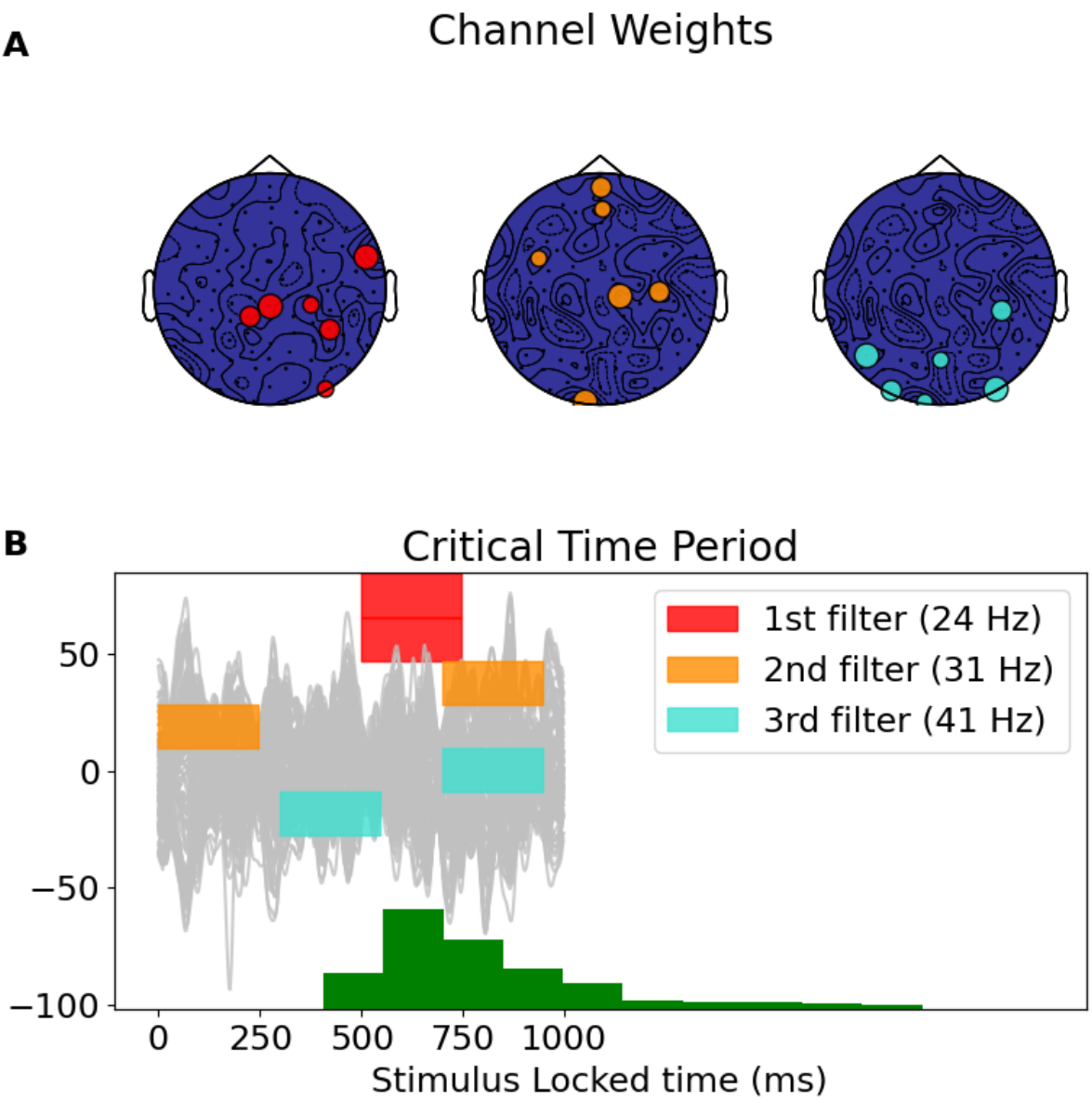


Figure 3.2: Post-hoc analysis of the trained model. Three colors represent the most critical frequency bands obtained from the normalized gradients. The corresponding central frequencies are given in parenthesis. Panel A shows the three most important pairs of weights from the spatial filter. Weights are in pairs due to the depth of 2 in the spatial convolution layer. Panel B shows the three time windows labeled by their corresponding frequencies that are most critical to predictions.

frequency bands (28-31 Hz, 20Hz-30Hz, 11Hz-17Hz, 11-14Hz) using the exact same model from the same subject as in Figure 3.3. Each frequency band has three sets of spatial filters to weight the electrodes, resulting in different linear combinations of the signals going into

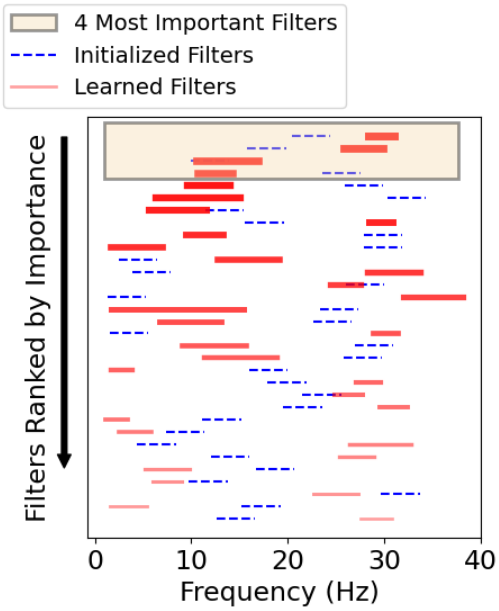


Figure 3.3: Comparison of filters ranked by importance from a subject in Dataset 1. Dotted blue lines represent the random initialization of the bandpass filters before the model has been trained. Red lines represent the bandpass filters after the model has been trained.

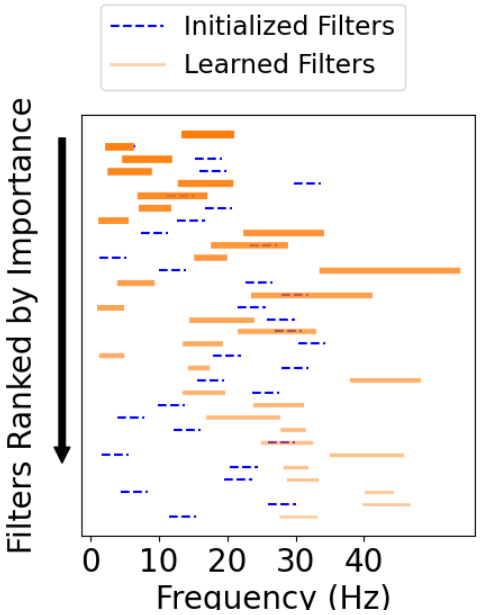


Figure 3.4: Comparison of filters ranked by importance from a subject in Dataset 2. Orange lines represent the bandpass filters after the model has been trained.

the next layer. We calculated a weighted sum of the spatial weights for each frequency to visualize the overall distribution. Results from the first and second row, as well as results from the third and fourth row, demonstrated that frequencies that are closer tend to have similar electrode weight patterns.

We can also visualize the activation map of the spatial layer and the weights from the last Fully Connected Layer . Figure 3.6 illustrates how signals with a 28Hz-31Hz bandpass filter (identified as the most important band) in Beta Band are weighted by 3 different sets of spatial filters over time window that were pooled. The weighted signals differ from the unweighted signal such that it focuses on more localized activities.

Fig 3.7 shows how the spatially filtered, and temporally pooled signals from the most important frequency band (28-31Hz) are weighted differently in the time domain to separately

make the final predictions for drift and boundary parameters.

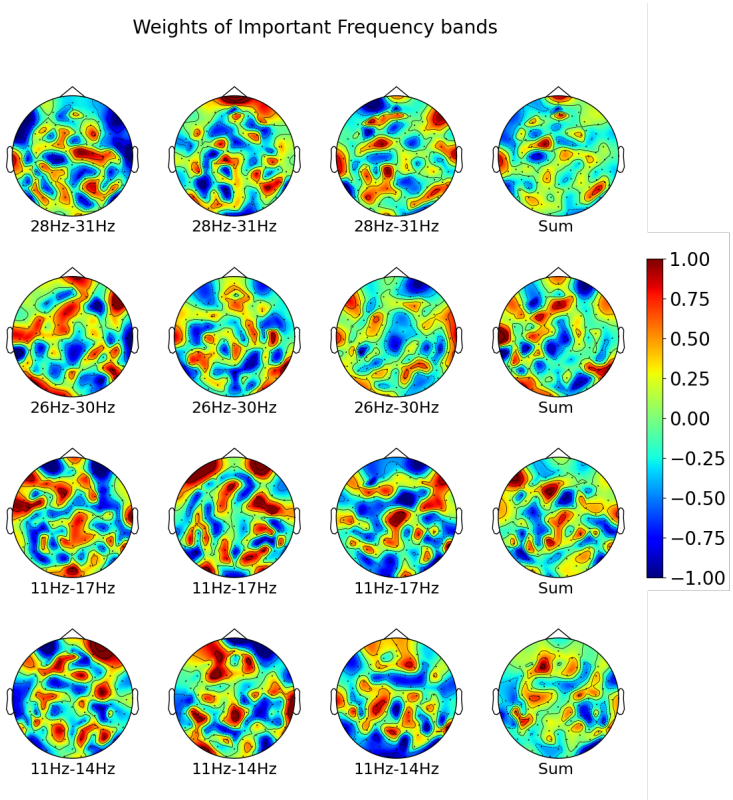


Figure 3.5: Sets of the Spatial weights from subject in Dataset 1 from the 4 most important temporal filters.

Dataset 2: Feature Maps as ERPs

Figures 3.9 and 3.10 show the ERPs using the two most important frequency bands (14Hz-20Hz, 3Hz-6Hz) for the example subject from Dataset 2, using the same exact model as in Figure 3.4. Figure 3.9 demonstrates low beta waves slowly suppressed prior to response, and Figure 3.10 demonstrates theta waves with a clear P200 signal.

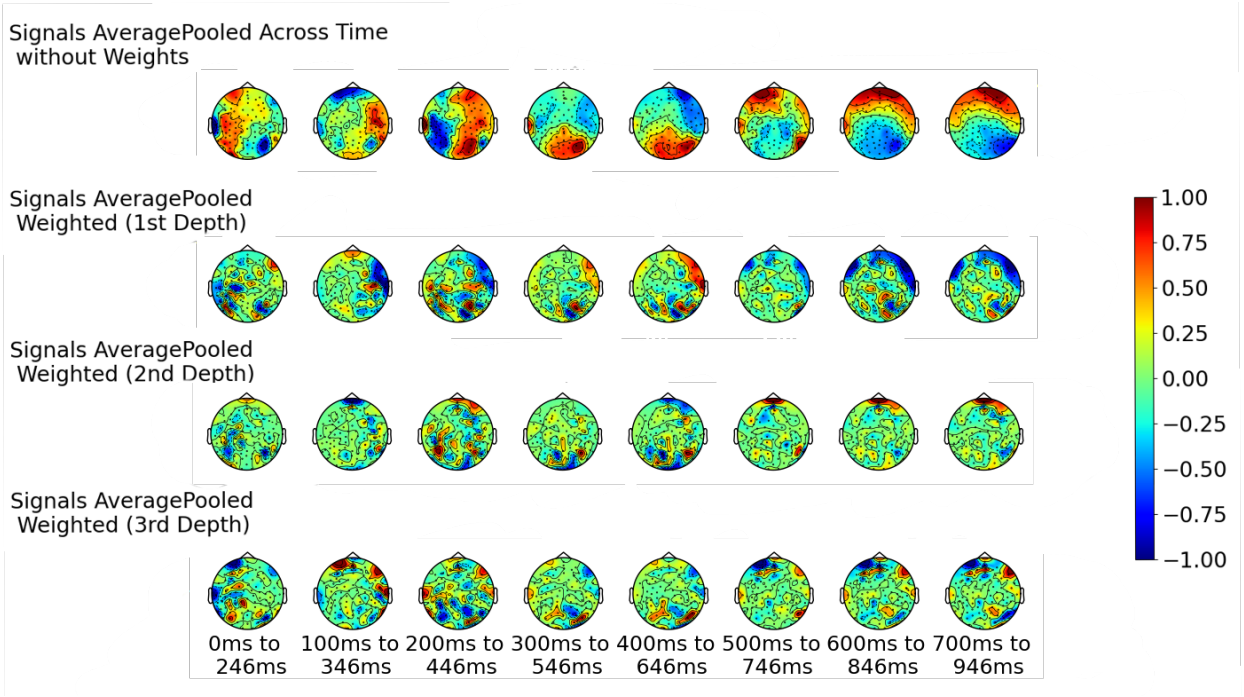


Figure 3.6: Topographic maps on 28-31Hz signal with and without spatial weights. Top row is the unweighted signal. The second, third and fourth row are the signals weighted by different sets of spatial weights.

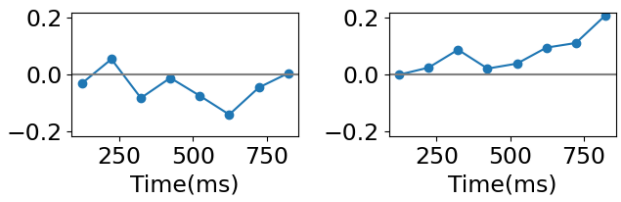


Figure 3.7: Weights from Fully Connected Layer averaged over depth dimension for 28-31Hz signal. Left: weights for predicting drift parameter. Right: weights for predicting boundary parameter.

Figure 3.8: Example subject from Dataset 1.

3.3.4 Seperate temperal and spatial features for each parameter

To further improve scientific interoperability, we present visualization from a model architecture where sinc convolutional layer and spatial convolutionall layer are both trained separately for each parameter. When the depth parameter of the spatial convolutional layer was set to 1, only 1 set of spatial weights were learned, so that we can visualize a single

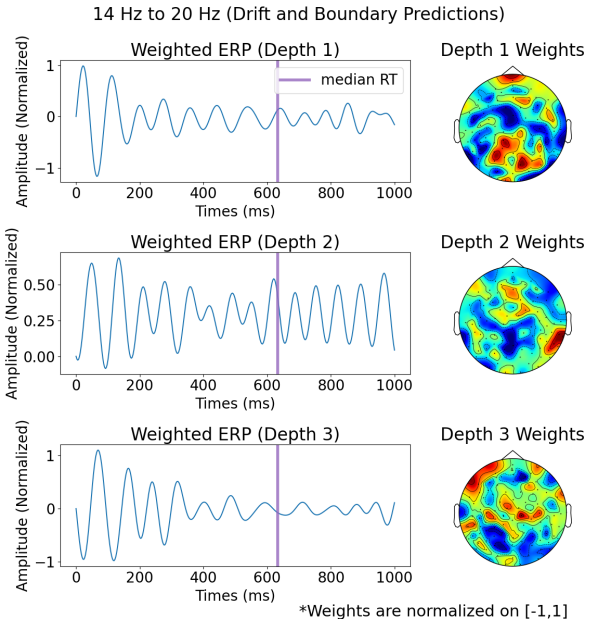


Figure 3.9: ERPs and spatial weights under the most important sinc filter (14Hz-20Hz)

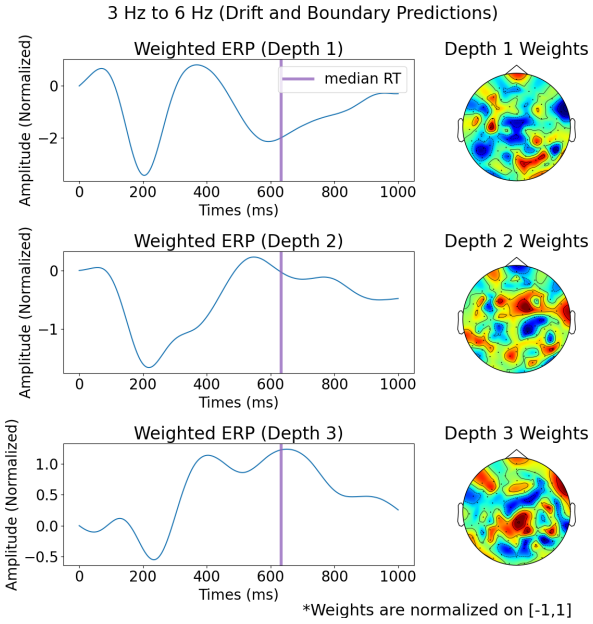


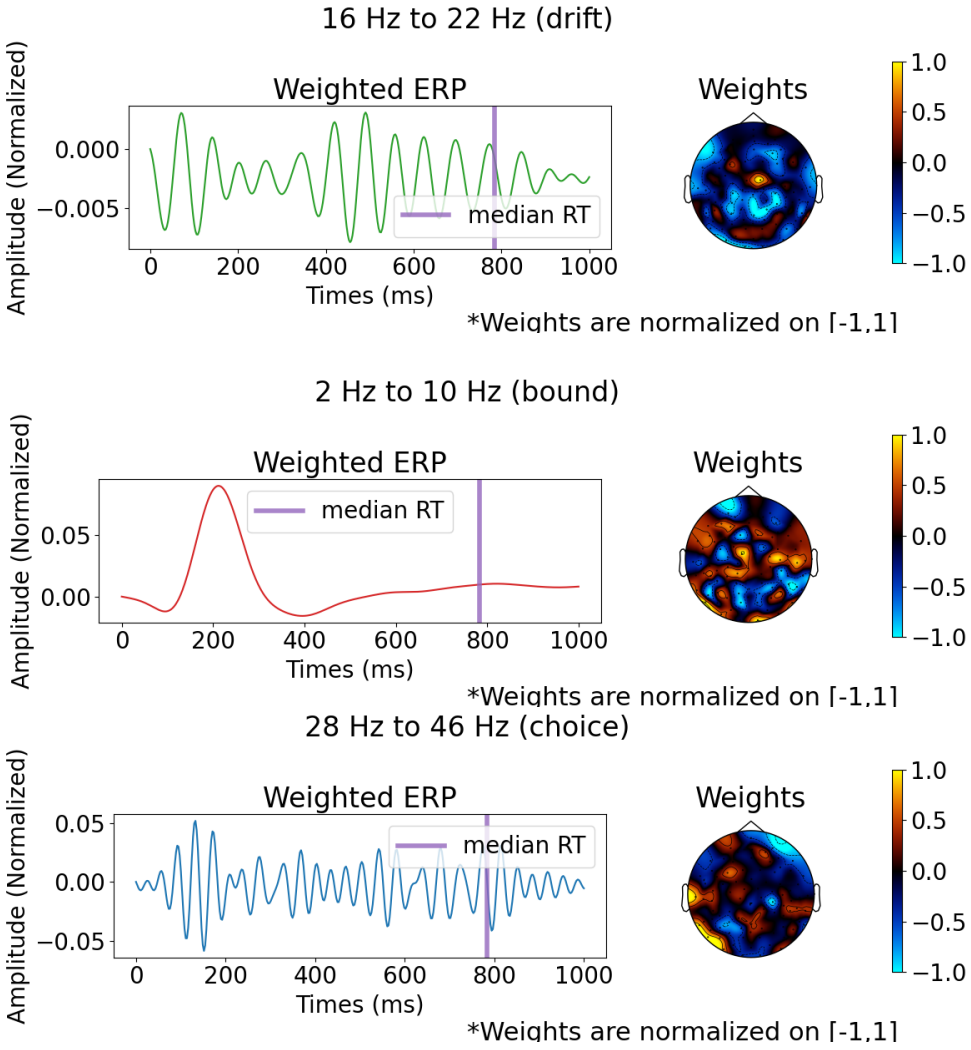
Figure 3.10: ERPs and spatial weights under the second most important sinc filter (14Hz-20Hz)

Figure 3.11: Example subject from Dataset 2 using a model where spatial weights are shared. ERPs obtained from spatial convolution weights (shown in Topographic maps) using the two most important filters used to predict DDM parameters.

waveform that can best capture the signals for each parameter within each frequency. The model has an additional objective to predict the observed choice of each trial. Therefore, the full model architecture has three separate sets of sinc filters and spatial weights for drift, bound, and choice predictions.

Figure 3.12 demonstrate the ERPs under the most important frequency band for each predictions. For this subject from Dataset 2, low beta activities were identified to be the most relevant signals. Theta and low Alpha activities were most useful for boundary prediction . High Beta and Gamma band activities were most useful for choice prediction.

Figure 3.13 shows spectra of the weighted sum of sinc kernels for each predictions. The filter learns to focus below 40Hz, suggesting that the model did not focus on noise. Figure 3.14 demonstrates the spectra of the kernels normalized by the bandwidth. No overwhelmingly



(a) ERP for predicting Choice.

Figure 3.12: Example subject from Dataset 2 using a model where neither sinc filters nor spatial weights is shared. Spatial layer only learns one set of weights. Top, middle and bottom figure each represents signals for predicting Drift, Boundary and Choice. ERPs were obtained from spatial convolution weights of the most important filters identified by the Attention vector.

distinctions can be found, but some patterns are observed across subjects. We can see weak effects showing that sinc filters for boundary tend to have filters in lower frequency bands up to 10Hz. Filters for Drift Rate tend to focus on high Alpha and Beta Band (10-25Hz), as well as 30-40Hz at which the SSVEP stimulus were flickered. Across all four subjects, we can consistently observe the patterns such that boundary filters focus on lower frequency bands, while drift filters concentrate on the beta band above 13Hz. The results for the other

subjects are included in appendix.

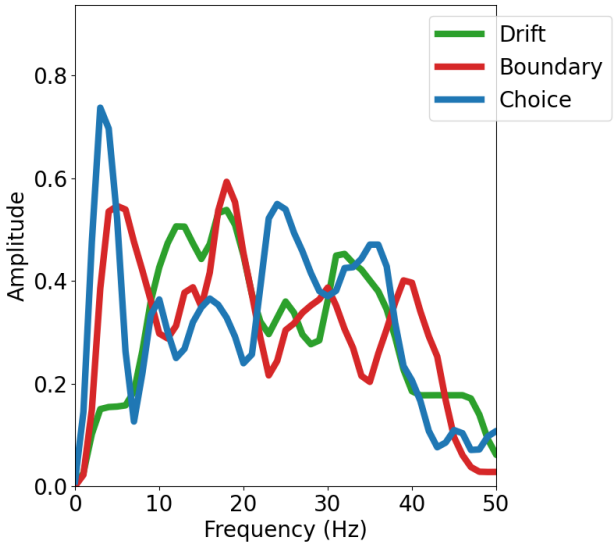


Figure 3.13: FFT results of weighted summ of sinc filters.

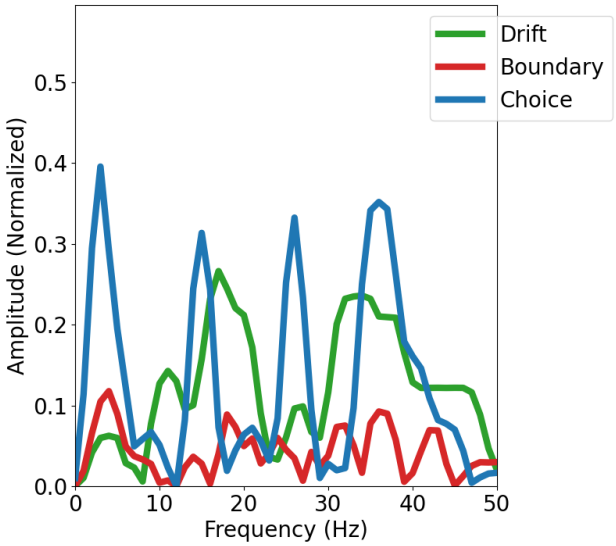


Figure 3.14: FFT results of weighted summ of sinc filters normalized by bandwidth for visualization.

;

Figure 3.15: Example of the FFT Spectrum obtained from the Weighted Sum of Sinc Filters for subject from Dataset 2.

3.4 Discussion

Visualization of Decision SincNet utilizing various tools from neural network interoperability on two different datset helped us better understand how EEG features are extracted. Incorporating attention module enabled us to examine what features are more relevant to estimate latent parameters of cognitive processes.

We demonstrated that the sinc filters learned to move from the initial values. The spectra of the learned filters showed important frequency bands for DDM parameters. Across subjects, boundary tend to be predicted from lower frequency bands (up to 10Hz), and drift rate tend to be predicted from high Apha and Beta Band. The association between Beta band activities

and evidence accumulation rate is consistent with previous research [11, 67]. Feature maps of spatial filters with depth dimension show how multi-channel EEG signals are linearly combined seperately to predict DDM pararmeters. When only one set of temporal and spatial weights are allowed for each parameter, the model learns to identify different important brain dynamics in time and frequency domain.

We presented results by computing and visualizing the gradients as saliency maps. Even though feature importance can be obtained, we found that the results are unstable. Therefore, it is hard to draw consistent and reliable scientific conclusions using this method. This finding is consistent with previous research calling for issues with repeatability and interpretability using saliency map [22, 46], especially in medical imaging [3].

An important limitation of these results is the qualitative interpretation of the features. There is no uniform statistical methods performed to quantify the features that were identified to be relevant. Future work should try to validate the relationships between neural signals and cognitive processes. For example, one can use the features extracted by SincNet to fit the Neurognitive Hierarchical Model (HDDM), and see if the model performs better than the existing method, where single-trial EEG features were extracted a priori [30].

Chapter 4

Directly Testing Sequential Sampling Models with Evidence Chains

4.1 Introduction

4.1.1 Different Forms of Stochastic Models in Perceptual Decision Making

Perceptual decision making is a critical cognitive process in everyday life. Extensive experimental research has documented patterns of behavior (accuracy and response time) with manipulations of stimulus properties [64] and response contingencies [32] in laboratory tasks using forced-choice designs. This wealth of behavioral data has motivated the development of models of computational mechanisms of perceptual decision making. The dominant approach has focused on sequential sampling models[40] that assume that people make rapid decisions by gradually accumulating noisy evidence based on a sequence of observations until a threshold quantity of evidence (or relative evidence) triggers a response. Variants of these

models include Drift-Diffusion Models (DDM [41]), race models [4, 23] , or Leaky Competing Accumulator (LCA) models[59], all prescribe that a decision is made based on integrated evidence, at a time when a threshold is crossed. One alternative approach, response urgency [9], allows for an independent mechanism for the timing of the response, while the choice is based on the properties of either instantaneous or integrated the evidence. We will briefly review sequential sampling models, and discuss the assumptions and questions we are interested in investigating.

Random Walk Models:

One of the most popular class of models in sequential sampling for 2AFC is random walk model[52, 24, 25], representing the idea of computing the differences in signals associated with each alternative.Some evidence [49] suggests that the brain tracks a Decision Variable (DV) accumulating evidence like a random walk, and a decision is generated when the integrated value hits a certain boundary. The purest form of this process is an unbiased Simple Random Walk if we use discrete time steps. An unbiased Simple Random Walk takes either one step to the right or one step to the left with an equal probability on a 1-D lattice. The location at each moment is thus the difference in number of right steps and number of left steps. When the number of time steps are large, the Simple Random Walk can be approximated by a random walk with Gaussian steps where the steps are samples from continuous normal.

Brownian Motion or Wiener process (the simplest form of diffusion process) is the limiting distribution of scaled Gaussian Random Walk. In a Drift Diffusion Model [5], the change of evidence supporting one choice is consisted of a Wiener Process term with mean 0, a variance $\varsigma^2 dt$, and a constant drift term δ as in Equation 2.1. A non-zero drift rate can be represented in a biased Simple Random Walk if the probability of stepping to one direction is higher than stepping to the other direction. Numerically, if a simple random X_n takes +1 with a probability of p and -1 with a probability of $1 - p$ at each step, the theoretical

constant drift rate is essentially the expectation of each step X_i during the random walk ($S_n = \sum X_i$). The expectation is calculated as $\mathbb{E}(X_i) = (2p - 1)$,

The most widely used stopping rule under the random walk procedure is a fixed threshold. Evidence accumulates during diffusion process until a threshold is met, triggering an overt response to terminate the process. A fixed threshold is an optimal decision maker and can be shown in Sequential Probability Ratio Test (SPRT) [65], an earlier method in sequential analysis to test among hypothesis h_1 and h_2 (see Equation 1.2).

The Likelihood ratio keeps being updated and it's calculated by the running sum of weights obtained by the logLR of probabilities towards $h1$ or $h2$ as each piece of evidence comes in. Samples accumulate until the sum reaches a positive or negative criterion (the decision bound Z_2 and Z_1) at time n:

$$\log Z_2 < \log \frac{P(e_1 \mid h_1)}{P(e_1 \mid h_2)} + ... + \log \frac{P(e_n \mid h_1)}{P(e_n \mid h_2)} \quad < \log Z_1. \tag{4.1}$$

It can be shown that the SPRT can be mathematically equivalent to random walk models, and it will converge on the DDM, when the discrete logLR scaled to a continuous time-dependent variable [5]. Previous work has shown that a fixed threshold is optimal as it requires the least samples to achieve a given error [65, 5]. A natural question to ask is: are humans optimal decision makers? Recent experiments have started considering variable bounds, where boundary changes as a function of time [63, 26]. Therefore, one of the objectives of the current experiment is to examine whether boundary is a fixed threshold across trials, or whether boundary changes within a trial as a function of time.

The O-U (Ornstein–Uhlenbeck) model is similar to the DDM (see Equation 2.1) except that the drift rate is now a linear function of value of the accumulated evidence from stimulus δ

and a decay rate λ :

$$dX_t = (\lambda x + \delta)\, dt + \varsigma dW_t,\ x(0). \tag{4.2}$$

The rate of change in x depends on its current value, and the value of λ will decide how much it accelerate or decelerate toward either the threshold. When $\lambda = 0$, the model is equivalent to a DDM. When $\lambda < 0$, dx will be closer to the fixed point as a function of the current state, while when $\lambda > 0$, dx will move further away from the fixed point. Experimentally, $\lambda < 0$ produces a recency effect as the later inputs will be weighted more during evidence accumulation, and $\lambda > 0$, produces a primacy effect as the earlier samples will be be weighted more [7].

The O-U process can be thought of as a weighted random walk in discrete time, and the state of evidence at each step is the weighted sum of integrated evidence. The current experiment will examine whether O-U process can better represent the data, and will inspect the weights throughout the course of evidence accumulation.

Separate Counters: Race Model and Leaky Integrator

Race Models (also referred to as Recruitment Models [23]) , is similar to Random Walk models as evidence is sequentially integrated, but the difference is that there exists a separate counter for each choice, and evidence supporting the two choices is accumulated independently to fixed thresholds. The model is referred to as accumulator model [4, 61] when time is discrete, and Poisson counter model [56] when time is continuous.

Leaky Competing Accumulator (LCA) [59] is a neurally-inspired model. LCA also uses two separate counters, and has two separate diffusion processes. Each process consists of a decay term and a mutual inhibition term. A strong inhibition would behave similarly to diffusion

or O-U models.

Urgency Gating and Runs Model

Urgency Gating Model is different from integration-to-threshold model mainly in two regards: moment-by-moment evidence is rapidly leaking while being integrated like a low-pass filter, and the samples are multiplied by an evidence-independent "urgency" signal that grows with time [9].

Consecutive Runs Model was once briefly mentioned in the history of sequential sampling [4], but was not further pursued. Runs model assumes that there is a counter of consecutive samples supporting the same choice.

In all of these models both internal neural noise and variable states of the physical world contribute to the stochastic process of decision making. Such assumptions have inspired much development in fitting various sequential sampling models to response time and choice probabilities during decision making tasks [38]. Complementing this approach are neuroscience studies that identify signals that carry the property of evidence accumulation to a bound.

However, one of the current challenges in the field is that albeit many variations of models, these computations are difficult to assess because how and when the noisy evidence are integrated can not be validated on *each trial* in a quantitative manner. One can not directly quantify how well a model describes the hypothesized evidence accumulation process on *each trial*. Instead, one can only indirectly infer the assumptions by assessing how well the model predictions fit the patterns of behavioral data in an aggregated manner.

4.1.2 Evidence Chain Paradigm to test sequential sampling models

In this study, we introduced the probabilisitic Pereceptual Decision Making (pPDM) task, a paradigm to present stimuli as a succession of random samples, allowing us to experimentally control the stochasticity during decision process, while minimizing momentary internal noise.

The probabilistic Perceptual Decision Making (pPDM) task aims to examine how people accumulate evidence under the sequential sampling framework by presenting one bit of information at a time until a decision is made. Subjects were asked to perform a two-alternative forced choice (2AFC) task based on the sequence of stimuli presented. The stimuli on the screen will alternate between one of the two choices A or B from a Bernoulli trial. The trial will be drawn from a distribution either with $p_A > p_B$ or $p_A < p_B$ producing a sequence of samples that potentially can be integrated into a chain of evidence favoring one of the choices. Subjects were instructed to respond by determining which choice is more probable on that trial.

Externalizing the sequential samples in an *evidence chain* will enable us to explicitly test the fundamental questions of *whether* and *how* the brain accumulates evidence over time to make decisions. Given a sequence of Bernoulli samples, we can quantify the stream of information as ground truth at termination in three aspects: the number of samples, the amount of accumulated evidence at any given point, and the sequence in which evidence is presented. In the continuous space, the number of samples would conceptually represent the course of time Δt. The current experiment provides the unique opportunity to 1) explore the function(s) to represent the amount of evidence at any given time point to best fit the empirical data, and 2) test whether such representation changes over time, varies across trials, or is associated with other factors.

4.2 Methods

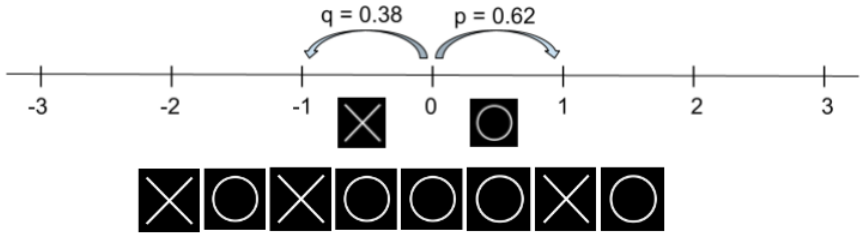
4.2.1 Experimental Design

We ran a two-alternative forced choice (2AFC) task as follows. For each experimental trial, we generated a stochastic sequence $\{X_n\}$ $n \in [0, 30]$ where $X_0 = 0$ and $X_n \sim Bern(p)$ where $X_n = \pm 1$, with $\mathbb{P}(X_n = 1) = p$ and $\mathbb{P}(X_n = 1) = 1 - p = q$. Within a sequence, We fixed $p = 0.62$ to make the Bernoulli trials biased. To visualize the stream of information as units of evidence, let $+1$ represent displaying one stimulus (on one trial, stimulus "O"), and let -1 represent displaying the other stimulus (on the same trial, stimulus "X"). Such a sequence of evidence was labeled "O" dominant; there were an equal number of "O" and "X" dominant trials presented to the observers. Mathematically, such sequence generated can be considered as a random walk on the integer line starting from 0 and moves either $+1$ or -1 at each step. Figure 4.1a shows an example of the random walk with a probabilities of $p = 0.62$ moving $+1$ and $q = 0.38$ moving -1. Figure 4.1b shows an example of the beginning of a sequence for a given trial.

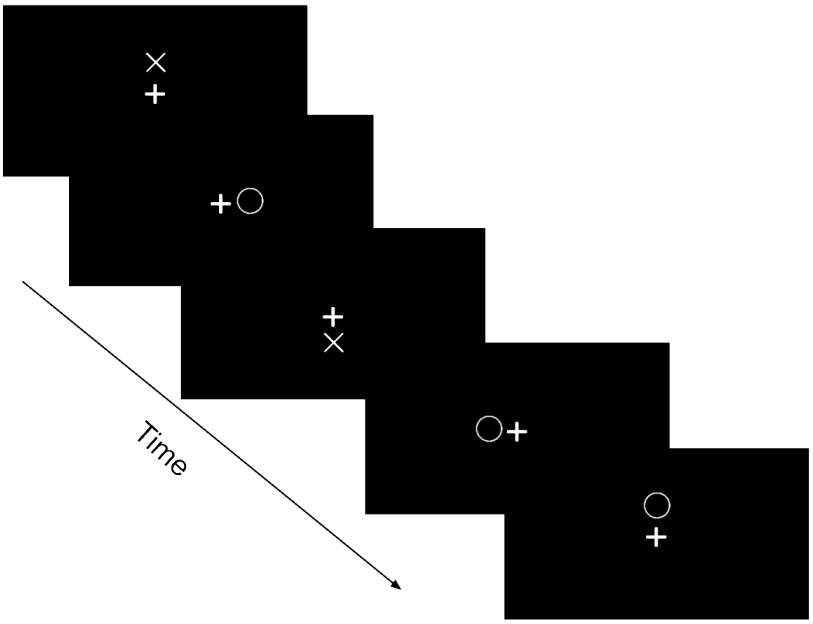
At the beginning of each experimental session, human subjects were instructed to determine whether the chain is "X" dominant or "O" dominant as the sequence is unfolding after a non fixation point ranging from 400ms to 800ms. with a nondeterministic duration. Subjects were asked to respond by pressing a keypad with two buttons using their index fingers. The left button is associated with choice "X" and the right button is associated with choice "O" throughout the experiments. Once a decision is made and either button was pressed, the sequence terminated without presenting the rest of the samples, and the following trial would begin after a brief mask for 1s.

We define each display a sample, and a maximum of 30 samples will be presented for each trial. Each Bernoulli sequence will end if a response is made, or it will end after all 30 samples

are displayed. The following experimental trial would only begin if a response is collected. To counterbalance the stimulus, half of the trial sequences are drawn have $\mathbb{P}(X_n = 1) = 0.62$, where stimulus "O" is more likely to be presented, and half of the trial sequences are drawn from $\mathbb{P}(X_n = -1) = 0.62$ where stimulus "X" is more likely to be presented.



(a) One-Dimensional Random Walk



(b) Example of Stimulus.

Figure 4.1: Demonstration of a 1-D Random Walk with the stimuli "X" and "O" (a) Random walk with a probability of $p = 0.62$ to move $+1$ (representing an "O" stimulus) and a probability of $q = 1 - p$ to move -1 (representing an "O" stimulus) at each stepand a sequence of stimuli presented at the beginning of a trial during pPDM (b). Each display from the sequence shown in (b) is generated from a 1-D random walk.

We introduced two experimental conditions by manipulating the duration of samples (inter-stimulus interval; ISI) . The shorter duration condition displayed each sample for 100ms

and the longer duration condition displayed each sample for 250ms. With the 250ms ISI condition, we want to ensure that the stochasticity from the stimulus samples would outweigh any contributions from internal noise. With then 100ms ISI, we expect that the processing of the individual samples will overlap and we will begin to observe internal noise effects, quantified by lower accuracy and more behavioral variability.

Properties of the evidence chain as decision criterion

Different properties of the evidence chains can be calculated as a trial-level measure of evidence accumulation process. For example for a Drift-Diffusion model, boundary can be obtained from the integrated evidence upon response, or some fixed period prior to response to account for non decision time, shown in Fig 4.4 indicated by the purple line. To obtain the rate of evidence accumulation, a measure of δ, we can estimate the slope m of the line fitted to the starting point and end position of the random walk. For walks that have the same boundary and slope but different walk paths, ς can be obtained to calculate the variance of walk: $\sum(S_t - m)^2$, where S_t is the walk at time t and m is the slope. For a Consecutive Runs Model, we can obtain the number of consecutive runs using the outcome from Bernoulli trials.

4.2.2 Behavior and EEG Data Acquisition

Upon arrival, each subject completed a training block containing sequences generated from the same Bernoulli distribution, such that their performance would converge when the real experimental blocks begin. Each experimental block consists of 50 trials followed by a break. A total of 20 blocks were introduced to each subject to complete across 2 sessions. The duration of each sample stimuli display was systematically manipulated across blocks with an ISI (inter-stimulus interval) of 50ms, 100ms, 250ms and 500ms. There were 2 blocks

of 50ms trials, 2 blocks of 500ms trials, 8 blocks of 100ms trials and 8 blocks of 250ms trials. The ISI were chosen as they can be potentially tracked in in the brain using EEG as frequencies at 20Hz, 10Hz, 4Hz, and 2Hz. The 50ms and 500ms blocks were designed as a pilot measure, and thus were not included for the following analyses. For behavioral data analysis, we backtracked 1 sample for the 250ms condition (-250ms) and 2 samples for the 100ms condition (-200ms) to account for motor execution time.

To test how the brain uses sensory evidence to make a decision, we simultaneous collect EEG recording while subjects complete the task after the training session. Compumedics Neuroscan128-channel Quik-Cap were used. All electrodes were placed according to the International 10-20 electrode placement standard. The EEG data has not been analyzed for this chapter.

4.2.3 Evaluation Framework

Notation and Terminology

The stimulus a subject was presented on a given trial depends on when a subject makes a decision to terminate a given trial. We consider the dataset we collected for each display duration condition $\mathcal{D}$ to consist of N time series. Elements of $\mathcal{D}$ are tuples, i.e., $\mathcal{D} := \{(X_1, z_1, y_1), \ldots, (X_N, z_N, y_N)\}$ where X_i or $X^{(i)}$ denotes the i^{th} stochastic sequence $X_i \in \{-1, 1\}$, $z_i \in \{1, 2, ..., 30\}$ denotes the index position at which a decision is made to terminate the sequence, and $y_i \in \{-1, 1\}$ denotes the choice of the response. The i^{th} sequence starting at index 1 and ending at index at z_i can thus be represented as $X^{(i)}_{1:z} := \{X^{(i)}_1, ..., X^{(i)}_z\}$. We define each observation $X^{(i)}_j$ as a sample representing a unit evidence displayed ("O" or "X").

For general data analyses regarding the entire chain in 4.3.1 and 4.3.2 , we recoded each

sample so that a positive sample +1 represents a sample towards the higher probability stimulus. We define task accuracy as whether the final choice y_i is consistent with evidence S_i. A trial is correct if $\text{sign}(S_i) = \text{sign}(y_i)$ and incorrect if $\text{sign}(S_i) \neq \text{sign}(y_i)$. Note that $S_i = 0$ is considered a correct trial as there is equal amount of evidence towards both choices. For example, if a subject decides that a sequence is dominated by "X" after seeing three "X" stimulus and two "O" stimulus, a trial is considered to be correct with $S_i = 1$, indicating 1 unit evidence towards the choice response. A negative value of evidence $Sn < 0$ would indicate an incorrect trial, as there is evidence away from the choice response. We define evidence as the difference between positive and negative samples. Mathematically, evidence is the sum of all samples seen upon termination $S_i = \sum_{j=1}^{z_i} X_j^{(i)}$.

Problem Formation

To examine how the evolution of the stochastic process gives rise to a decision at sample position $\mathcal{Z}$, we propose a classification framework and formally define the problem as follows:

Given a stochastic sequence of i.i.d. rv $\{X_j\}$, a random variable T taking values in $\{1, 2, ..., 30\}$ is considered to be a stopping time unitized by sample with respect to $\{X_j\}$. If for each $j \in \{1, 2, ..., z\}$ there exists a decision function $G^z : \mathbb{R}^z \to \{0, 1\}$, such that:

$$1_{\{T=z\}} = G^z \left(X_0, X_1, ..., X_z\right), \tag{4.3}$$

G_z can be thought of a function which takes the values of the process $X_{1:z}^{(i)}$ observed up to the present point and outputs 0 if process proceeds and 1 if the process is terminated. Our objective is to find out the G^z decision function that can best describe human behavior. G^z can be linear models such as sum or weighted sum of samples X_j, or it can be non-linear models that capture different aspects of X_i.

As the length of the sequences X_j vary, we created datasets that has equal length at each sample position, and use a classification framework to evaluate how well each decision function G^z fit the human behavioral data at different sample length. This is similar to the idea of asking the model to complete human decision making tasks, but instead of seeing how accurate the task performance is, the end goal is to see how well a model can mimic human behavior. In the next subsections, we will discuss the creation of datasets and evaluation framework in detail.

Dataset Creation

We created sub-datasets $\mathcal{D}_k$ for $k = 3, 4, ..., 15$ to only include trials that terminated at each sample position $\{(X^{(i)}_{1:z}, m^{(i)}) \mid X^{(i)} \in \mathcal{D}, z_i = r\}$, and trials that went beyond the sample position $\{(X^{(i)}_{1:z}, m^{(i)}) \mid X^{i} \in \mathcal{D}, z_i > r\}$. We used the labels $m_i \in \{0, 1\}$ to denote the whether the trial terminated ($m^{(i)} = 1$) or continued ($m^{(i)} = 0$). The number of trials available for each sub-dataset $\mathcal{D}_k$ is visualized in Figure 4.5D and E.

We also created new dataset $\mathcal{D}'$ resampled from original Dataset $\mathcal{D}$. By applying a sliding window of size 4 with a step size of 1 on all the sequences, we created input vectors $\{X_j, X_{j+1}, X_{j+2}, X_{j+3}\}$ taking values $+1$ and -1. There is a total of 16 (2^4) possible combinations of the random variables. For each input $X_j{}'$, We also created the level of evidence $S_{0:j+3}$ and position Z_{j+3} using information from the chain to which the vector belongs. Binary labels $m_{i+3} \in \{0, 1\}$ were created to denote the whether the trial terminated ($m^{(i)} = 1$) or continued ($m^{(i)} = 0$) after the last entry of the vector X_{i+3}. Each element in dataset $\mathcal{D}'$ thus have 6 input features and 1 label as $\mathcal{D} := \{X^{(i)}_j, X^{(i)}_{j+1}, X^{(i)}_{j+2}, X^{(i)}_{j+3}, S^{(i)}_{0:j+3}, Z_{j+3}, m^{(i)}_{i+3}\}$. Dataset $\mathcal{D}'$ has a total of 16931 samples for 250ms condition, and 29828 samples for 100ms condition.

It is worth mentioning that under the classification framework, a positive sample is coded

as +1 towards the ground truth at the position where the chain is truncated, rather than the ground truth of the entire chain. We used this approach because the goal is to model the data at each time point using only the information available up to that point, without considering any future information. For simplicity, we used the absolute value of evidence for all the models.

Receiver Operating Characteristic Curve analysis

As both 100ms and 250ms condition demonstrates similar patterns, we will prioritize the results on 250ms.

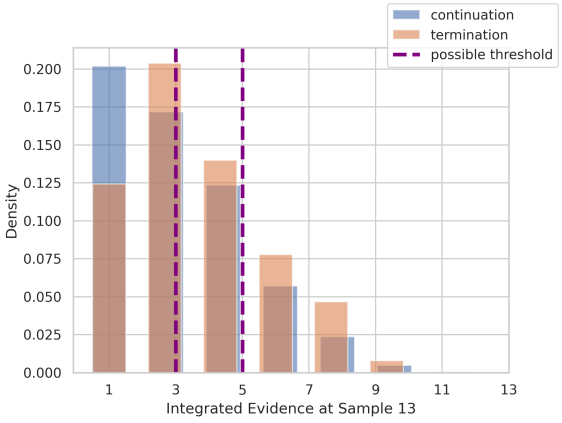


Figure 4.2: Distributions of Sum of Evidence from two classes. Dotted purple lines indicate the possible thresholds used to classify the two classes.

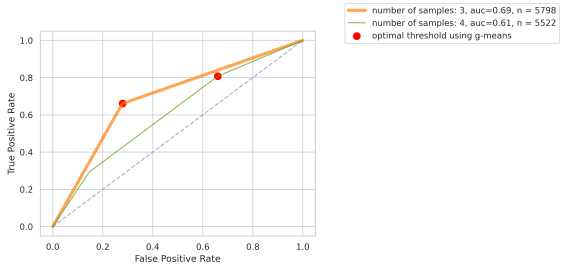


Figure 4.3: Demonstration of ROC Curves for two classifiers each with sample length of 3 and 4. Red dot implies the optimal threshold calculated using G-Means.

To test various hypotheses, at each sample length, we use use hypothesis-driven Decision Function G^z to classify whether subjects terminate or proceed given the sequence X_i with the same length. The resulting predicted values calculated from sequences in both classes (e.g., cumulative sum as shown in Figure 4.2) can be viewed as two different distributions. The classification problem can be thought of as how to find the best criterion to distinguish the two classes, similar to the idea in signal detection theory. At each possible threshold, we have True Positive Rate and False Positive Rate. If we plot these results on a 2D plane,

we will receive a curve for each classifier. A perfect classifier will have a curve at the left upper right corner, where TPR is 1 and FPR is 0. Therefore, the closer the curve is to the left upper corner, the better the classifier is. The area under the curve is an AUC curve. AUC higher than 0.5 indicates that the hit rate (TPR) is always larger than the false alarm (FPR), suggesting a reasonable classifier.

To obtain the best threshold to classify the two distributions, we use the Geometric Mean (G-mean) to measure the classification performance to balance sensitivity (True Positive Rate) and specificity (False Positive Rate):

$$GSS = \sqrt{\text{sensitivity} * \text{specificity}} = \sqrt{(TPR * (1 - FPR)}. \tag{4.4}$$

ROC Curve in orange shown in Figure 4.3 shows two examples of ROC curve for two different classifiers. The red dot is obtained using G-Means, representing the best threshold given that classifier. In the current framework, we have a separate classifier at each sample length, and therefore will obtain different optimal thresholds to examine if they change over time when they are interpretable.

For each hypothesis, we considered two questions to be answered. First, we evaluated how well the model can classify trials that proceeded and trials that terminated at each position. Second, we examined how the function produces readout given the input sequence, resulting in stopping criterion used for classification.

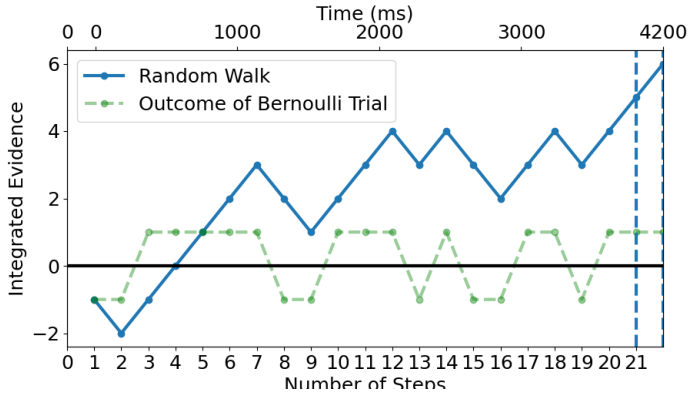


Figure 4.4: Integrated random walk as evidence accumulation.

4.2.4 Testing hypotheses using traditional Sequential Sampling Models

Hypothesis 1: Sum of Evidence

Under this hypothesis, we assume that subjects integrate evidence as information comes in, such that for a given trial $S_i = \sum_{j=1}^{z_i} X_j^{(i)}$, and $\{S_0^{(i)}, S_1^{(i)}, ..., S_z^{(i)}\}$ is therefore a random walk with a hypothetical starting point at 0. Figure 4.4 demonstrates an observation sequence (green) generated with a probability of 0.62 to move +1 and the corresponding hypothetical trajectory (blue).

We tested whether subjects used a deterministic fixed level of evidence, a fixed boundary, or a boundary that varies as a function of time, by considering the two hypotheses:

- H_0: The stopping criteria is deterministic, a fixed bound b_0 is used regardless number of samples.

$$G_z\left(X_0, X_1, ...X_z\right) = \begin{cases} 1, & S_z = b_0 \\ 0, & S_z \neq b_0 \end{cases} \tag{4.5}$$

- H_1: The stopping criteria changes as the number of samples increase n , where $f\left(S_n\right))$ is a function of S_n and n.

$$G_z\left(X_0, X_1, ...X_z\right) = \begin{cases} 1, & S_z = f(S_z, n) \\ 0, & S_z \neq f(S_z, n) \end{cases} \tag{4.6}$$

A fixed threshold of a random walk is equivalent to a scaled LLR test . According to the Neyman-Pearson Lemma, this method is the most powerful for hypothesis testing. Let $X_1, X_2, ...X_n$ be rv's from a Bernoulli trial such that $X_i \in \{0, +1\}$. Suppose subjects observed $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$ and they know the Bernoulli trials are biased, with either probability of p or $1-p$ of getting $X_i = +1$. They are testing between two simple hypotheses a with a parameter θ, where $H_0 : \theta = p$ and $H_1 : \theta = 1-p$. For instance, during the experiment θ could be the probability of a stimulus "X". Mathematically, given the PMF of Bernoulli distribution, the Likelihood Ratio test can be written as:

$$\begin{aligned} LR &= \frac{L_1\left(\theta \mid x_1 \ldots x_n\right)}{L_2\left(\theta \mid x_1 \ldots x_n\right)} \\ &= \frac{p^{\sum x_i} \cdot (1-p)^{n-\sum x_i}}{(1-p)^{\sum x_i} \cdot p^{n-\sum x_i}} \end{aligned} \tag{4.7}$$

we can then take the log and get:

$$LLR = 2\sum\left(x_i - n\right) - n \cdot \log(\frac{p}{1-p})$$

Since $\log(\frac{p}{1-p})$ is a constant, it is the same as random walk expressed as $S_n = 2\sum x_i - n$ with a different y-intercept. The proof is shown in the appendix (see Equations C.4 and C.5).

Given the training session prior to the experiment, another possibility of the LLR subjects could be testing is whether there is evidence higher than the known probability, assuming

that learned the random walk is biased with $p = 0.62$. Is it possible that people look for additional evidence to discount the learned probability?

The slope of the random walk obtained from $\frac{S_z}{n}$ can be used to approximate the drift parameter in the DDM, representing the evidence accumulation rate. Given the probability p, expectation of evidence accumulation rate is 0.24 obtained from $E(X_i) = 2p - 1$, and expectation of sum of evidence at each time step is $\{0.24, 0.48, ..., 6.96\}$ obtained from $E(Sn) = n(2p - 1)$. Therefore, at each step, the amount of evidence accounting for the learned probability is $S'_n = \sum X_i - E(S_i)$. If this hypothesis is true, we should expect to see evidence increases with samples, but stays at a fixed level of evidence corrected by expectation across samples.

Hypothesis 2: Rectified Evidence

Note that under the current evaluation framework, we can not directly distinguish between the Random Walk Model and the Rectified Evidence Model because accumulating on on counter is naturally inhibiting the other counter. This is because for m observations of positive samples out of n total samples, evidence from a random walk is $S_n = 2m - n$, and the counter towards final decision on a correct trial is m. Therefore, at each sample n, the classification task is essential the same except for a different numerical value of threshold b_0 for, as shown in Eqaution 4.9 and 4.11. For instance, an integrated evidence at $+5$ terminating at sample $n = 7$ after hitting the boundary would be the same as having one counter towards final answer (the up counter) at $+6$ and the alternative counter (the down counter) at -1. The only difference is the numerical value of threshold. For visualization, we still present the total numbers of up counter and down counter for this hypothesis.

Hypothesis 3: Max Consecutive Run

One of the advantage of an observable stream of information is that we can examine that given the same amount of information, whether the order in which the samples are presented has an effect on evidence accumulation. Under this hypothesis, we test whether the max value of uninterrupted occurrences of samples a stopping rule, defined as M_k:

$$\text{Max}\left(\sum_{j=i}^{i+k-1}\delta(X_j, X_{j-1})\right) \text{ for } 1 \leq i \leq n,\ k \geq 1, \tag{4.8}$$

Where the function $\delta(X_j, X_{j-1}) = 1$ if $X_j = X_{j-1},$ and 0 otherwise

- H_0: The stopping criteria is deterministic, a fixed bound b_0 is used regardless number of samples.

$$G_z\left(X_0, X_1, ...X_z\right) = \begin{cases} 1, & M_k = m_0 \\ 0, & M_k \neq m_0 \end{cases} \tag{4.9}$$

- H_1: The stopping criteria changes as the number of samples increase n , where $f\left(M_k\right))$ is a function of S_n and n.

$$G_z\left(X_0, X_1, ...X_z\right) = \begin{cases} 1, & M_k = f(M_k, n) \\ 0, & M_k \neq f(M_k, n) \end{cases} \tag{4.10}$$

Hypothesis 4: Weighted Random Walk

If evidence accumulation follows an O-U model, the samples can be modeled as a running weighted sum of a random walk to account for the effects of the decay term and the time-varying drift rate. When we fit a separate simple linear regression model at each sample $Y = \beta_0 + \beta_1 S_z$ for sample $k = 3, 4, ..., 15$, the coefficients $beta_1$ vary across models with a similar intercept shown in C.3. Therefore, for every unit increase in sum of evidence, the log-odds ($\log\left(\frac{P(Y=1)}{P(Y=0)}\right)$) of termination is different as sample increases. We hypothesis that the sum of evidence is weighted as each sample comes in:

$$G_z\left(X_0, X_1, ...X_z\right) = \begin{cases} 1, & \text{logit}(P) > 0 \\ 0, & \text{logit}(P) <= 0 \end{cases} \tag{4.11}$$

where the log odds $\text{logit}(P) = \ln\left(\frac{p}{1-p}\right) = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_z X_z$.

We fit the model using the training data, and applied the coefficients on test data. For interpretability of the coefficients, we fit the logistic regression without the intercept b_0 as it is fair to assume that the probability of observing a success ($Pr(G_z = 1)$) is equal to 0.5 when all predictor variables are zero. This allows us to directly compare the coefficient across models with input at different length.

Thus, we will test the hypothesis on whether the sum of evidence is weighted the same or differently over time:

- $H_0 : \beta_1 = \beta_2 = \cdots = \beta_z$.
- $H_1 : \beta_1 \neq \beta_2 \neq \ldots \neq \beta_z$.

Mathematically, we can show that H_0 is theoretically equivalent to weighted sum of each sample in the same direction but linearly decreasing $(k\beta_1 X_1, (k-1)\beta_1 X_2, ..., \beta_1 X_k)$, whereas H_1 represents that each sample is weighted differently $((\beta_1+\beta_2+...\beta_k)X_1, (\beta_2+...\beta_k)X_2, ..., \beta_k X_k)$.

4.2.5 Non-Linear Machine Learning Model: Extreme Gradient Boosting Decision Tree

XGBoost (Extreme Gradient Boosting) is a gradient tree boosting technique that has demonstrated state-of-the-art in supervised learning [8] . The model uses an ensemble of regression trees and sums up the continuous scores of each leave to combine weak learners and compute the final prediction. Gradient boosting is a sequential modeling approach, during which a subsequent model focuses on learning the prediction errors from the previous model. Through this iterative process, XGBoost can capture non-linearity between input and target response, as well as interaction among features.

To train the model, some parameters needed to be tuned using the validation set. The original dataset $\mathcal{D}$ was divided into three subsets: 60% for training, 20% for validation, and 20% for testing. The final parameters after a grid search were: colsample_bytree (subsampling of columns) = 0.75, gamma (regularization) = 1.5, learning rate = 0.1, max_depth = 5, subsample = 0.75. An early stopping of 10 is used and number of estimator is set ot be 100.

After XGBoost models are trained, we used Partial Dependency Plots (PDP) to visualize the dependence between the target response and input features while marginalizing over the values of all other input features [13]. PDP is defined as

$$\begin{aligned} pd_{X_S}(x_S) &\overset{\text{def}}{=} \mathbb{E}_{X_C}\left[f(x_S, X_C)\right] \\ &= \int f(x_S, x_C)\, p(x_C)\, dx_C \end{aligned} \tag{4.12}$$

, where $f(x_S, x_C)$ is the response function given an input sample and x_S and x_C are the two features.

4.3 Results

4.3.1 Accuracy and Trial Distribution

Accuracy for pPDM is overall high. Figure 4.5A shows the accuracy by subject. We excluded subject 109 for all further analyses due to low accuracy. Figure 4.5B shows the accuracy by sample at which a trial is terminated by a response. Performance is overall high except for trials that end at the last four samples. We will mainly present results using correct trials that end before sample 15. Figure 4.5C shows the accuracy by evidence at termination. Performance is high for $1 <= S(i) <= 15$.

At any given sample position, indecision would require asking for more samples, and sufficient sampling would result in an overt response, defined as trials proceeded and trials terminated respectively. Figure 4.5D and E show the number of trials proceeded and terminated (colored in red) at each sample position. Figure4.5F and G show the distribution of trials terminated at each level of evidence. These results suggest that across subjects, neither condition has a fixed sample position, or a fixed level of evidence used as a deterministic decision rule. Figure4.5F and G shows that both conditions have central values of evidence $S_i = 3$ for odd number of samples. The 100ms condition has central values of evidence $S_i = 2$ for even number of samples, and the 200ms condition has central values of evidence $S_i = 2$ for even number of samples. These behavioral data suggest that humans are not always optimal decision-makers, as the distributions of evidence at termination for either condition are not narrow.

Figure 4.5: (A) Accuracy by subjects. (B) Accuracy by sample at which the trials are terminated. (C) Accuracy by level of evidence. (D) & (E) Number of trials that proceeded and terminated at each sample. (F) & (G) Distribution of trials by level of evidence at response. X-axes are truncated given the low occurrences of extreme values.

4.3.2 Variability of Sum of Evidence and Evidence Accumulation Rate

To investigate what contributes to the variability in boundary, we looked at the relationship between number of samples and evidence at termination (bound). Figure 4.6A shows that overall bound increases as sample increases, however, the upward trend has a similar slope as the expectation of the random walk $\mathbb{E}S_n = n(\mathbb{E}) = n(2p-1)$. Figure 4.6B shows the Evidence Accumulation Rate at different level of evidence. We can see that evidence accumulation rate increased as the level of evidence at termination increased. Figure 4.6 C shows that when evidence at termination subtract the expected value of the random walk with $p = 0.62$, the evidence level is slightly more flat.

These results suggest the complexity of sequential sampling process. First, bound increases as number of samples goes up, partially due to the expectation of random walk given a learned probability. Second, we can still see that evidence accumulation rate vary across trials at each level of evidence after correction, suggesting variability in evidence accumulation rate across trials. This finding is consistent with the DDM models that captures inter-trial drift variability [40].Third, we can also observe that when the evidence accumulation rate is 1, that is, when there are only consecutive samples towards one choice, subjects still respond at the various bounds.

4.3.3 Model Comparison

Model 1: Sum of Evidence

Figure 4.7 demonstrates the distribution of evidence from the two classes of trials $\{X_i \mid m_i = 0\}$ and $\{X_i \mid m_i = 1\}$ at each sample position. The distributions highly overlap with slight

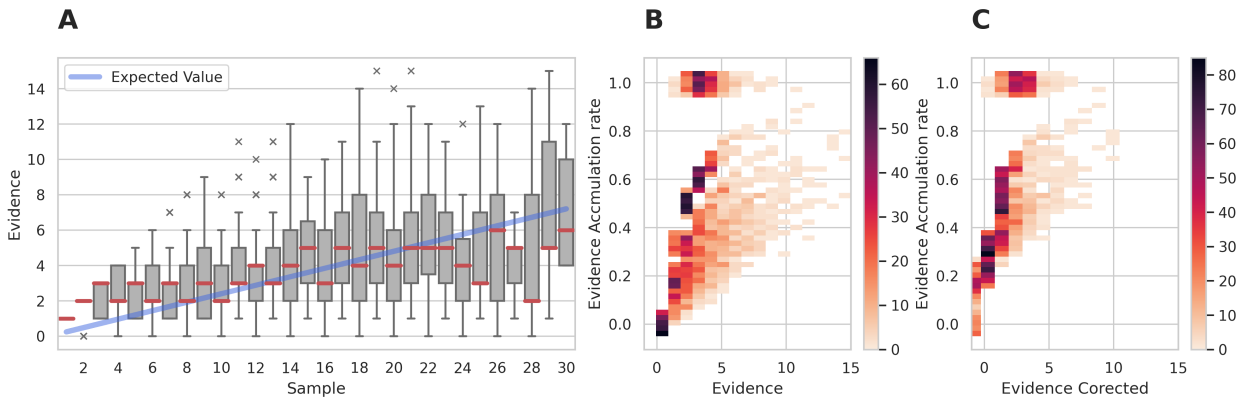


Figure 4.6: Evidence Accumulation Rate at different level of evidence for 250ms condition. Values are slightly jittered for better visualization.

different spread, leading to high false alarm and false negative rate. 4.8A and B show the ROC curve for training and test data. Note that training and test data under the current hypothesis is used as a reference for the following analyses, as no model parameter is needed to learn for G_z here.

Figure 4.8C shows the AUC score up to $z = 15$. The results suggesting that sum of evidence is a reasonable hypothesis as most AUC scores are above 0.5 for both training and test data. Figure 4.8 D shows the optimal threshold used at each sample position to classify the trials. Note that level of evidence can only be odd number if number of samples is odd, and level of evidence can only be even number if number of samples is even given the nature of simple random walk. We find some evidence favoring H_1, and that threshold increases as the number of samples increase. Similar results supporting H_1 can be shown by the results from the 100ms condition in Figure **??**.

Integration of Evidence The current hypothesis is directly testing the classical Drift Diffusion Model (DDM). The AUC scores from both conditions demonstrate that sum of evidence is a reasonable criterion, but the criterion does not generalize to all trials and all sample positions. Only samples 3-7 and 15 for 250ms, and samples 4,5,6,14 for 100ms show robust results for both training and test data. 4.7

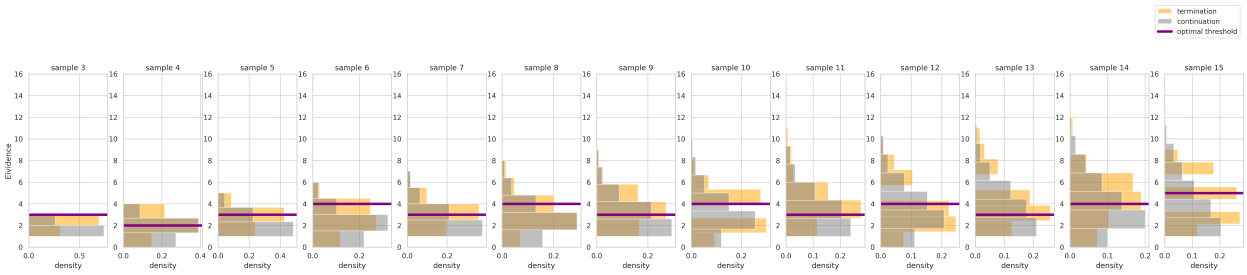


Figure 4.7: Distributions of evidence for proceeding trials and terminated trials at each sample for 250ms.

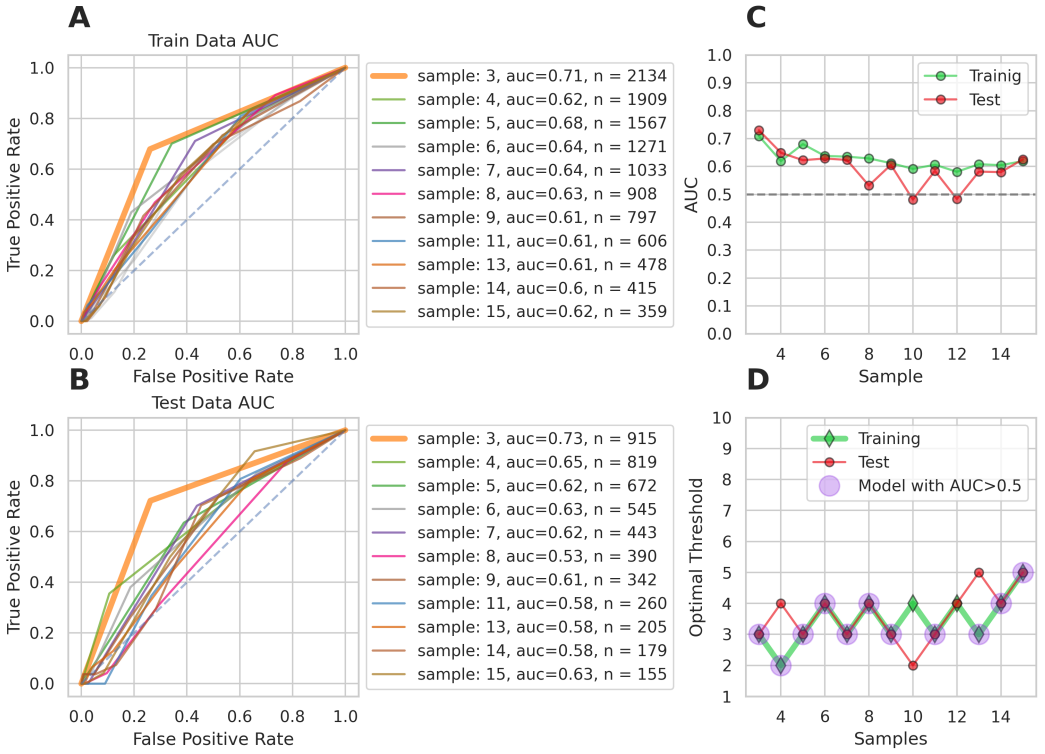


Figure 4.8: Model Performance and optimal threshold for 250ms condition. A & B,

Optimal threshold We show that the optimal threshold used to distinguish proceeding trials and terminated trials increases as number of samples increases. However, when we account for the expectation of random walk $\mathbf{E}(S_n)$, the optimal threshold decreases, suggesting a fixed and slightly collapsing threshold when a prior if formed to best fit the data.

Note that the current framework of classification can not distinguish the race model and

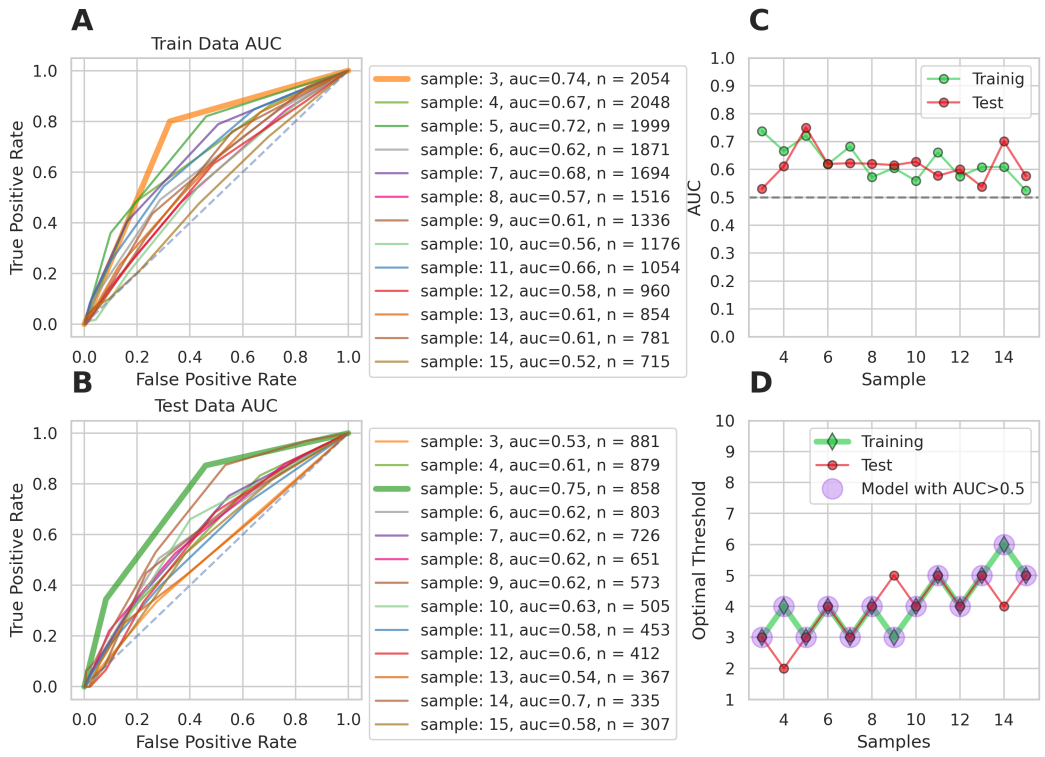
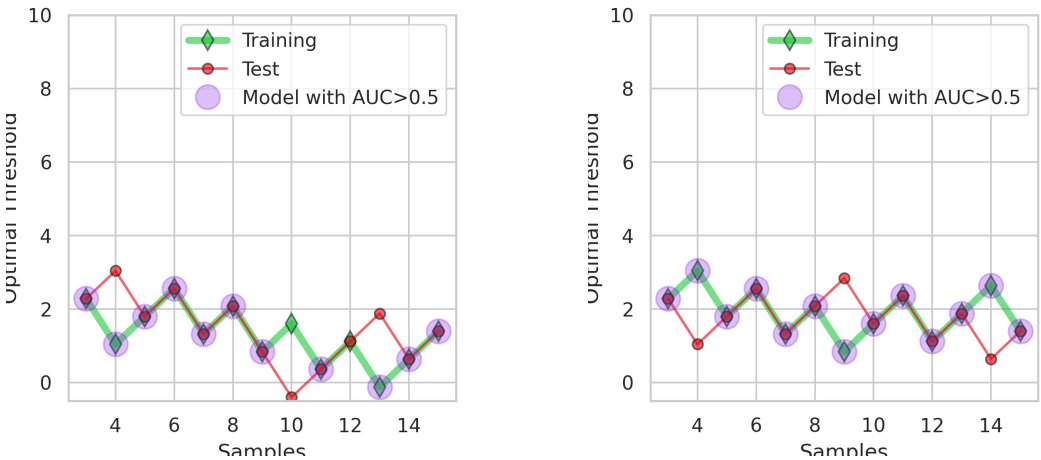


Figure 4.9: Model Performance and optimal threshold for 100ms condition. A & B,



(a) 250ms condition.

(b) 100ms condition.

Figure 4.10: Optimal threshold corrected by expectation of random walk.

the sum of evidence model because the Bernoulli trials make the two alternatives perfectly inhibiting each other.

Model 2: Max Runs

In Figure 4.7, we observe that within a sample position and at each level of evidence, subjects still exhibit variable behaviors, suggesting that the level of evidence and number of samples can not fully explain the decision process. It is possible that, in addition to the level of evidence, the path through which the evidence is reached during accumulation should also be taken into account.

Figure 4.11 are the histograms of number of max consecutive runs for proceeding trials and terminated trials, showing a slight difference between the two classes of trials up to sample 13. Figure 4.12 shows the performance of the max runs model. It is not surprising that the model performs better than Model 1 at sample 3 and 4, suggesting that for certain trials that begin with consecutive runs, subjects tend to stop asking for more samples. This effect could also be seen in Figure 4.6 when y-axis is 1. The Max runs model performs similarly for the 100m condition provided in Figure C.4 in appendix. Therefore, we conclude that subjects integrate evidence and keep track of consecutive runs depending on sample position.

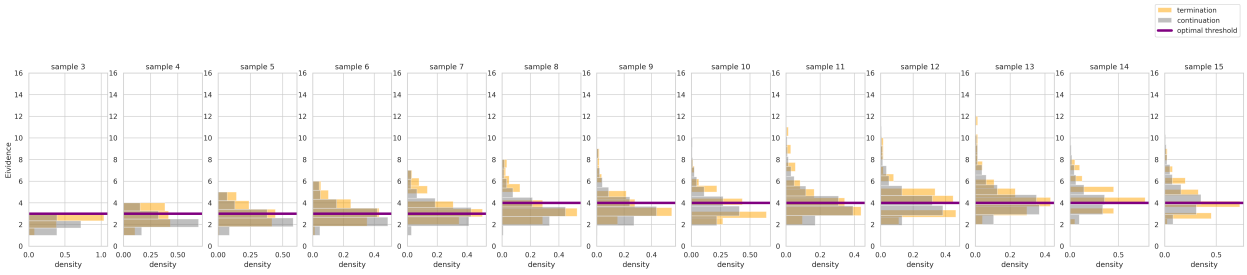


Figure 4.11: Distribution of max consecutive runs at each sample sample for 250ms condition.

Model 3: Weighted Sum of Evidence

The weighted sum of evidence model using logistic regression is the best model while still preserving linear classification boundary. Figure 4.13 demonstrates the model performance for 250ms condition, especially for samples from 7-12. When we inspect the coefficients of

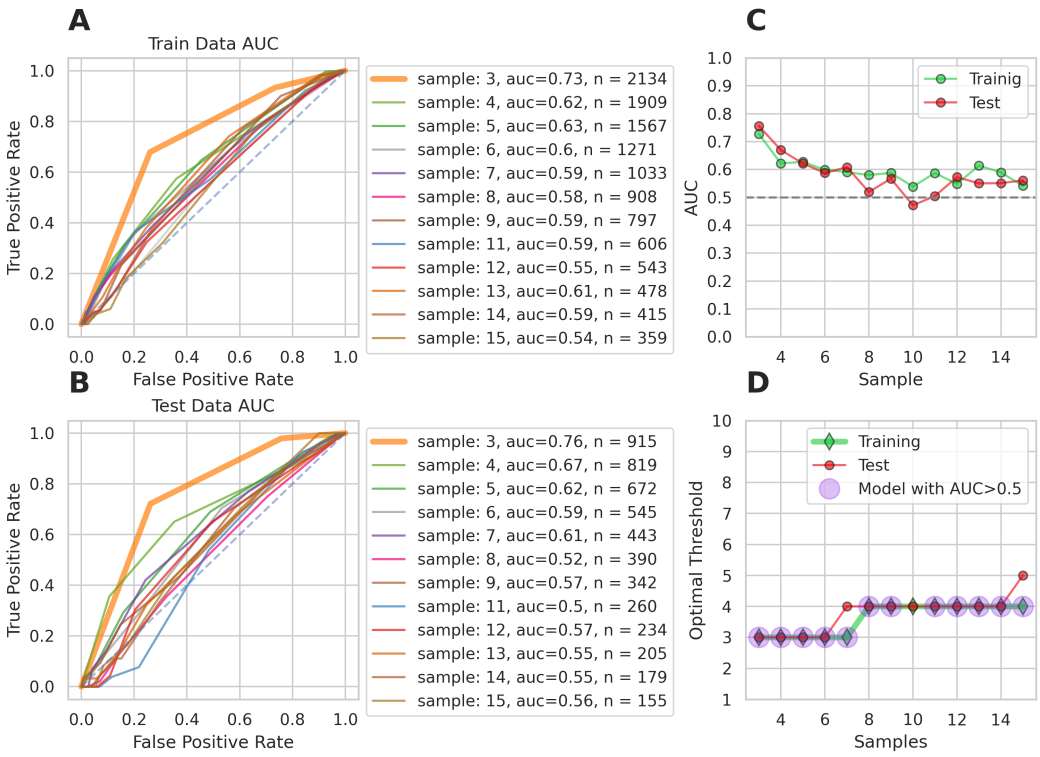


Figure 4.12: Model performance of max runs model for 250ms condition.

each model, the coefficients consistently show a recency effects, such that as samples come in, subjects upweight the recent evidence and downweight the prior evidence. If they made a decision early, that is because they accrued sufficient evidence (also supported by the max runs model for samples at 3 and 4), if they chose to wait for more samples, they will compensate the passing of earlier samples, such that an increase in recent evidence has a higher log odds of an overt response. Note that coefficient for X_1 is skipped as the first sample is always +1 under this current framework. Small improvement of model performance can also be observed in the 100ms condition shown in Figure 4.15, and the recency effect can be observed in Figure 4.16 with one sample delay, as only 2 samples were backtracked in order to correct for nondecision time.

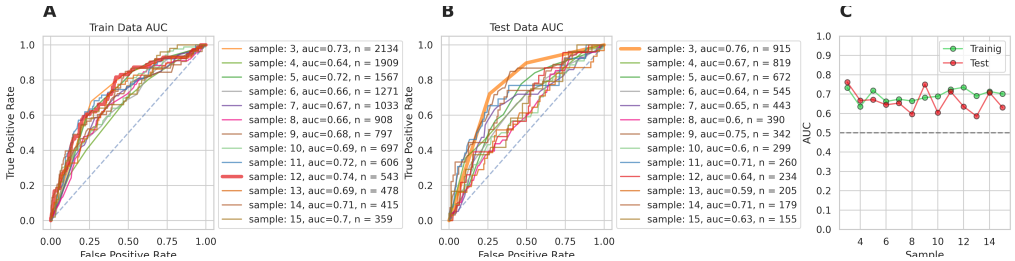


Figure 4.13: Weighted Sum of evidence model performance for 250ms condition.

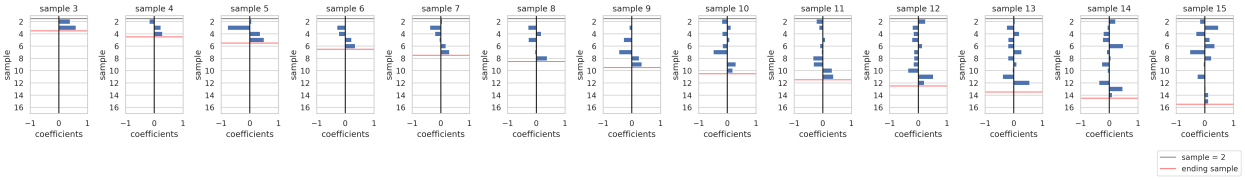


Figure 4.14: Coefficients from logistic regression fit at sample for 250ms condition.

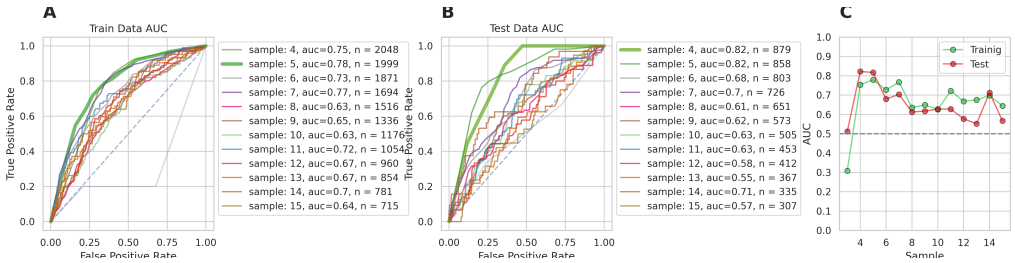


Figure 4.15: Weighted Sum of evidence model performance for 100ms condition.

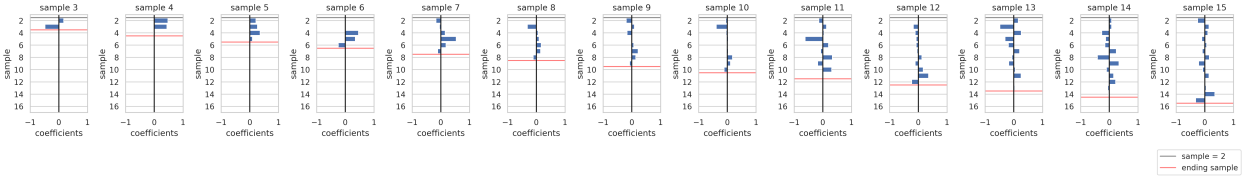


Figure 4.16: Coefficients from logistic regression fit at sample for 100ms condition.

4.3.4 Local signals with XGBoost Tree Ensemble Model

Combinatorial Representation of Local Signal

We used dataset $\mathcal{D}'$ for Figure 4.17 to show the number of occurrences for all possible combinations and the corresponding probability of termination on X_{i+3}. The sequence $+1, +1, +1, +1$ is most likely with the highest number of occurrences because given $p = 0.62$, $\Pr(4 \text{ Positives}) = p^4 = 0.148$ followed by probability of 3 positive, 2 positives, 3 negatives

and 4 negatives. Figure 4.17A shows all the possible combinations sorted from highest to lowest probability of being followed by a termination response. There is a systematic pattern, where sequences with more positives are more likely to result in termination. Addition, a subtle effect of sequence order is can be observed especially when $\sum X_{i:i+3} = 3$. Figure 4.17B shows all the occurences where the level of evidence at the final entry regarding the entire sequence is 0, sorted the same order as in A. While other patterns are no longer systematic, in the rare cases when $\sum X_{i:i+3} = 4$, subjects show the highest probability of responding. Figure 4.17C and D show the occurrences where level of evidence is low ($<= 2$) compared to high (> 2). There is a small effect of such that, given the same sequence order, subjects are more likely to respond at a higher evidence across all combinations.

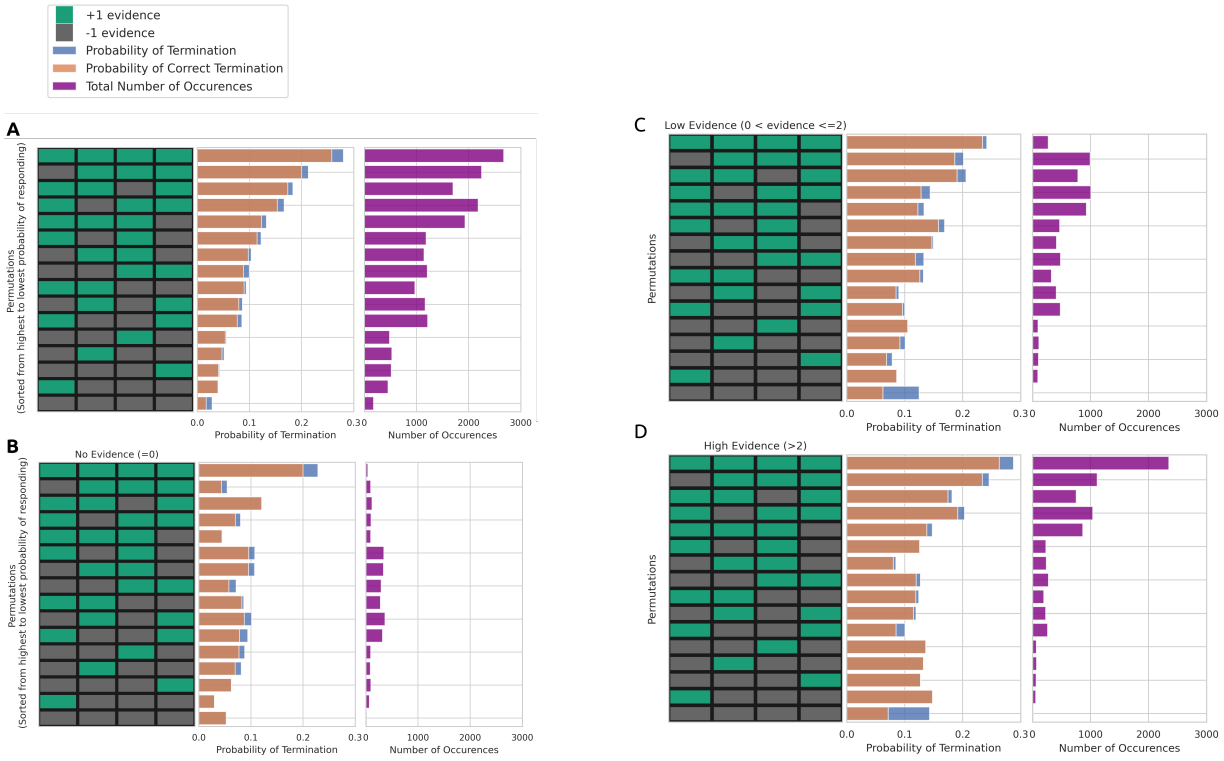


Figure 4.17: Local signal for 250ms.

XGBoost Results

Our results demonstrate that subjects adopt complex behaviors when performing the task, guided by evidence, sample, local signals with a length of 4, and the interactions among these factors. We used XGBoost to model the non-linear relationships between the 6 predictors and the response, in order to account for the effects from all previous models, assuming that local signals can capture the signals in max run model and logistic regression model.

Figure 4.18 shows the loss functions on train and validation data, and the performance of XGBoost model using all the occurrences for each condition, resulting in an overall performance measured by AUC score at 0.68 and 0.7.

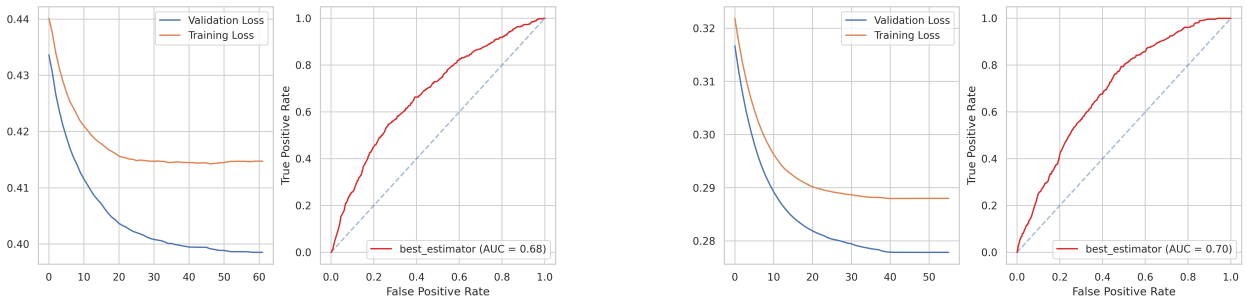


Figure 4.18: Training and validation Loss of XGBoost Model for 250ms (left) and 100ms (right) conditions each with all samples combined.

Figure 4.19 and 4.20 show the performance of the trained model tested the same train/test split matched to previous models for a fair comparison. Sequences with a length of 3 were omitted. For the 250ms condition, the model seem to perform better for earlier and later samples, showing a similar but better pattern as Max Runs Model. The model is more robust as it performs more consistently between training and test set. For the 100ms condition, the model performs the best among all existing models.

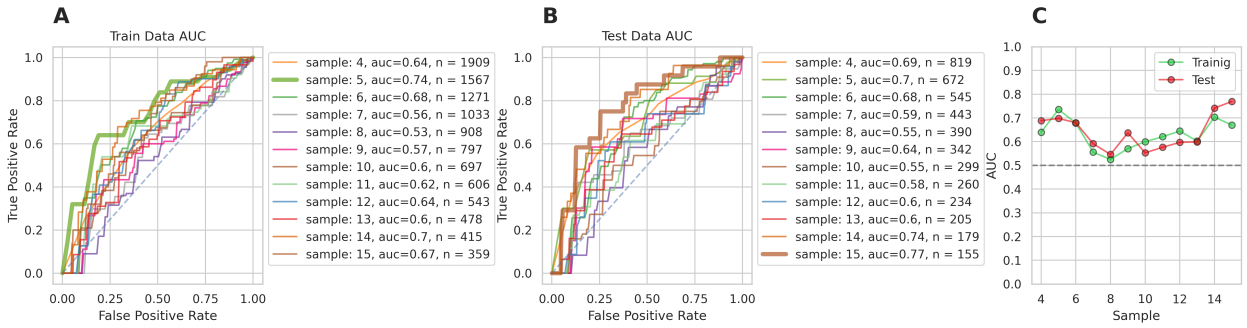


Figure 4.19: Performance of trained XGBoost model on training and test data for 250ms condition.

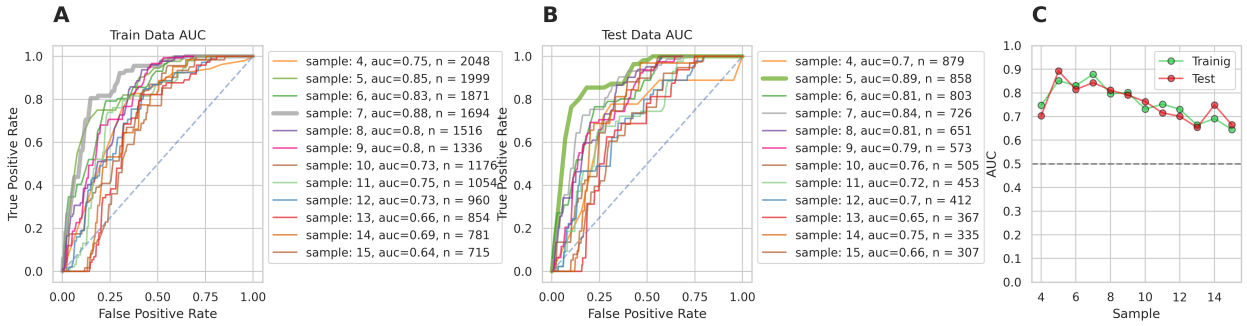


Figure 4.20: Performance of trained XGBoost model on training and test data for 100ms condition.

Interpretating XGBoost Model with Partial Dependence and Feature Importance

We used PDP to investigate the probability of termination as a function of each feature for a causal interpretation. For the 250ms condition shown in Figure 4.21A-D, each local signal increases the probability of termination if it is positive, consistent with the patterns shown in Figure 4.17A. Probability as a function of level of evidence in Figure 4.21B shows an overall increasing trend. Interestingly, the function has a global maximum at 3 and 4, followed by a decrease as evidence level goes up at 5 and 6, indicating a non-linear relationship. A similar pattern can also be seen in Figure 4.22.

The 2D PDP plot representing the relationship between sample position and evidence in Figure 4.21D shows that subjects has the highest probability of responding when trials hike

up quickly in the first several samples (reaching evidence at 3 and 4), and in the trials when more than 15 samples are seen. Figure 4.23 shows the probability to termination as a function of sample by each level of evidence using the average from Partial Dependence. Both conditions show that subjects are more likely to respond when the initial evidence strong. For the 250ms condition, subjects increase the probability of responding again after seeing half of the samples. For the 100ms condition, the threshold to increase the probability again is around 20 samples.

Lastly, Figure 4.21H and Figure 4.22H shows the feature importance, computed by the number of times a feature is used to split the data across all trees. For both models, Evidence is used the most, followed by sample position. These two features far outweigh the importance of local signals.

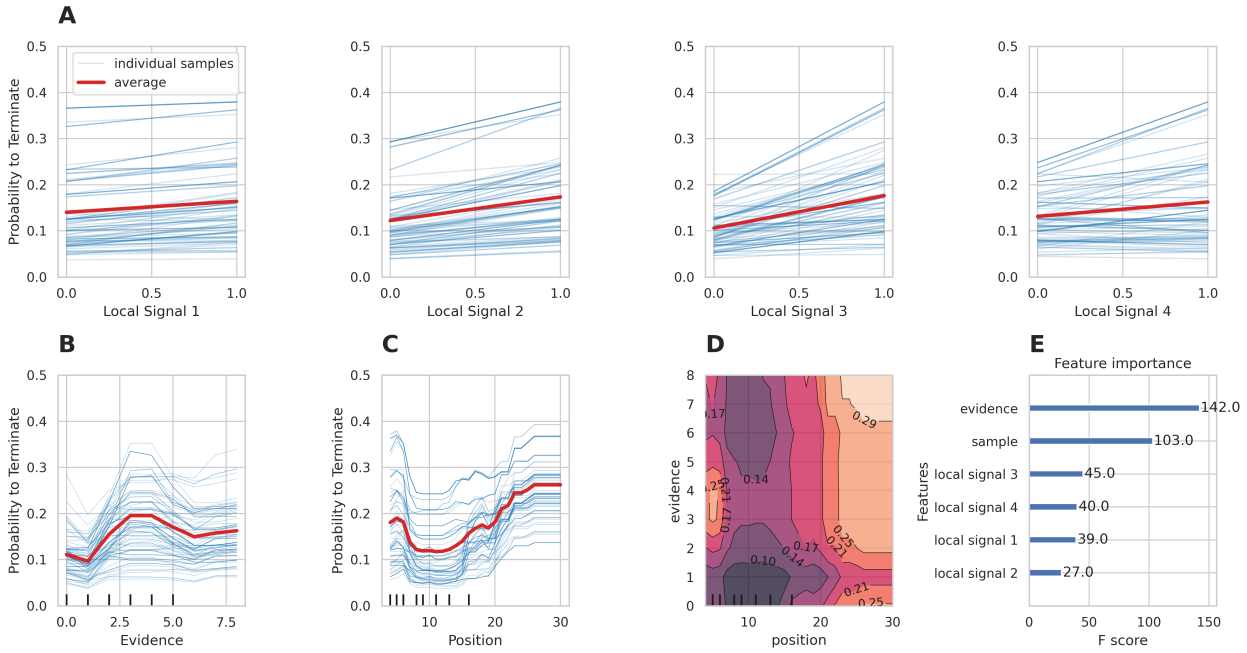


Figure 4.21: Partial Dependency Plots (A-D) and feature importance (E) for 250ms condition. Individual samples were plotted 100 input samples randomly drawn from the dataset.

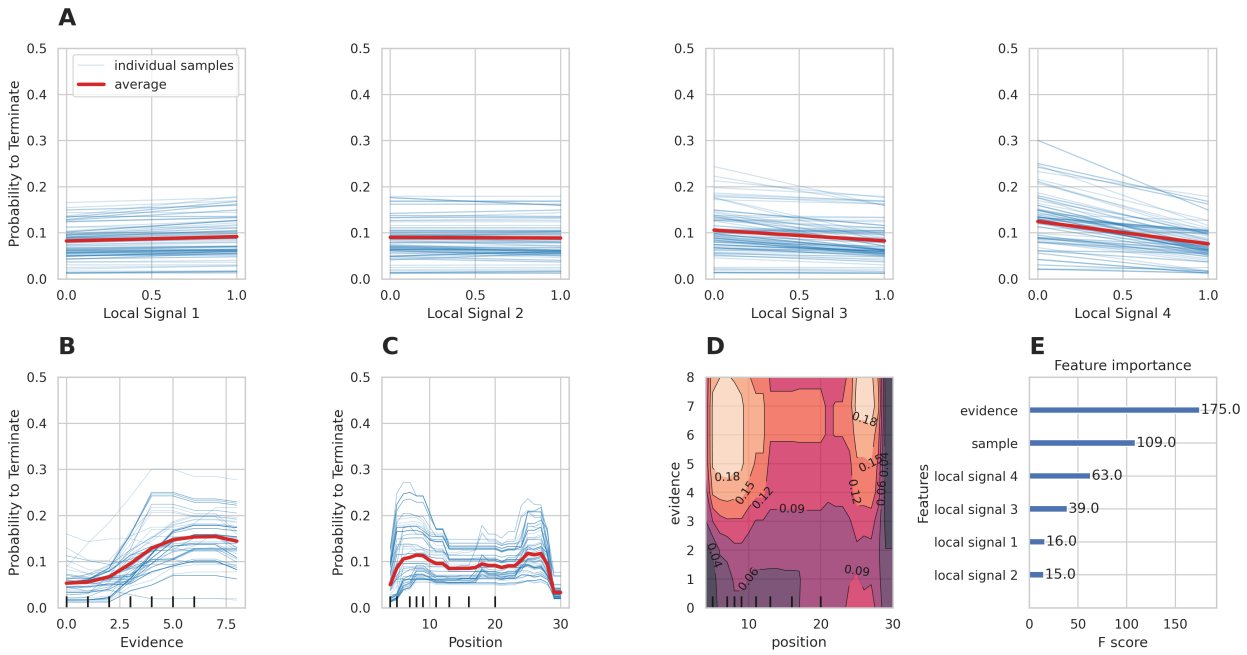
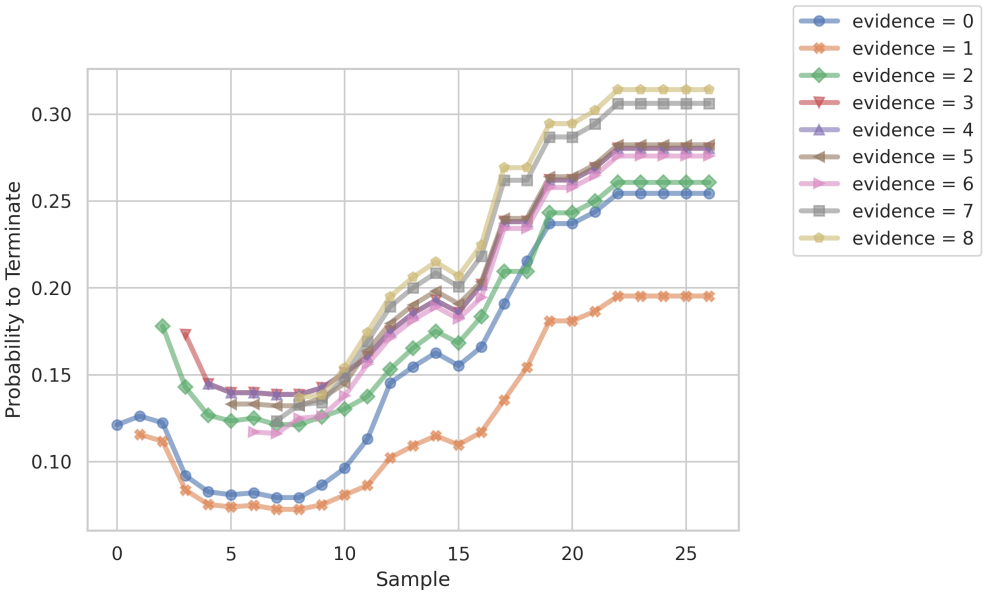


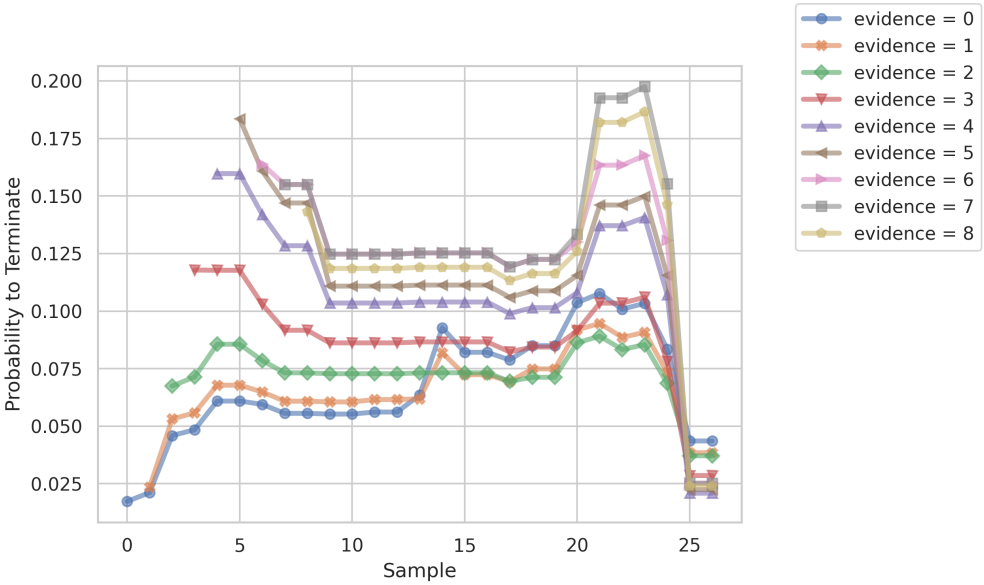
Figure 4.22: Partial Dependency Plots (A-D) and feature importance (E) for 100ms condition.

4.3.5 ROC AUC Summary

We compared the performance of four major models shown in Table for 250ms condition and in Table for 100ms condition. AUC scores higher than 0.6 are in bold, and cells are colored in yellow if both training and test data are the highest among all four models. Weighted Sum of Evidence Model and XGBoost are the two better models. For the 250ms condition, XGBoost model can better capture behavior in earlier and later samples, whereas logistic regression can best model the behavior from sample 7 to 12, where the recency effect is the most prominent.



(a) 250ms condition.



(b) 100ms condition.

Figure 4.23: Probability to terminate as a function of sample by each level of evidence for 250ms condition (a) and 100ms condition (b) calculated by the average effect from Partial Dependence.

Model	Sample												
	3	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
sum	**0.71**	**0.62**	**0.68**	**0.64**	**0.64**	**0.63**	**0.61**	0.59	**0.61**	0.58	**0.61**	**0.6**	**0.62**
	0.73	**0.65**	**0.62**	**0.63**	**0.62**	0.53	**0.61**	0.48	0.58	0.48	0.58	0.58	**0.63**
runs	**0.73**	**0.62**	**0.63**	**0.6**	0.59	0.58	0.59	0.54	0.59	0.55	**0.61**	0.59	0.54
	0.76	**0.67**	**0.62**	0.59	**0.61**	0.52	0.57	0.47	0.5	0.57	0.55	0.55	0.56
logit	**0.73**	**0.64**	**0.72**	**0.66**	**0.67**	**0.66**	**0.68**	**0.69**	**0.72**	**0.74**	**0.69**	**0.71**	**0.7**
	0.76	**0.67**	**0.67**	**0.64**	**0.65**	**0.6**	**0.75**	**0.6**	**0.71**	**0.64**	0.59	**0.71**	**0.63**
xgb	-	**0.64**	**0.74**	**0.68**	0.56	0.53	0.57	**0.6**	**0.62**	**0.64**	**0.6**	**0.7**	**0.67**
	-	**0.69**	**0.7**	**0.68**	0.59	0.55	**0.64**	0.55	0.58	**0.6**	**0.6**	**0.74**	**0.77**

Table 4.1: Table with AUC scores for training and test data for 250ms. The first row in each cell shows the AUC score for the training data, and the second row shows the AUC score for the test data. AUC scores greater than 0.6 are shown in bold. Cells colored in yellow indicate that the model performs the best for both training and test data.

Model	Sample												
	3	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
sum	**0.74**	**0.67**	**0.72**	**0.62**	**0.68**	0.57	**0.61**	0.56	**0.66**	0.58	**0.61**	**0.61**	0.52
	0.53	**0.61**	**0.75**	**0.62**	**0.62**	**0.62**	**0.62**	**0.63**	0.58	**0.6**	0.54	**0.7**	0.58
runs	**0.76**	**0.7**	**0.72**	**0.64**	**0.67**	0.55	0.55	**0.6**	0.58	0.55	**0.6**	0.58	0.53
	0.49	**0.67**	**0.76**	**0.65**	**0.63**	0.59	**0.6**	**0.63**	**0.62**	0.57	0.52	**0.67**	0.56
logit	0.31	**0.75**	**0.78**	**0.73**	**0.77**	**0.63**	**0.65**	**0.63**	**0.72**	**0.67**	**0.67**	**0.7**	**0.64**
	0.51	**0.82**	**0.82**	**0.68**	**0.7**	**0.61**	**0.62**	**0.63**	**0.63**	0.58	0.55	**0.71**	0.57
xgb	-	**0.75**	**0.85**	**0.83**	**0.88**	**0.8**	**0.8**	**0.73**	**0.75**	**0.73**	**0.66**	**0.69**	**0.64**
	-	**0.7**	**0.89**	**0.81**	**0.84**	**0.81**	**0.79**	**0.76**	**0.72**	**0.7**	**0.65**	**0.75**	**0.66**

Table 4.2: Table with AUC scores for training and test data for 100ms.

4.4 Discussion

4.4.1 Conclusion

We used the pPDM experiment to quantitatively test the fundamental theory in sequential sampling in human decision making. We formally tested hypotheses on mechanisms during evidence accumulation with four models, and tried to identify a model that best describe the way people behaved by evaluating the performance as a classification task. Subjects show high accuracy in completing the task. We observed complex patterns of behavior and introduced models that can reasonably account for multiple mechanisms. More specifically,

new models we introduced provide interpretable results to better explain decision making process.

Integration of evidence Our results suggest that humans continuously integrate evidence, but the threshold is not fixed. Decision criterion given the level of evidence is affected by time (quantified as number of samples), the amount of evidence contained in the last several samples (local evidence), and the expectation of evidence accumulation rate as a prior belief.

Evidence as a function of time include two factors. One is that as time goes by, evidence needed to make a decision generally increases, as demonstrated by the optimal threshold identified by Model 1. However, this effect can be eliminated by accounting for the expected evidence accumulation rate, suggesting that there might be a fixed level of significance subjects used to on Likelihood Ratio to test hypothesis. The fixed criterion is not demonstrated in the form of a fixed decision bound Z_1 and Z_2, but instead an increased bound as a function of time$Z(n)$. This can be considered as *increasing bound*, rather than collapsing bound in the literature [17]. If the idea of increasing bound holds true, the findings would indirectly support Urgency Gating model, as there would theoretically require a separate signal that pushes for a response to conclude a trial rather than waiting indefinitely. What makes it more complicated is that the effect of increasing bound on probability is non-linear, such that lower evidence increase the probability more, and higher evidence increases the probability less.

Time as a function of response More samples also increases the probability of making a decision. However, we found two strategies given the non-linear results. For earlier samples subjects have a higher chance of responding if the signal is strong towards one choice. This result is supported by the success of Model 2 Runs Model with shorter chains. For samples

in the middle of the chains, we see that subjects either respond at a lower chance, or wait to sample more and discount the prior ones. For samples closer to the end of the chain, subjects tend to respond even if evidence is low.

Among the four models, two models that we constructed best fit the data observed. Both models are inspired by existing mechanisms in sequential sampling. One of the better models is integration of evidence weighted overtime. We found a recency effect such that subjects focus on recent samples and suppress prior samples. This model represents a process similar to O-U process, rather than a Standard DDM. The other model is a non-linear ensemble model using gradient-boosted trees that utilizes amount of evidence, number of samples (course of time), local signals and the interactions between them. The results suggest that computations during decision making include all of the features and the interactions among them. The resulting behavior demonstrated support for a combination of O-U Model, Urgency Gating Model, and Runs Model, but none of these models by itself can best explain the data. These results suggest that cognitive scientists should consider developing models that have the ability to account for multiple mechanisms, and design experiments that investigate how these mechanisms are employed in different scenarios.

4.4.2 Future Work

Future work should focus on three directions. First, we should assess the robustness of each model using cross-validation, as the behavioral data is highly variable. Robustness of model among all subsets of data should be considered as an evaluation metric in addition to AUC scores. Second, EEG data should be analyzed given the current behavioral results. We should try to search for brain signals that track evidence and sample position. Current classification framework can be used to compare brain signals on trials with and without a response at each sample. In addition, we should try to construct models that can incorporating both brain

signals and behavioral features. A unifying framework needs to be developed to formally describe these mechanisms.

Even though drawing samples from a Bernoulli Distribution gave us the full control and granularity to test the fundamental theories, it limited our ability to distinguish models that assume separate processes for the two alternatives, such as Race model and LCA. Future experiments generate each sample from a Gaussian distribution or from two separate distributions, so that the sample at each step is not binary. Similar studies should be able to adapt the current framework to test hypotheses. Future experiments should also manipulate the probability of the Bernoulli trials or other distributions, such that there isn't a fixed expectation on the stochastic process.

Chapter 5

Conclusion

5.1 Interpretable Neural Network for Neurocognitive Modeling

One of the challenges in fitting the Drift Diffusion Model (DDM) is that cognitive components vary across trials, leading to different trial-level behaviors. Individuals exhibit varying information processing speeds and levels of caution momentarily. Estimating these components using the Wiener likelihood given RT observed is not solvable on a trial-level as we must infer two latent parameters—drift rate and boundary based on only one observable value. The development of Decision SincNet simultaneously predict the parameters of the DDM and discover the neural correlates of these parameters in an end-to-end manner.

Decision SincNet provides power of single-trial inference as it uses brain signals as an additional source of information to constrain the latent cognitive parameters at the trial level. This approach allows for better estimation of the evidence accumulation rates and decision boundaries that would most likely produce the observed behaviors. At the same time, the

model provides valuable insights into the relationship between brain activity and cognitive processes. The design of Decision SincNet is different from classical neural network models as it was constructed as a fully interpretable model with a training objective to predict parameters of a hypothesized process. Therefore, interpreting how the model processes input data is intrinsically linked to analyzing EEG data and discovering neural correlates identified by the model. The success of the current modeling framework and visualization methods can open up new opportunities for neuroscience research. Future work can build upon this line of research and introduce different neuroimaging modalities and likelihood functions.

5.2 Complex Behaviors during Human Decision Making

We developed a 2AFC task pPDM using an observable stream of evidence to examine how unit samples of evidence given rise to an overt response that terminates the sampling process. We formally tested different existing hypotheses during evidence accumulation, and developed a classification framework to use data-driven methods to evaluate which hypothesis-guided model best describe the data we collected. We concluded that under the sequential sampling framework, humans employ complex strategies that involve multiple mechanisms to make decisions. By using a non-linear machine learning ensemble model, where all features were carefully constructed to account for relevant factors, we were able to better disentangle the processes. Interpretability methods applied to the tree-based ensemble model show that humans consider the integration of evidence, the passage of time, the sequence in which samples are presented, and the interplay between these elements to guide decision making.

5.3 Summary

In summary, our work investigated fundamental questions in decision-making by utilizing both computational modeling and empirical approaches, with a particular focus on trial-level predictions. Our results suggest that evidence integration plays an important role in decision-making, and models should account for across-trial variability. We used deep learning and machine learning techniques, including the analysis of high-dimensional EEG data, solving non-convex functions using gradient-based methods, and modeling complex human behaviors. Interpretability methods serves as an example of leveraging ML/AI for scientific research. We believe our work will inspire future research in this field to adopt new methods and address fundamental questions on brain and behavior.

Bibliography

[1] S. U. Amin, M. Alsulaiman, G. Muhammad, M. A. Mekhtiche, and M. S. Hossain. Deep learning for eeg motor imagery classification based on multi-layer cnns feature fusion. *Future Generation computer systems*, 101:542–554, 2019.

[2] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang. Filter bank common spatial pattern algorithm on bci competition iv datasets 2a and 2b. *Frontiers in neuroscience*, 6:39, 2012.

[3] N. Arun, N. Gaw, P. Singh, K. Chang, M. Aggarwal, B. Chen, K. Hoebel, S. Gupta, J. Patel, M. Gidwani, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6):e200267, 2021.

[4] R. Audley and A. Pike. Some alternative stochastic models of choice 1. *British Journal of Mathematical and Statistical Psychology*, 18(2):207–225, 1965.

[5] R. Bogacz, E. Brown, J. Moehlis, P. Holmes, and J. D. Cohen. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological review*, 113(4):700, 2006.

[6] D. Borra, S. Fantozzi, and E. Magosso. Interpretable and lightweight convolutional neural network for eeg decoding: Application to movement execution and imagination. *Neural Networks*, 129:55–74, 2020.

[7] J. R. Busemeyer and J. T. Townsend. Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review*, 100(3):432, 1993.

[8] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[9] P. Cisek, G. A. Puskas, and S. El-Murr. Decisions in changing conditions: the urgency-gating model. *Journal of Neuroscience*, 29(37):11560–11571, 2009.

[10] A. Damianou and N. D. Lawrence. Deep gaussian processes. In *Artificial intelligence and statistics*, pages 207–215. PMLR, 2013.

[11] T. H. Donner, M. Siegel, P. Fries, and A. K. Engel. Buildup of choice-predictive activity in human motor cortex during perceptual decision making. *Current Biology*, 19(18):1581–1585, 2009.

[12] D. Erhan, Y. Bengio, A. Courville, and P. Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

[13] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[14] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[15] J. I. Gold and M. N. Shadlen. The neural basis of decision making. *Annu. Rev. Neurosci.*, 30(1):535–574, 2007.

[16] D. M. Green, J. A. Swets, et al. *Signal detection theory and psychophysics*, volume 1. Wiley New York, 1966.

[17] G. E. Hawkins, B. U. Forstmann, E.-J. Wagenmakers, R. Ratcliff, and S. D. Brown. Revisiting the evidence for collapsing boundaries and urgency signals in perceptual decision-making. *Journal of Neuroscience*, 35(6):2476–2484, 2015.

[18] H. R. Heekeren, S. Marrett, and L. G. Ungerleider. The neural systems that mediate human perceptual decision making. *Nature reviews neuroscience*, 9(6):467–479, 2008.

[19] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[20] C. Ieracitano, N. Mammone, A. Hussain, and F. C. Morabito. A novel explainable machine learning approach for eeg-based brain-computer interface systems. *Neural Computing and Applications*, pages 1–14, 2021.

[21] S. P. Kelly and R. G. O'Connell. Internal and external influences on the rate of sensory evidence accumulation in the human brain. *Journal of Neuroscience*, 33(50):19434–19441, 2013.

[22] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim. The (un) reliability of saliency methods. *Explainable AI: Interpreting, explaining and visualizing deep learning*, pages 267–280, 2019.

[23] D. LaBerge. A recruitment theory of simple behavior. *Psychometrika*, 27(4):375–396, 1962.

[24] D. R. J. Laming. Information theory of choice-reaction times. 1968.

[25] S. W. Link. The relative judgement theory of two choice response time. *Journal of Mathematical Psychology*, 12:114–135, 1975.

[26] G. Malhotra, D. S. Leslie, C. J. Ludwig, and R. Bogacz. Time-varying decision boundaries: insights from optimality analysis. *Psychonomic bulletin & review*, 25:971–996, 2018.

[27] J. M. Mayor-Torres, M. Ravanelli, S. E. Medina-DeVilliers, M. D. Lerner, and G. Riccardi. Interpretable sincnet-based deep learning for emotion recognition from eeg brain activity. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 412–415. IEEE, 2021.

[28] D. J. Navarro and I. G. Fuss. Fast and accurate calculations for first-passage times in wiener diffusion models. *Journal of mathematical psychology*, 53(4):222–230, 2009.

[29] M. D. Nunez, K. Fernandez, R. Srinivasan, and J. Vandekerckhove. A tutorial on fitting joint models of m/eeg and behavior to understand cognition. *Behavior Research Methods*, pages 1–31, 2024.

[30] M. D. Nunez, A. Gosai, J. Vandekerckhove, and R. Srinivasan. The latency of a visual evoked potential tracks the onset of decision making. *Neuroimage*, 197:93–108, 2019.

[31] M. D. Nunez, J. Vandekerckhove, and R. Srinivasan. How attention influences perceptual decision making: Single-trial eeg correlates of drift-diffusion model parameters. *Journal of mathematical psychology*, 76:117–130, 2017.

[32] O. Odoemene, S. Pisupati, H. Nguyen, and A. K. Churchland. Visual evidence accumulation guides decision-making in unrestrained mice. *Journal of Neuroscience*, 38(47):10143–10155, 2018.

[33] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev. The building blocks of interpretability. *Distill*, 3(3):e10, 2018.

[34] R. G. O'Connell, M. N. Shadlen, K. Wong-Lin, and S. P. Kelly. Bridging Neural and Computational Viewpoints on Perceptual Decision-Making. *Trends in Neurosciences*, 41(11):838–852, Nov. 2018.

[35] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[36] M. G. Philiastides, R. Ratcliff, and P. Sajda. Neural representation of task difficulty and decision making during perceptual categorization: a timing diagram. *Journal of Neuroscience*, 26(35):8965–8975, 2006.

[37] M. M. Rahman, V. D. Calhoun, and S. M. Plis. Looking deeper into interpretable deep learning in neuroimaging: a comprehensive survey. *arXiv preprint arXiv:2307.09615*, 2023.

[38] R. Ratcliff. A theory of memory retrieval. *Psychological review*, 85(2):59, 1978.

[39] R. Ratcliff and G. McKoon. The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, 20(4):873–922, 2008.

[40] R. Ratcliff and J. N. Rouder. Modeling response times for two-choice decisions. *Psychological science*, 9(5):347–356, 1998.

[41] R. Ratcliff and P. L. Smith. A comparison of sequential sampling models for two-choice reaction time. *Psychological review*, 111(2):333, 2004.

[42] M. Ravanelli and Y. Bengio. Speaker recognition from raw waveform with sincnet. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 1021–1028. IEEE, 2018.

[43] J. Rennie. On l2-norm regularization and the gaussian prior, 2003.

[44] J. D. Roitman and M. N. Shadlen. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *Journal of neuroscience*, 22(21):9475–9489, 2002.

[45] S. Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

[46] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

[47] R. Saleem, B. Yuan, F. Kurugollu, A. Anjum, and L. Liu. Explaining deep neural networks: A survey on the global interpretation methods. *Neurocomputing*, 513:165–180, 2022.

[48] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[49] M. N. Shadlen and W. T. Newsome. Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology*, 86:1916–1936, 2001.

[50] H. Shimizu and R. Srinivasan. Improving classification and reconstruction of imagined images from eeg signals. *Plos one*, 17(9):e0274847, 2022.

[51] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[52] M. Stone. Models for choice-reaction time. *Psychometrika*, 25(3):251–260, 1960.

[53] C. Summerfield and F. P. De Lange. Expectation in perceptual decision making: neural and computational mechanisms. *Nature Reviews Neuroscience*, 15(11):745–756, 2014.

[54] Q. J. Sun, K. Vo, K. Lui, M. Nunez, J. Vandekerckhove, and R. Srinivasan. Decision sincnet: Neurocognitive models of decision making that predict cognitive processes from neural signals. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2022.

[55] A. Topic and M. Russo. Emotion recognition based on eeg feature maps through deep learning network. *Engineering Science and Technology, an International Journal*, 24(6):1442–1454, 2021.

[56] J. T. Townsend and F. G. Ashby. *Stochastic modeling of elementary psychological processes*. CUP Archive, 1983.

[57] B. M. Turner, B. U. Forstmann, B. C. Love, T. J. Palmeri, and L. Van Maanen. Approaches to analysis in model-based cognitive neuroscience. *Journal of Mathematical Psychology*, 76:65–79, 2017.

[58] B. M. Turner, B. U. Forstmann, E.-J. Wagenmakers, S. D. Brown, P. B. Sederberg, and M. Steyvers. A bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage*, 72:193–206, 2013.

[59] M. Usher and J. L. McClelland. The time course of perceptual choice: the leaky, competing accumulator model. *Psychological review*, 108(3):550, 2001.

[60] J. Vandekerckhove, F. Tuerlinckx, and M. D. Lee. Hierarchical diffusion models for two-choice response times. *Psychological methods*, 16(1):44, 2011.

[61] D. Vickers. Evidence for an accumulator model of psychophysical discrimination. *Ergonomics*, 13(1):37–58, 1970.

[62] K. Vo, Q. J. Sun, M. D. Nunez, J. Vandekerckhove, and R. Srinivasan. Deep latent variable joint cognitive modeling of neural signals and human behavior. *NeuroImage*, 291:120559, 2024.

[63] A. Voss, V. Lerche, U. Mertens, and J. Voss. Sequential sampling models with variable boundaries and non-normal noise: A comparison of six models. *Psychonomic bulletin & review*, 26:813–832, 2019.

[64] A. Voss, K. Rothermund, and J. Voss. Interpreting the parameters of the diffusion model: An empirical validation. *Memory & cognition*, 32(7):1206–1220, 2004.

[65] A. Wald and J. Wolfowitz. Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, pages 326–339, 1948.

[66] T. V. Wiecki, I. Sofer, and M. J. Frank. Hddm: Hierarchical bayesian estimation of the drift-diffusion model in python. *Frontiers in neuroinformatics*, 7:55610, 2013.

[67] V. Wyart, V. De Gardelle, J. Scholl, and C. Summerfield. Rhythmic fluctuations in evidence accumulation during decision making in the human brain. *Neuron*, 76(4):847–858, 2012.

[68] Y. Zhang, J. Chen, J. H. Tan, Y. Chen, Y. Chen, D. Li, L. Yang, J. Su, X. Huang, and W. Che. An investigation of deep learning models for eeg-based emotion recognition. *Frontiers in Neuroscience*, page 1344, 2020.

[69] Y. Zhang, D. Hong, D. McClement, O. Oladosu, G. Pridham, and G. Slaney. Grad-cam helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. *Journal of Neuroscience Methods*, 353:109098, 2021.

[70] Y. Zhang, P. Tiňo, A. Leonardis, and K. Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, 2021.

[71] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

Appendix A

Supplementary materials Chapter 2

A.1 Supplementary Figures

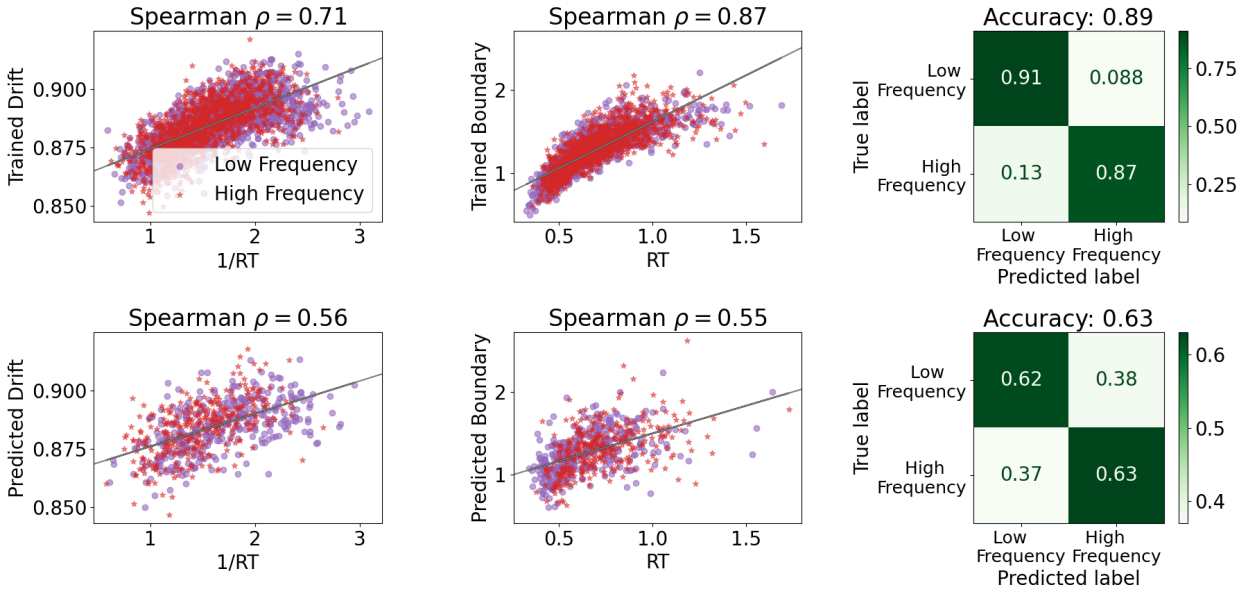


Figure A.1: Model performance from subject s59 in Dataset 2.

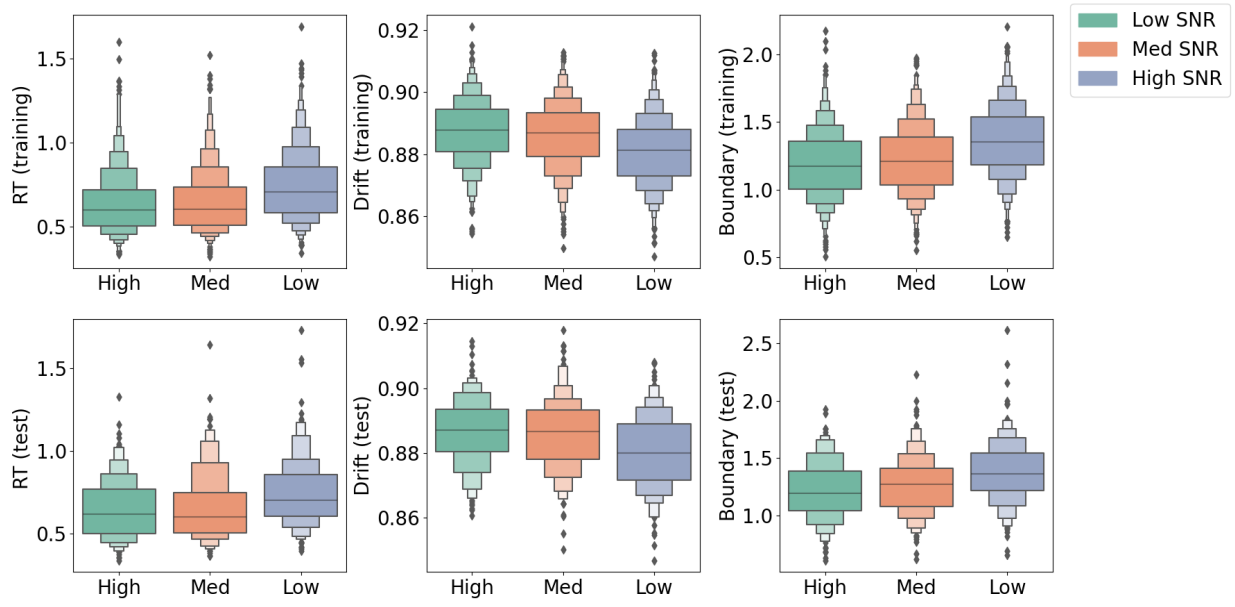


Figure A.2: Distributions of RT, Drift, and Boundary by experimental condition for subject s59 in Dataset 2.

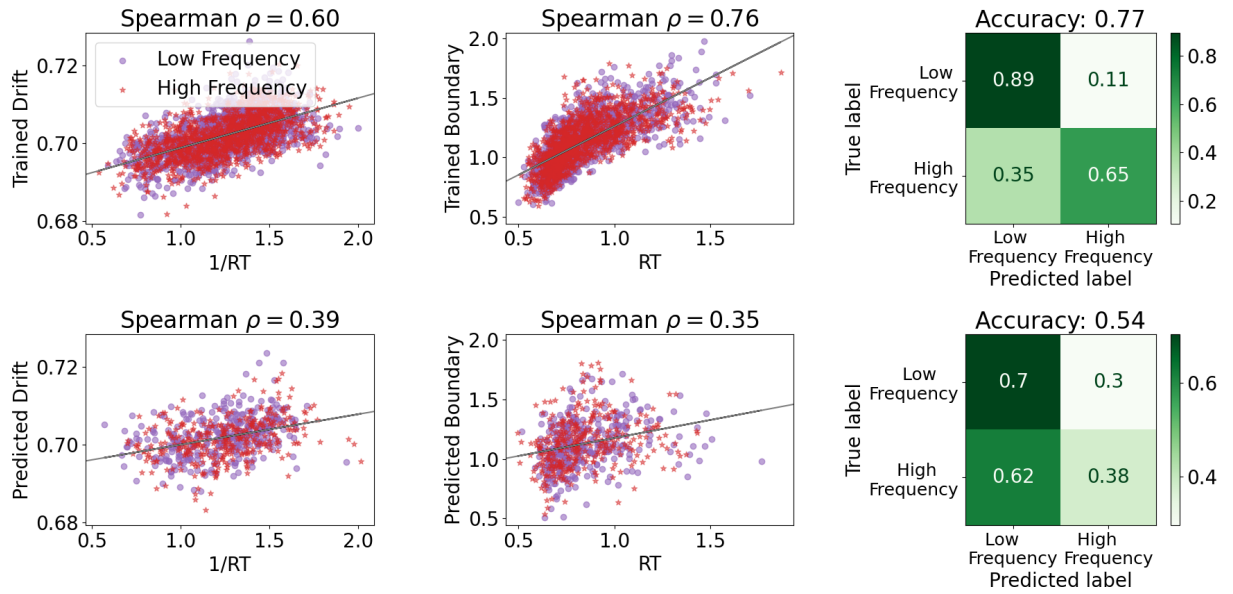


Figure A.3: Model performance from subject s110 in Dataset 2.

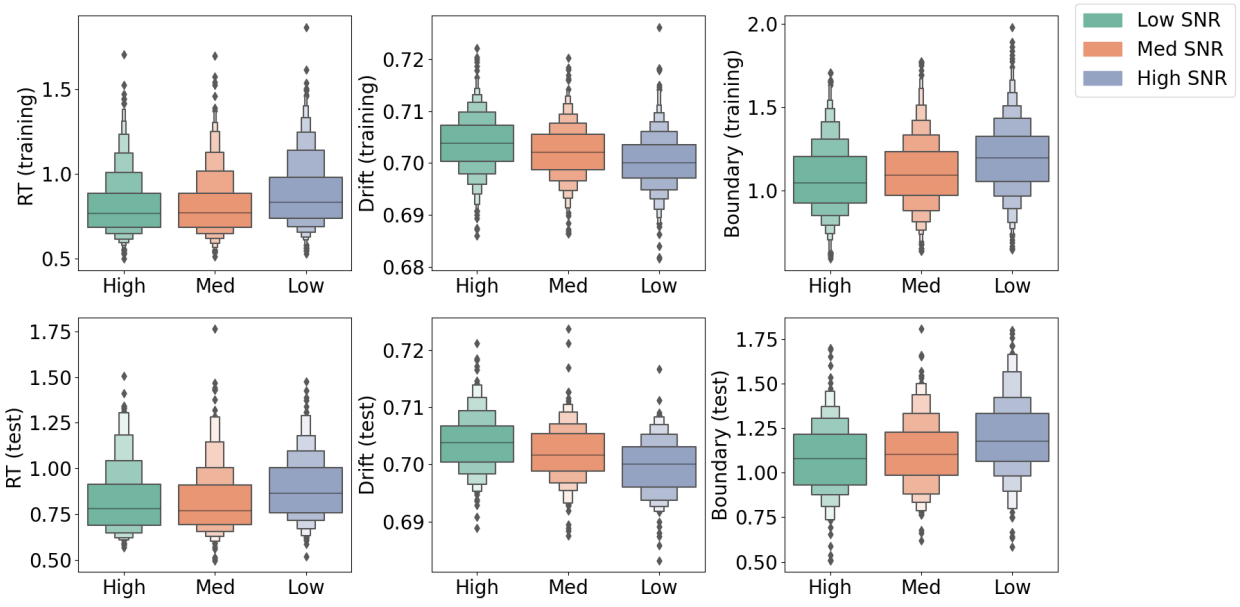


Figure A.4: Distributions of RT, Drift, and Boundary by experimental condition for subject s110 in Dataset 2.

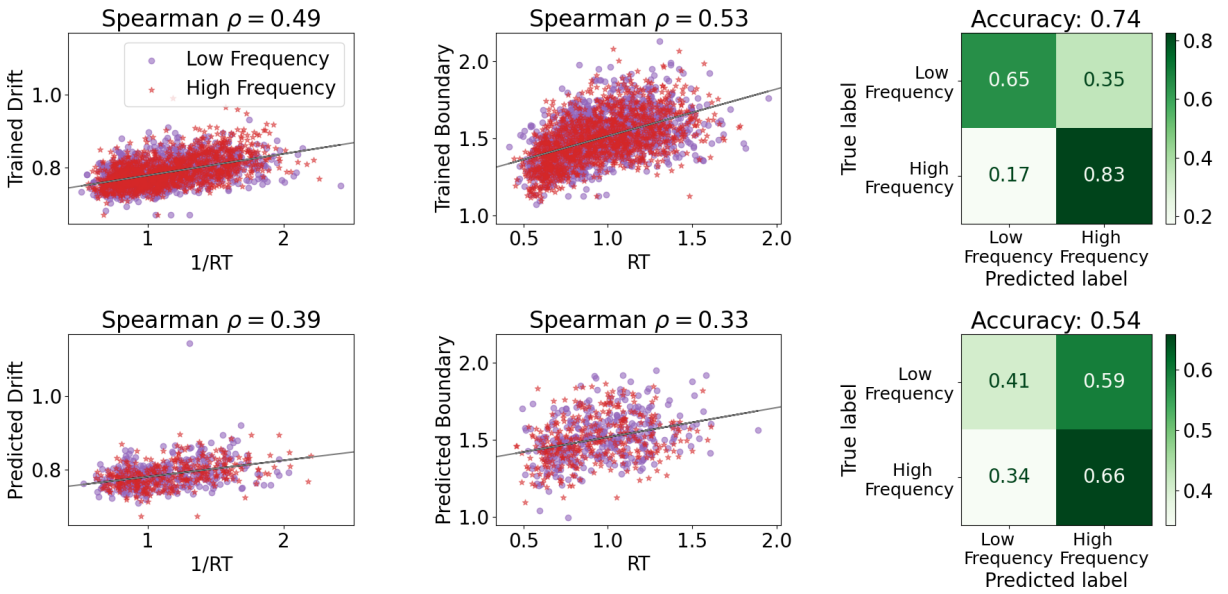


Figure A.5: Model performance from subject s110 in Dataset 2.

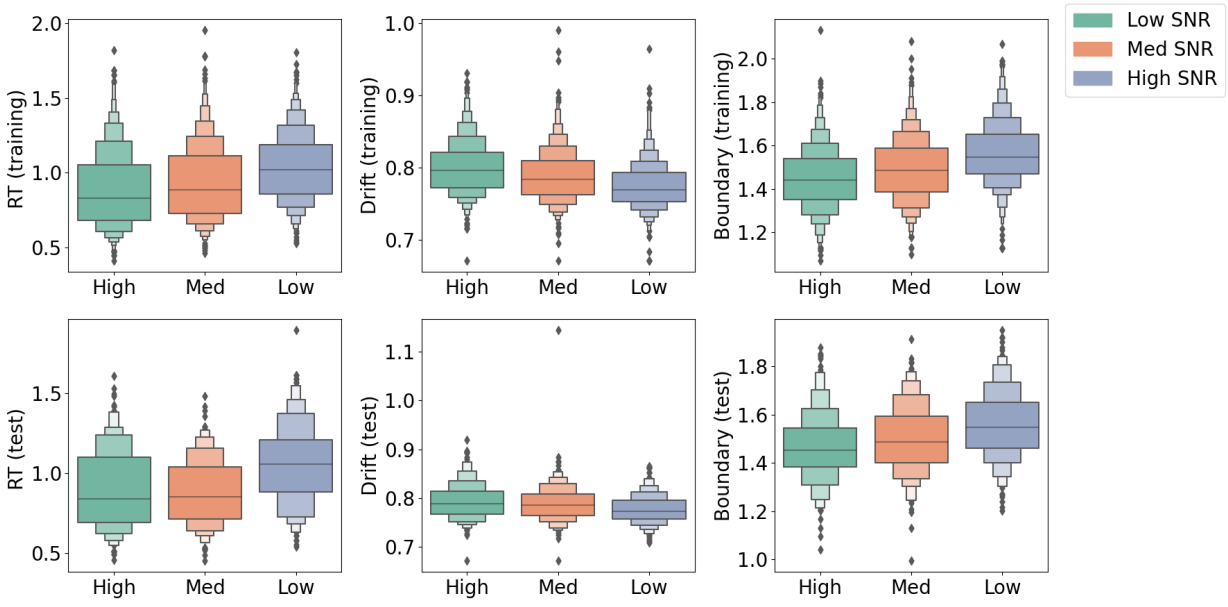


Figure A.6: Distributions of RT, Drift, and Boundary by experimental condition for subject s100 in Dataset 2.

Appendix B

Supplementary materials Chapter 3

B.1 Supplementary Figures

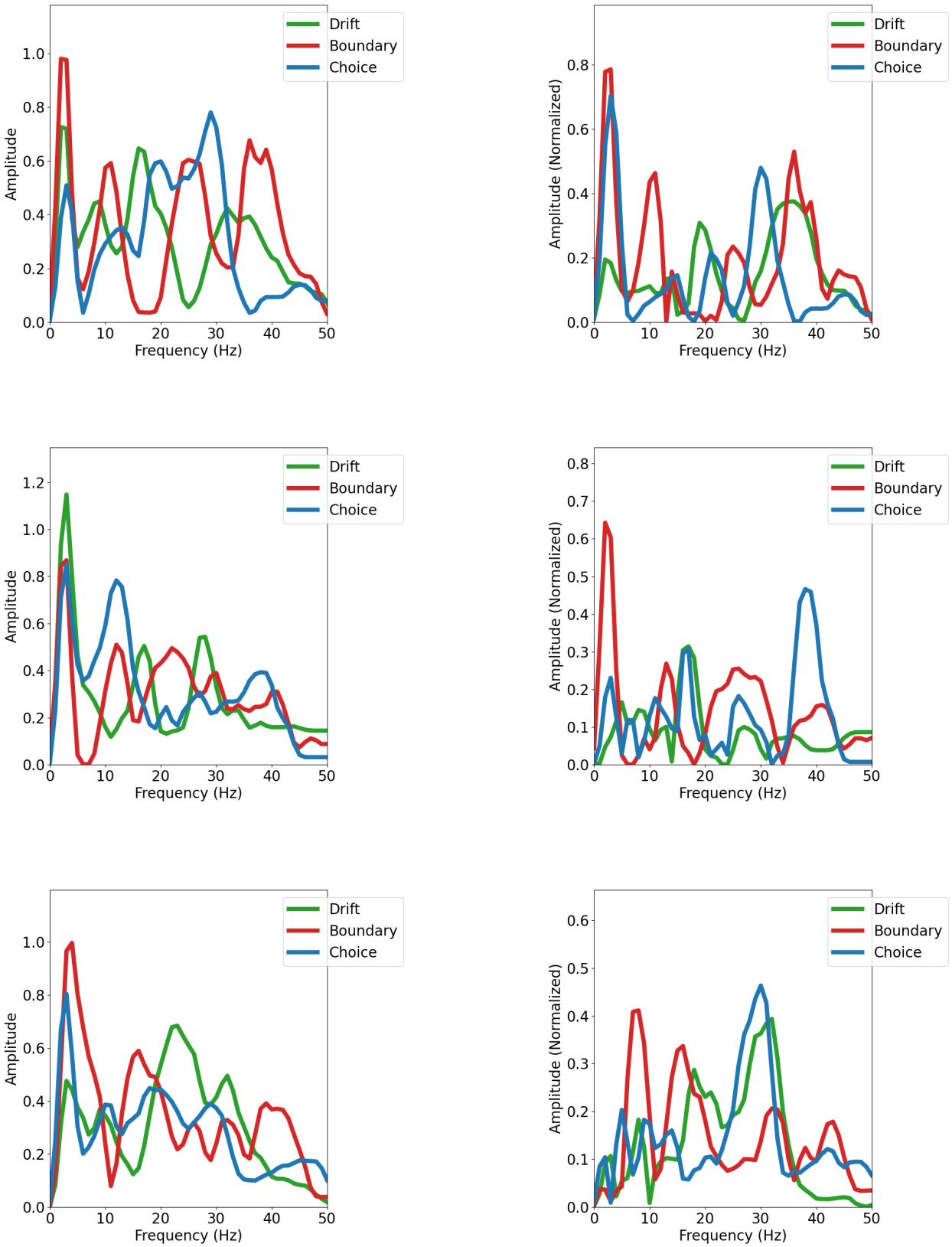


Figure B.1: Results of spectra of sinc kernels of three other subjects from Dataset 2.

Appendix C

Supplementary materials Chapter 4

C.1 Supplementary Equations

C.1.1 Coefficients for Weighted Random Walk Model

Equation for H_0:

$$\begin{aligned} y &= \beta_1 \left(S_1 + S_2 + \cdots + S_k\right) \\ &= \beta_1 \left(kx_1 + (k-1)x_2 + \cdots + x_k\right) \\ &= k\beta_1 x_1 + (k-1)\beta_1 x_2 + \cdots + \beta_1 x_k. \end{aligned} \tag{C.1}$$

Equation for H_1:

$$
\begin{aligned}
y &= \beta_1 S_1 + \beta_2 S_2 + \cdots + \beta_k S_k \\
&= \beta_1 (x_1) + \beta_2 (x_1 + x_2) + \cdots + \beta_k (x_1 + x_2 + \cdots x_k) \\
&= (\beta_1 + \beta_2 + \cdots + \beta_k) x_1 + (\beta_2 + \cdots \beta_k) x_2 + \beta_k x_k.
\end{aligned} \tag{C.2}
$$

Let $X_1, X_2, ...X_n$ be a random sample from a Bernoulli trial with a parameter θ such that $X_i \in -1, +1$. Suppose subjects observed $X_1 = X_1, X_2 = x_2, ..., X_n = x_n$ and they know the Bernoulli trials are biased, with either probability of p or $1 - p$ for getting $X_i = +1$. They are testing between two simple hypotheses:

$$
\begin{aligned}
&H_0 : \theta = p, \\
&H_1 : \theta = 1 - p
\end{aligned} \tag{C.3}
$$

$$
\begin{aligned}
LR &= \frac{L_1 (\theta \mid x_1 \dots x_n)}{L_2 (\theta \mid x_1 \dots x_n)} \\
&= \frac{p^{\sum x_i} \cdot (1-p)^{n - \sum x_i}}{(1-p)^{\sum x_i} \cdot p^{n - \sum x_i}} \\
&= \frac{p^{\sum x_i}}{p^{n - \sum x_i}} \cdot \frac{(1-p)^{n \sum \delta x_i}}{(1-p)^{\sum x_i}} \\
&= p^{2 \sum x_i - n} \cdot (1-p)^{n - 2 \sum x_i} \\
&= p^{2 \sum x_i - n} \cdot (1-p)^{-(2 \sum x_i - n)} \\
&= \left(\frac{p}{1-p} \right)^{2 \sum x_i - n}.
\end{aligned} \tag{C.4}
$$

The Log Likelihood Ratio is thus

$$
\begin{aligned}
LLR &= \log\left(\frac{p}{1-p}\right)^{2\sum x_i - n} \\
&= 2\sum (x_i - n) - n \cdot \log\left(\frac{p}{1-p}\right)
\end{aligned}
\tag{C.5}
$$

C.2 Supplementary Figures

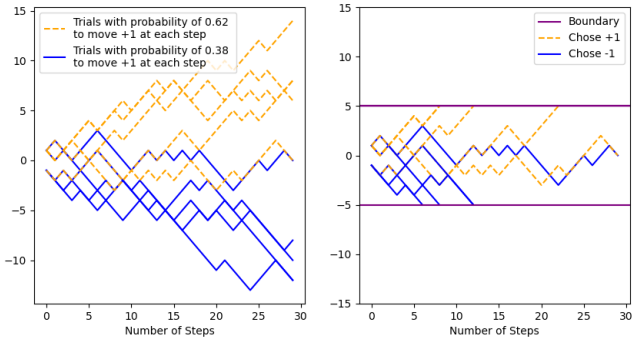


Figure C.1: Left: Evidence chains where each step throughout the walk is generated from a Bernoulli trial with a probability of $p = 0.62$ to move $+1$ (indicated in orange) and $p = 0.38$ to move $+1$ (indicated in blue). Right: Demonstration of termination of random walks if a decision threshold is met (indicated in purple).

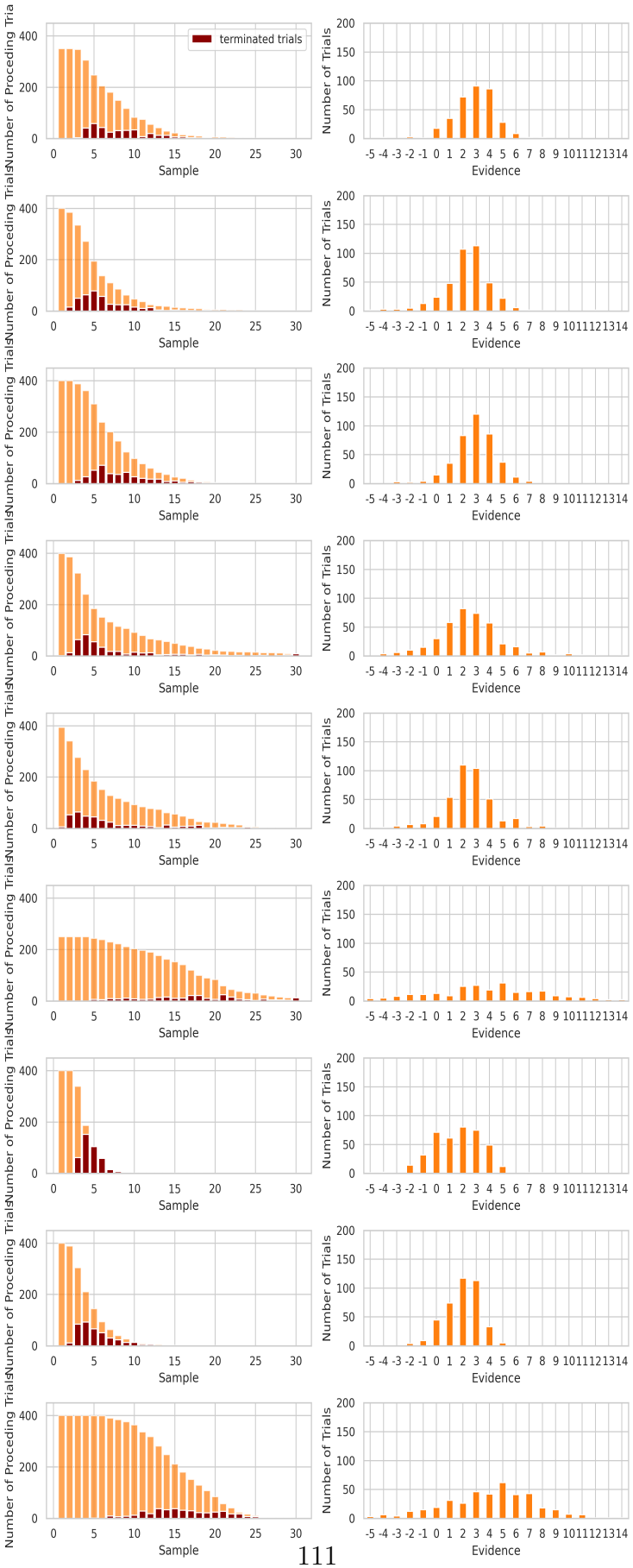


Figure C.2: Trial distribution under 250ms condition by sample (left column) and be evidence (right column) for all subjects.

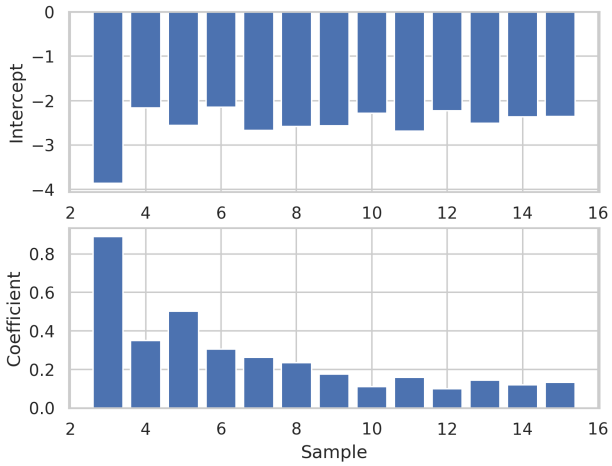
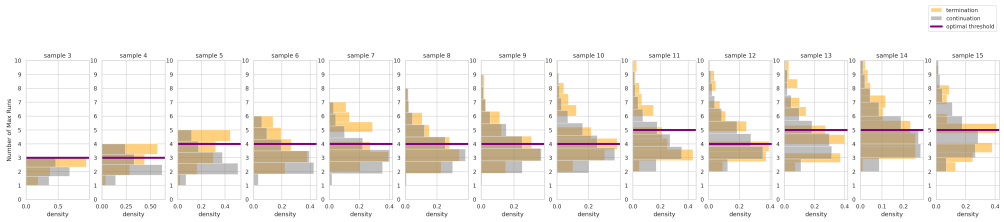
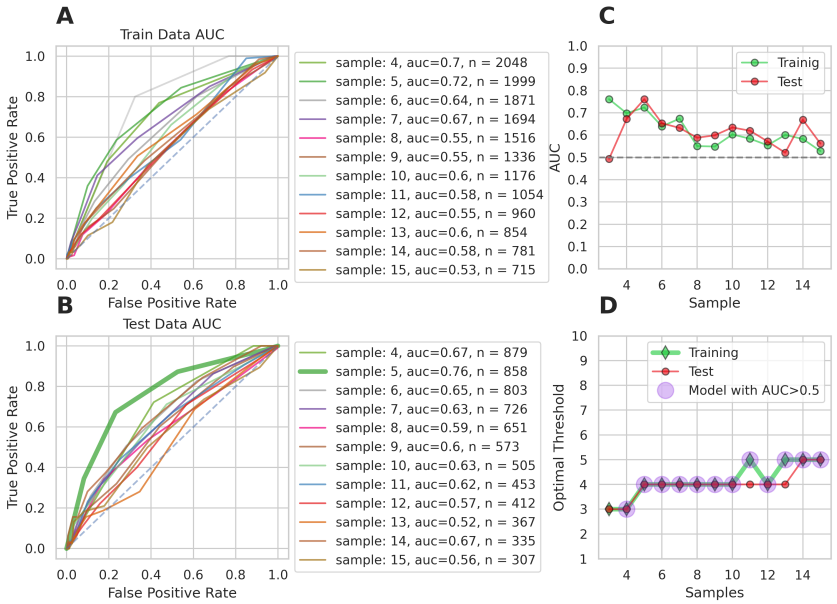


Figure C.3: Visualization of Intercept and Coefficients for a weighted random walk classifier. Coefficients are different at different sample position while intercept remains to be similar values.



(a) Distribution of number of max runs for two classes of trials at different sample position for 100ms condition.



(b)

Figure C.4: Max run model for 100ms condition.

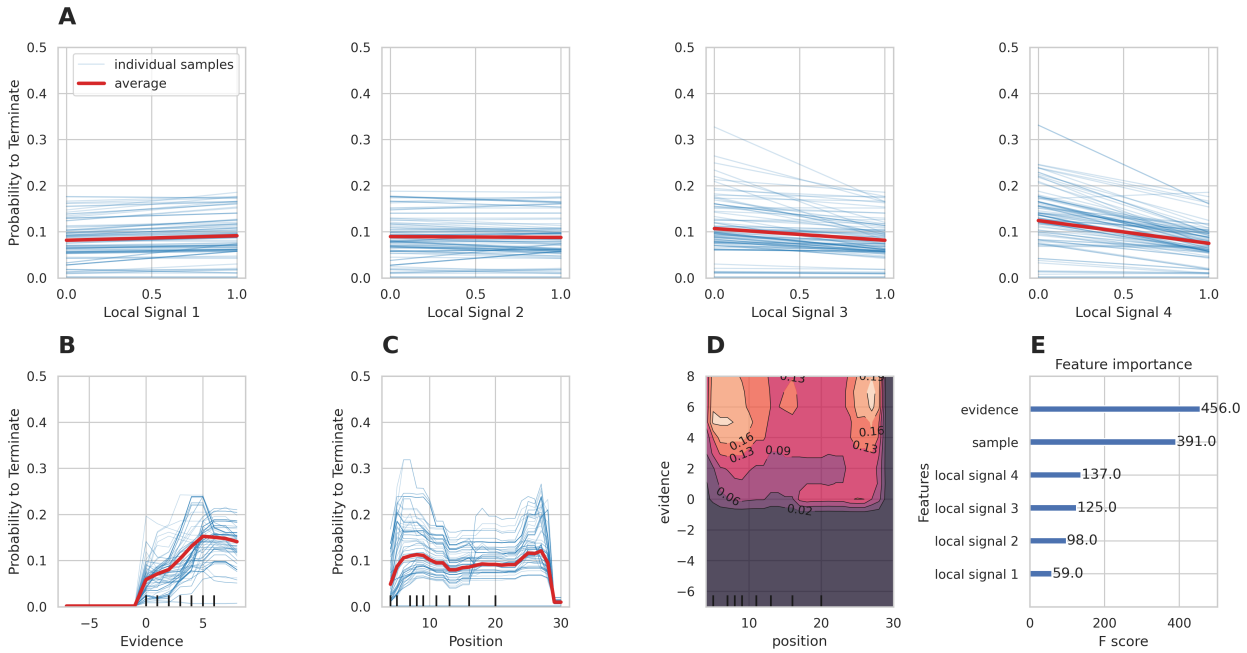


Figure C.5: Partial Dependency Plots (A-D) and feature importance (E) for 250ms condition signed evidence.

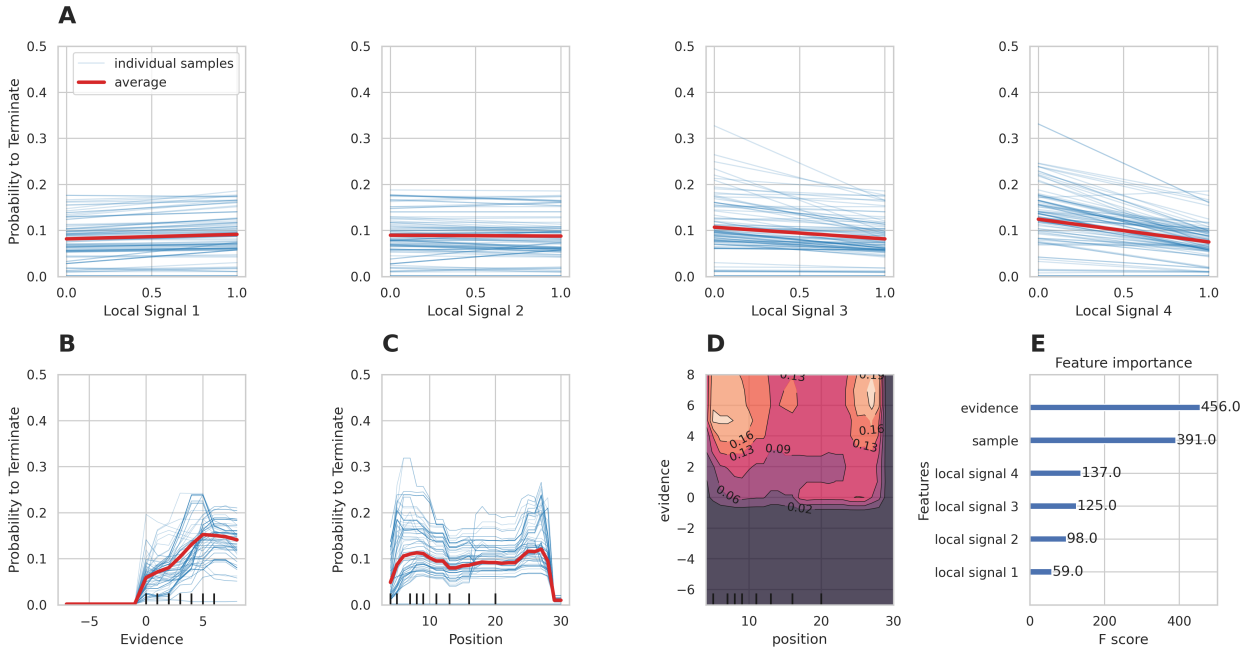


Figure C.6: Partial Dependency Plots (A-D) and feature importance (E) for 100ms condition using signed evidence.