1-1-2012

An SSVEP Brain-Computer Interface: A Machine Learning Approach

Fei Teng
University of Mississippi

AN SSVEP BRAIN-COMPUTER INTERFACE: A MACHINE LEARNING APPROACH

A Dissertation
submitted in partial fulfillment of requirements
for the degree of Ph.D
in the Department of Computer and Information Science
The University of Mississippi

by

FEI TENG

July 2012

ABSTRACT

A Brain-Computer Interface (BCI) provides a bidirectional communication path for a human to control an external device using brain signals. Among neurophysiological features in BCI systems, steady state visually evoked potentials (SSVEP), natural responses to visual stimulation at specific frequencies, has increasingly drawn attentions because of its high temporal resolution and minimal user training, which are two important parameters in evaluating a BCI system. The performance of a BCI can be improved by a properly selected neurophysiological signal, or by the introduction of machine learning techniques. With the help of machine learning methods, a BCI system can adapt to the user automatically.

In this work, a machine learning approach is introduced to the design of an SSVEP based BCI. The following open problems have been explored:

1. *Finding a waveform with high success rate of eliciting SSVEP.*

   SSVEP belongs to the evoked potentials, which require stimulations. By comparing square wave, triangle wave and sine wave light signals and their corresponding SSVEP, it was observed that square waves with 50%

duty cycle have a significantly higher success rate of eliciting SSVEPs than either sine or triangle stimuli.

2. *The resolution of dual stimuli that elicits consistent SSVEP.*

   Previous studies show that the frequency bandwidth of an SSVEP stimulus is limited. Hence it affects the performance of the whole system. A dual-stimulus, the overlay of two distinctive single frequency stimuli, can potentially expand the number of valid SSVEP stimuli. However, the improvement depends on the resolution of the dual stimuli. Our experimental results showed that 4 Hz is the minimum difference between two frequencies in a dual-stimulus that elicits consistent SSVEP.

3. *Stimuli and color-space decomposition.*

   It is known in the literature that although low-frequency stimuli ($<$ 30Hz) elicit strong SSVEP, they may cause dizziness. In this work, we explored the design of a visually friendly stimulus from the perspective of color-space decomposition. In particular, a stimulus was designed with a fixed luminance component and variations in the other two dimensions in the HSL (Hue, Saturation, Luminance) color-space. Our results showed that the change of color alone evokes SSVEP, and the embedded frequencies in stimuli affect the harmonics. Also, subjects claimed that a fixed luminance eases the feeling of dizziness caused by low frequency flashing objects.

4. *A machine learning approach.*

   Machine learning techniques have been applied to make a BCI adaptive to individuals. An SSVEP-based BCI brings new requirements to machine learning. Because of the non-stationarity of the brain signal, a classifier should adapt to the time-varying statistical characters of a single user's brain wave in realtime. In this work, the potential function classifier is proposed to address this requirement, and achieves 38.2bits/min on offline EEG data.

ACKNOWLEDGEMENTS

Thanks Dr. Yixin Chen for teaching me how to be a researcher. Thanks Dr. Scott Gustafson and Dr. Dwight Waddell for providing all the supports. Thanks Dr. Paul Ruth for helping me make the bluetooth work. Thanks Dr. Pamela Lawhead, Aik Min Choong and Chirstopher Reichley for great advises and helps on the design of the amplifier.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number Page

LIST OF TABLES

Table Number Page

Chapter 1

INTRODUCTION

1.1 Introduction

A brain-computer interface translates brain activities into commands that control external devices. The BCI research began in the early 1970s. At that time Jacques Vidal built a first BCI based on visual evoked potentials (134; 135).

This research field was initially motivated by the need for a new type of communication tools for paralyzed or elderly people, whose brains work perfectly but whose muscles do not (60; 18; 143). While they have lost all other communication abilities, the brain might be the last opportunity for them to communicate with the outside world. In recent years, researchers have investigated BCI for healthy people for computer gaming or entertainment applications (100; 72; 62; 101). However, the ability of existing BCIs is very limited and needs to be improved for healthy users (86).

To explain this, the structure of a typical BCI system is described in Section 1.2, and a general performance measure to evaluate BCIs is introduced in Section 1.3. In Section 1.4, we outline the open problems that have been explored to make progress toward a better SSVEP-based BCI.

1.2 A Brain-Computer Interface

A typical BCI system is shown in Figure 1.1. It has a measurement unit to collect the brain signal, a signal processing unit to extract features from the neurological activity and a classification unit to decode “thoughts” into control commands.

There are several measurement techniques in BCI systems. These techniques can be grouped into invasive methods, which place electrodes within the brain, and non-invasive methods, which place electrodes above the skull. Considering that a healthy person usually do not want to implant electrodes into her head, together with the fact that the EEG is technically easier and less expensive to realize (94), the non-invasive Electroencephalography (EEG) is preferred in many BCI designs.

EEG is a very weak electrical signal that needs to be amplified before it can be processed by a software. At the moment we started our BCI study, the commercial EEG collection device in the lab does not provide access to the raw data. Thus we made an EEG recording device.

Our design can be divided into the analog part and the digital part. The analog part amplifies the EEG signal, and is mainly based on “The OpenEEG

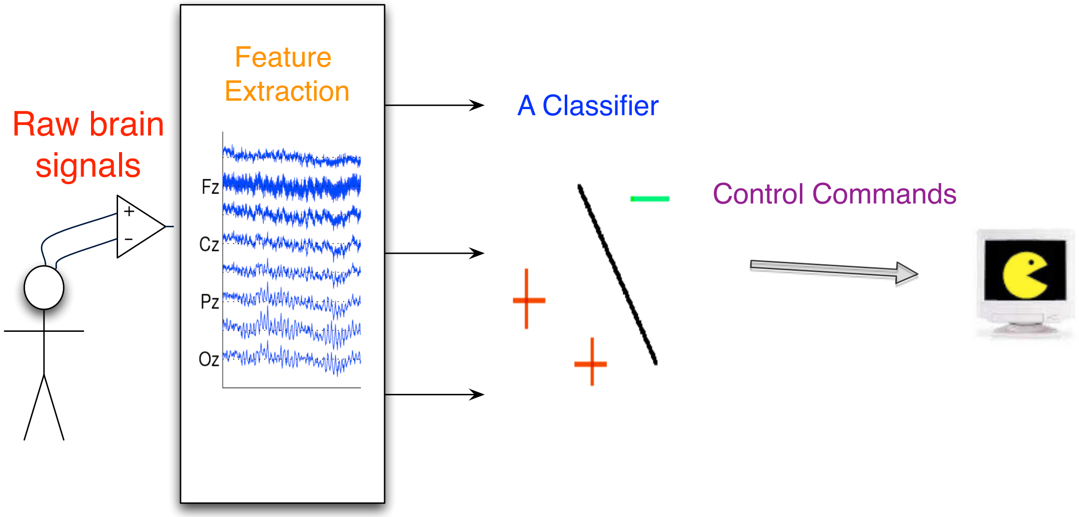


Figure 1.1: A BCI translates brain signals into commands. It collects raw brain activity, processes it into features, and then uses a classifier to decode these features.

Project", an open source project helping people build their own EEG devices for free (as in General Public License) (96). In our design, the digital part uses an ATMEGA8L microprocessor to digitize the amplified signals and control a bluetooth module to send data wirelessly to a computer.

Many neurophysiological features can be detected with EEG. It is beyond the scope of this study to supply a complete review of them. Instead, we present a brief review of three most widely used signals, namely SSVEP, motor -imagery and event related potentials.

- **Event-related potentials (ERP)**

  ERP refers to a positive deflection (P300 peak) appears after the user notices a rare or surprising event (19; 143; 101; 127). An ERP-based

BCI is described in Section 2.1.1.

- **Motor-imagery (MI) related brain activity**

  The activation in the user's motor cortex would increase if she simulates a given action in her brain without actual performance (76; 79). See Section 2.1.2 for an MI-based BCI.

- **Steady state visually evoked potentials (SSVEP)**

  SSVEP refers to signals that are natural responses to visual stimulation at specific frequencies. The user's EEG would contain periodic waveforms of the same frequency as the stimulus (84; 67; 68; 92; 34; 47; 125; 89; 126).

A major difference among these features is that the magnitude of the response varies across the brain, as different brain areas are responsible for different tasks. SSVEP, MI and ERP are usually detected at the visual cortex, the primary motor cortex and the parietal lobe, respectively. These locations are illustrated in Figure 1.2.

Other important differences among these features include the temporal resolution (response time) and user training time. Compared with the P300, whose response time is limited by the rareness of the event to evoke ERP, an SSVEP system detects SSVEP peaks that appears in the subject's brain after around 400ms (108). Compared with the motor-imagery, which requires the

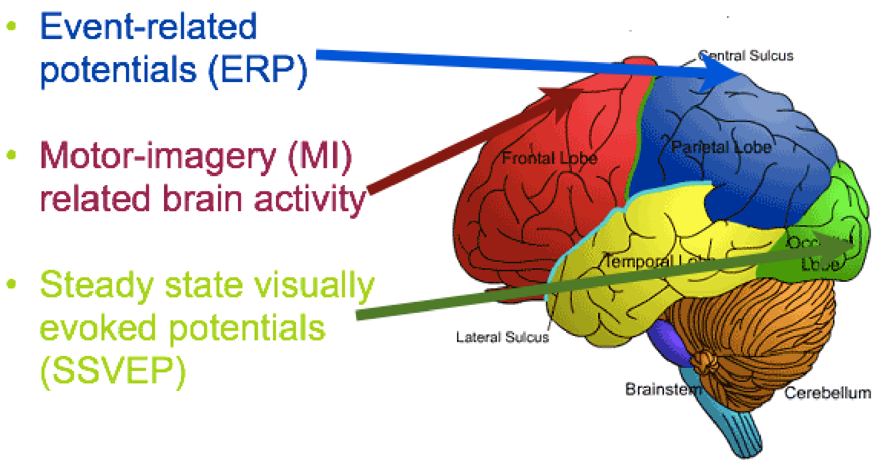

Event-related potentials (ERP)
Motor-imagery (MI) related brain activity
Steady state visually evoked potentials (SSVEP)
Central Sulcus
Parietal Lobe
Frontal Lobe
Occipital Lobe
Temporal Lobe
Lateral Sulcus
Brainstem
Cerebellum

Figure 1.2: A brief schematic of the brain by Young (146). The primary visual cortex is at the back of the brain in the occipital lobe. The primary motor cortex is located in the posterior portion of the frontal lobe. The strongest P300 signal is typically measured at the parietal lobe.

user to be trained beforehand (103), a user's SSVEP is naturally entrained to the frequency of a given light stimulus (108). Only minimal user training is needed to use an SSVEP-based BCI (95). Consequently, among these choices, SSVEP is viewed as a promising electrophysiological source for BCI systems (16). However, SSVEP does not outperform everything. Same as the P300, low-frequency flashing objects (with a frequency lower than 30 Hz) used by SSVEP BCI as stimuli may cause dizziness or even safety hazards linked to photo-induced epileptic seizures (46; 98; 51).

Finally, features extracted by the signal processing component are decoded by the classifier into control commands. Without the help of machine learning techniques, BCIs will have to use predetermined parameters, which require users to adjust themselves to the decision rules (134). A machine learning technique can adapt the system to a user through an initial training session. In addition, a realtime adaptation can be implemented to accommodate the non-stationary property of the EEG signal over time.

1.3 The Bit Rate of a BCI

Bit rate is a general performance measure to evaluate BCIs. Let us consider two BCI systems, BCI_1 and BCI_2. Assume that BCI_1 can choose one option out of 20 possible selections with an accuracy of 90%. BCI_2 can make a binary decision with an accuracy of 95%. If these two BCIs are used to pick a symbol from a set of 20 objects, BCI_1 will finish the task with one successful trial while BCI_2 will need to make five consecutive correct decisions. Therefore, BCI_1's success rate is still 90% while BCI_2's drops to $77.4\%(= 95\%^5)$. This fact suggests that the bit rate of BCIs needs to take into consideration the accuracy, the number of possible selections and the number of decisions per minute. Thus, the bit rate R of a BCI is computed as the product of the number of bits per decision (B) and the average number of decisions per minute (142). The number of bits per decision B is given by

$$B = \log_2 N + P \log_2 P + (1 - P) \log_2 \frac{1 - P}{N - 1}, \tag{1.1}$$

where N is the number of possible selections, and P is the accuracy.

1.4 Outline of this Dissertation

The goal of this work is to find a better stimulus and a machine learning approach, which introduce adaptiveness, accuracy and speed to an SSVEP BCI. Open problems that have been explored are illustrated in Figure 1.3. From the perspective of the stimulator, research was conducted on finding an effective stimulus (in Section 3), finding the resolution of dual stimuli (in Section 4) and what help color-space decomposition can provide to the design of a visually friendly stimulus (in Section 5). For the classification unit, the Potential Function Classifier (in Section 6) is designed to process the neurological features and adjust to the changes in realtime. These research topics are briefly described below.

- **Finding a waveform with high success rate of eliciting SSVEP.** Because the EEG is always mixed with background noises, the efficacy of an SSVEP-based BCI system relies heavily on the signal-noise ratio. Intuitively, SSVEP will be detected much easier and faster if the signal-noise ratio is high. The faster an SSVEP is identified, the more promptly a BCI system can respond correctly, hence a higher information throughput (1). Square wave (with different duty cycles), triangle wave, and sine wave were compared in Section 3 for their success rate of

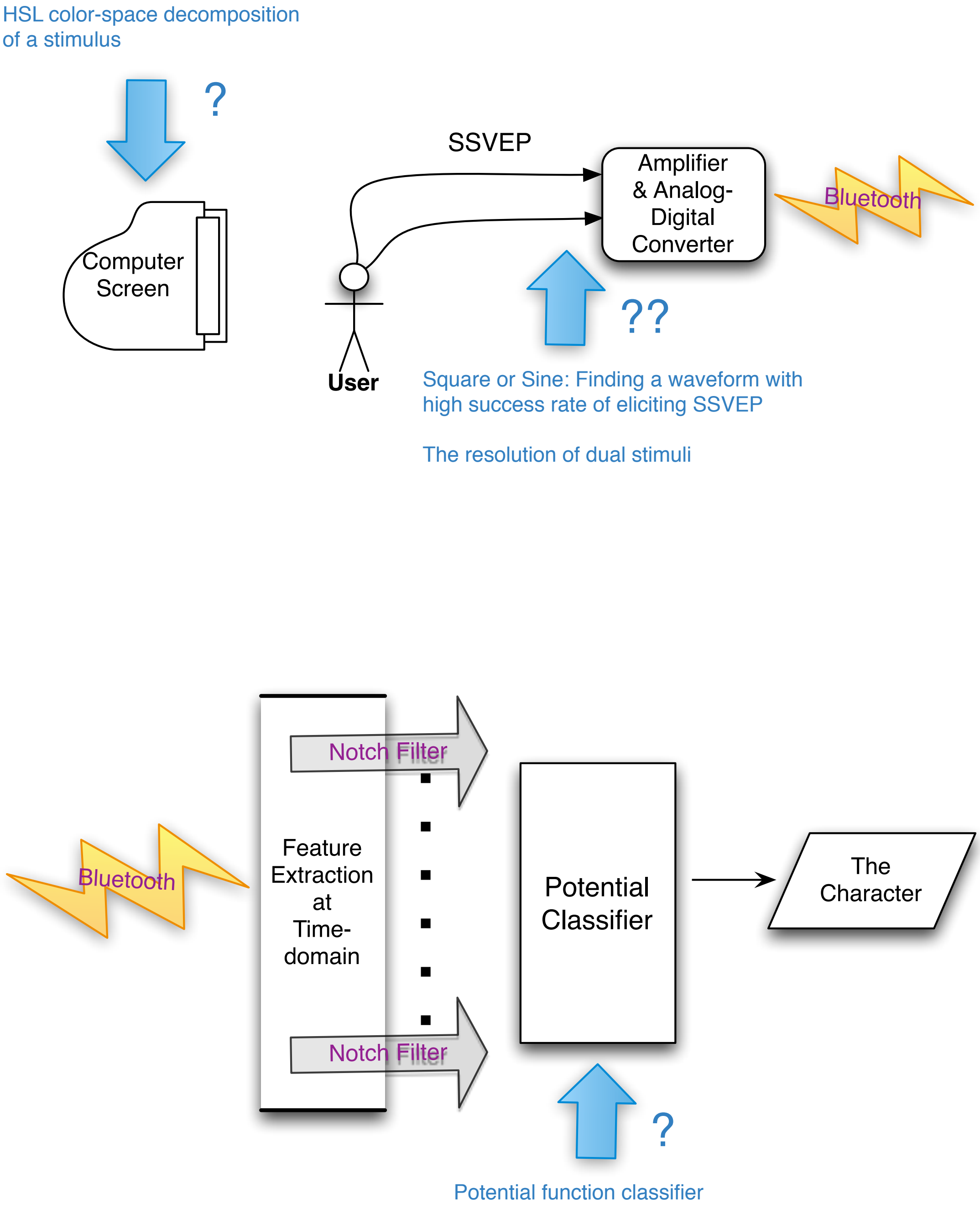

HSL color-space decomposition of a stimulus
?
Computer Screen
SSVEP
User
Amplifier & Analog-Digital Converter
Bluetooth
??
Square or Sine: Finding a waveform with high success rate of eliciting SSVEP
The resolution of dual stimuli
Bluetooth
Feature Extraction at Time-domain
Notch Filter
Notch Filter
Potential Classifier
The Character
?
Potential function classifier

Figure 1.3: Open problems investigated in this work are marked by question marks with arrows pointing to where they occur. A computer screen generates visual stimuli. The Amplifier collects the EEG signal and uses Bluetooth to send it to software-notch-filters and a potential function classifier, which outputs the character that the user wants to input.

eliciting SSVEP. It was observed that the choice of a square wave, triangle wave or sine wave light signal visual stimulus affects the strength of the elicited SSVEP. Square wave is with the highest success rate of eliciting SSVEP. Also, researchers observed that a stimulus at frequency f can elicit SSVEP not only at f, but also harmonics at $2f$, $3f$, or sometimes even at higher orders (17; 71). This seems to suggest that harmonics may be used in detecting the stimulating frequency. However, in order to take advantage of the harmonics in the design a BCI system, the following question needs to be addressed. Are the harmonics in SSVEP elicited by the fundamental frequency, i.e., f, or by the artifacts of the stimulus? It was observed that square waves with 50% duty cycle have a significantly higher success rate than either sine or triangle stimuli, and the success rate of getting harmonics is positively correlated with the strength of the artifacts in a stimulus.

- **The resolution of dual stimuli that provides consistent SSVEP.** It was reported that SSVEPs could be elicited in the range of 4–100Hz (106; 59; 50), while the strongest response was observed in the range of 5–20Hz (34; 47; 68). This fact limits the number of valid stimuli, hence affects the performance of an SSVEP-based BCI. In order to provide more stimuli options within $5 - 20$Hz, dual stimuli were proposed in the literature. For example, Cheng et al. (35) used multiple color stimuli to deliver two frequencies simultaneously. However, no research has been done on the resolution of the dual stimuli, i.e., what is the resolution of dual stimuli that provides consistent SSVEP? We use dual stimuli, generated by two sine waves on a light emitting diode (LED) to study the resolution needed for consistent responses (Section 4). Our experimental results showed that 4 Hz is the minimum difference between two frequencies.

- **Stimuli and color-space decomposition.**

  It is known that low-frequency stimuli (< 30 Hz) tend to elicit strong SSVEP but may cause safety hazards linked to photo-induced epileptic seizures (46; 47). Arakawa et al. (6) showed that both luminance and color patterns elicit SSVEP. However, in their experiments, the luminance was not completely isolated from the color. In this study, we explored a stimulus in the HSL (Hue, Saturation, Luminance) color-space. Stimuli were designed with a fixed luminance component and variations in the other two dimensions in the HSL space. We demonstrate this type of stimulator elicits SSVEP at the fundamental frequency, and the embedded frequencies affect harmonics. Furthermore, all subjects in our experiment felt that this color-space decomposition makes low-frequency stimuli more visually friendly than ordinary luminance stimuli.

- **A machine learning approach.**

  Machine learning techniques adapt the BCI to a subject. Considering the dynamic nature of EEG signals of one user, i.e., the structure of the data may vary over time, the classifier needs to adapt to the changes in realtime. In order to address this problem, we propose the Potential Function Classifier in Chapter 6.3. This algorithm has been tested with datasets from the UCI Machine Learning Repository and offline EEG datas.

Chapter 2
BACKGROUND

BCI is an interdisciplinary research area. Without understanding some important facts from neurophysiology, one cannot see the options and challenges in this field.

This chapter introduces neurophysiological background knowledges. In a BCI, the raw brain signals are processed by feature extraction methods, which are introduced in Section 2.1. Among brain signals, the P300, MI and SSVEP are reviewed with example applications in Section 2.1.1, Section 2.1.2 and Section 2.1.3, respectively. Methods to deliver accurate stimuli using a computer screen are shown in Section 2.1.3. Finally, machine learning techniques that have been deployed in BCIs are discussed in Section 2.1.4.

2.1 Feature Extraction

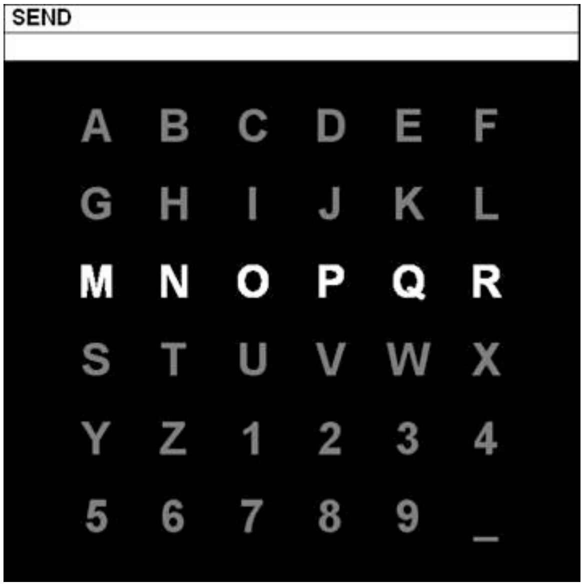
In most current BCI systems, features used were motivated from neurophysiological observations. For example, SSVEP BCIs are based on the fact that the users EEG would contain periodic waveforms of the same frequency as the stimulus, thus they use frequencies as their features (88; 90). Also because of the mechanism P300 and MI occur, BCIs based on them take P300 peak at the parietal, or MI peak lobe at the primary motor cortex, as features.

In applications where the frequency range of interest is given a priori, Fast Fourier Transform (FFT) is widely applied to extract discriminative features in the frequency domain (85; 77; 102; 111). Wavelets transform is another technique that combines spatial and frequency information (45). In time to frequency domain transforms, a high resolution in the frequency-domain can only be achieved using a long time window, i.e., a long data sequence in time-domain. FFT needs x seconds to achieve a $\frac{1}{x}$Hz resolution, for example, one second of data to achieve one Hz resolution. Considering that the SSVEP appears about 400ms after the stimulation (108), and SSVEP BCIs usually use 1Hz as the difference between stimuli, time-domain features were explored, to extract SSVEP peaks without waiting for a full second (for 1Hz resolution in FFT). For instance, Li et al. (78) used bandpass filters to extract independent features. Kalman filter was used by Neuper's team (93) and Gage's team (48).

Many feature extraction methods have been proposed to increase the signal to noise ratio. Among them, Independent Component Analysis (ICA) receives wide attention (120; 65). ICA is commonly used when multiple EEG reading are available. It interprets each channel of the recorded EEG data as a linear combination of n unknown but independent sources, then reconstructs the signals. Principal components analysis (PCA) is another technique that was used in (45; 48). It decomposes the EEG data into mutually orthogonal channels. In some applications, signal to noise ratio can be improved by a differential feature extraction approach. For example, common spatial patterns (CSP) are computed in motor-imagery systems (78; 43) to identify the source of neurophysiological events.

2.1.1 P300

P300 is popularly used for building BCI spellers (19). P300 peak is a positive deflection appears after the user notices a rare or surprising event. For example, a strong P300 peak is detectable near the parietal lobe when letter A is noticed by a user waiting for A but has been shown letter B for some seconds. Figure 2.1 shows a P300 interface used in the Brain-Computer Interface Laboratory at East Tennessee State University. A substantial but unsolvable problem of a P300 is that it is slow to make a P300 peak appear, thus affects the performance of BCIs based on them. This is because the event driving a P300 peak has to be rare enough, e.g., something shown once a second is not rare. Researchers usually improve P300 performance by using a relatively large number of possible selections (36 in (117; 42; 64)).



SEND

A	B	C	D	E	F
G	H	I	J	K	L
M	N	O	P	Q	R
S	T	U	V	W	X
Y	Z	1	2	3	4
5	6	7	8	9	_

Figure 2.1: A P300 interface from the Brain-Computer Interface Laboratory at East Tennessee State University. This P300 system highlights the characters randomly, and waits for the P300 peak that appears in the user's brain after she notices the wanted character being highlighted.

In literature, both online and offline P300 BCIs were explored. For example, online systems were developed in (104) and (141) with bit rates (calculated by Eq.(1.1)) of 9.48 bits/min and 10.88 bits/min, respectively. Offline systems reported in (42) achieved 20.1 bits/min, in (8) 2.65 bits/min, in (37) 5.64 bits/min and 23.75 bits/min in (117). Kaper et al. (64) showed the most promising result of 84.7 bits/min, as a special case on a single subject.

2.1.2 Motor Imagery

The activation in the user's motor cortex would increase if she simulates a given action in her brain without actual performance. This activity is called the motor-related brain activity. For example. if a user imagines to

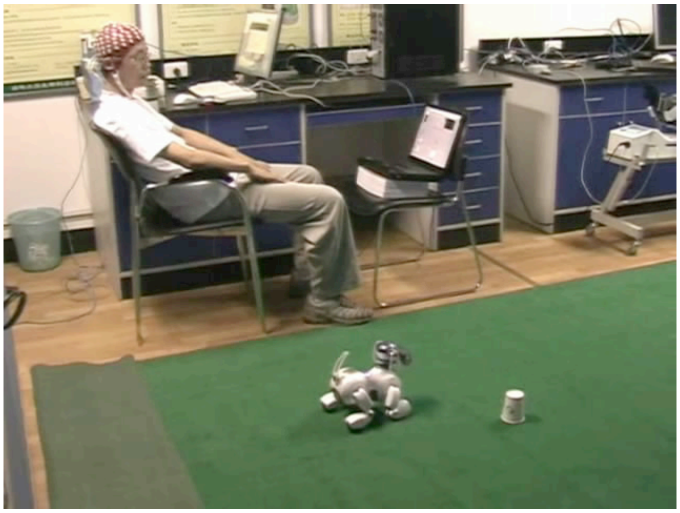


Figure 2.2: A motor-imagery system from the Tsinghua University. This system detects the activation in the user's motor cortex when the user simulates a given action in his brain, and translates this activity into commands to control the robot dog on the floor.

raise her left arm, the activation in her motor cortex will increase. Even better, this increase is distinguishable from imagining raising her right arm. Figure 2.2 shows the motor-imagery related brain activity system used in the Tsinghua University. An problem of MI BCI is that it is not intuitive and takes time (days or even weeks) to learn to imagine movement, thus to use the system (143).

Conversely to the P300, motor imagery systems can make a decision fast (22) but lack of possible selections (2 in (22; 136), 3 in (23; 25), 4 in (24)). In different applications, Blankertz et al. achieved bit rates of 23 bits/min, 6-15 bits/min, 15-35 bits/min and 12-35 bits/min in (21), (22), (23), and (25), respectively. A 4.3 bits/min was reported in an online system (136). Impressive MI-based BCIs were shown in "The BCI Competition III", in which the top three teams achieved 47.4 bits/min, 40.4 bits/min and 37.8 bits/min (24).

2.1.3 SSVEP

SSVEP refers to signals that are natural responses to visual stimulation at specific frequencies. The user's EEG would contain periodic waveforms of the same frequency as the stimulus (125; 89; 126). Compared to other neurophysiological features in EEG, SSVEP holds the advantage of short/no training time - a user's SSVEP is naturally entrained to the frequency of a given light stimulus.

Figure 2.3 shows an SSVEP system in the Institute of Automation, University of Bremen.

At present, no general conclusion on SSVEP stimuli can be drawn because many conditions have not been tested and variables interact with each other. In literature, the type of stimulation, the frequency, the luminance, the color, the embedded frequencies and the subject's attention have been considered as attributes affecting SSVEP.

- **Stimulation Type**

  Several types of SSVEP visual stimulators have been introduced and used for years (36; 106; 94; 47), based on the fact that both luminance and color patterns elicit SSVEP, while the power of the SSVEP response

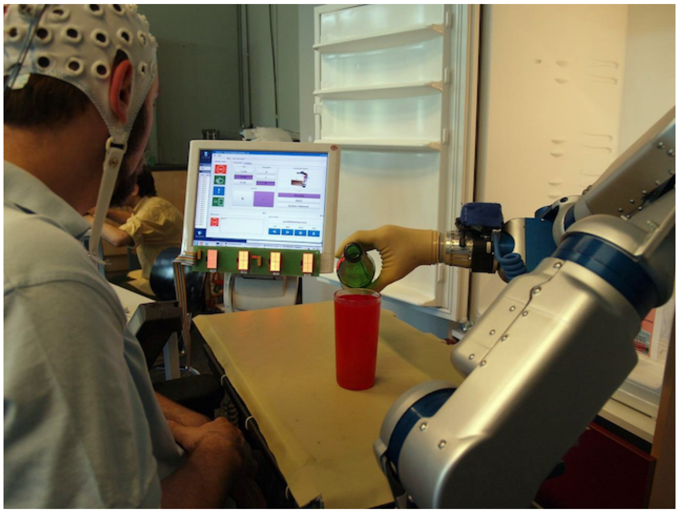


Figure 2.3: An SSVEP system at the University of Bremen. The user focuses on light sources blinking with different frequencies (the light-emitting diodes at the bottom of the screen). The frequency that is currently in the focus lets the neurons in the visual cortex of the brain synchronize with the same frequency. By detecting the frequency at which the user is looking, the system lets him control the robot arm.

is affected by them (107; 6). In 1989, Regan claimed that the SSVEP response for light stimuli was larger than that for pattern reversal in (106). Wu confirmed this statement by showing that SSVEP response elicited by an LED was larger than that by a rectangle stimulus on a computer screen. This explains why the bit rates of BCIs using LED stimuli are usually higher than those of BCIs using computer screens (29; 137). However, from the viewpoint of implementation, a computer screen is preferred as this type of stimulation mainly relies on software development and no hardware modification is necessary. Furthermore, researchers can set any attributes to any possible value of the stimuli on screens, no matter the luminance, contrast, color, saturation et al., compared to LEDs, over which no accurate control could be achieved[1]. It is also noteworthy that the PC hardware and operating system may affect the accuracy of the stimulation frequency on a screen (62). Sugiarto and Sutoyo claimed that DirectX, OpenGL and Matlab are effective in implementing an accurate stimulus with a computer screen (123; 124).

- **The Frequency**

  The stimulus frequencies used in SSVEP research are usually categorized into three bands: low (1-12Hz), medium (12-30Hz) and high (30-60Hz). SSVEP is strongest in the visual cortex, when the stimulus is flashing at around 15Hz (98).

- **The Luminance**

  Arakawa et al. showed that both luminance and color elicit SSVEP (6). However, in those experiments, the luminance was not completely isolated from the color.

- **The Color**

  In1966, Regan found out that red, yellow, and blue light stimuli, together with the chosen frequency, affect SSVEP responses (107). In 2001, Cheng's group first considered the color of the stimulus as a source

[1] Note that the stimulus frequency on a computer screen is restricted by the refresh rate of the screen.

of frequency instead of on/off lights (35). After them, many researcher explored the use of different colors, in which red, white and green are frequently used. Two BCI labs demonstrated that the best-performing color is green (49; 97). But no comparison has been done to show how color influences the SSVEP performance. In Section 5, we completely isolated luminance and color to check if color patterns elicit SSVEP.

- **The Embedded Frequencies**

  It is known that SSVEP has the same fundamental frequency as the visual stimulus. If two frequencies were delivered simultaneously, SSVEP would have both (89). Many methods have been used to embed frequencies in a single stimulus, for example, different colors (35), or the luminance in LEDs (126). In Section 3, we conclude that frequencies other than the fundamental frequency in a square wave may elicit SSVEP. In Section 4, we conclude that two embedded frequencies in an LED have to be at least 4Hz apart to elicit consistent SSVEP.

- **Attention on the stimuli**

  It has been proved that the SSVEP strength is strongly influenced by attention (91). If a subject moves her attention to something else than the flashing stimulus, no matter proactive or passive, the power of SSVEP will decrease. Most researchers solve this problem by moving the flashing objects along with the controlled elements (81; 132). Specifically, if two stimuli were presented, the SSVEP of the ignored one would decrease and the SSVEP of the selected one would be enhanced (112). Sometimes it is not favorable as we want to take the advantage of multiple stimuli. This problem could be solved by using a single flashing object to deliver multiple frequencies (126).

Despite dizziness or even safety hazards linked to photo-induced epileptic seizures caused by low-frequency flashing objects (with a frequency lower than 30 Hz) (46; 51), SSVEP-based BCIs achieve promising information transfer rates, with flexible number of possible selections, which may vary from 4 in (97), 11 in (137; 34) to 30 in (29). Interestingly, the four-class SSVEP achieved an impressive 51.5 bits/min (an average over 11 subjects), comparing with 11-class SSVEPs' 42 bits/min (137) and 27.15 bits/min (34), or a 17.4 bits/min with 30 classes (29). Promising results were also reported in (138) as 29-63 bits/min and in (63) as 66.7 bits/min.

2.1.4 Machine Learning Techniques in BCIs

Several groups applied machine learning techniques to BCI to adapt the system to users. For example, quadratic discriminant analysis (QDA) was implemented in (93; 115). It achieves the optimality if the data is Gaussian distributed. Linear discriminant analysis (LDA) is used in (93; 115; 43; 70).

It is similar to QDA with a stronger assumption that each class has a same covariance. Regression techniques are applied in (77; 144; 83; 48; 120) to find an optimum function mapping the data to their class labels. Fatourechi (45) and Kirby (70) tested the k-nearest-neighbors (KNN) classifier, which assigns an unknown data point to the majority class of its k-nearest neighbors. In (111), support vector machines (SVM) were used. SVM separates data with hyperplanes by maximizing the margin. There are also works using neural network classifies (4).

Chapter 3

AN EFFECTIVE STIMULUS

As shown in Figure 1.3, this study works toward a good SSVEP BCI system. Obviously, it needs an effective stimulus to evoke distinguishable SSVEP peaks for further processing. In this chapter, we find an effective stimulus, defined as a stimulus with high success rate of eliciting SSVEP.

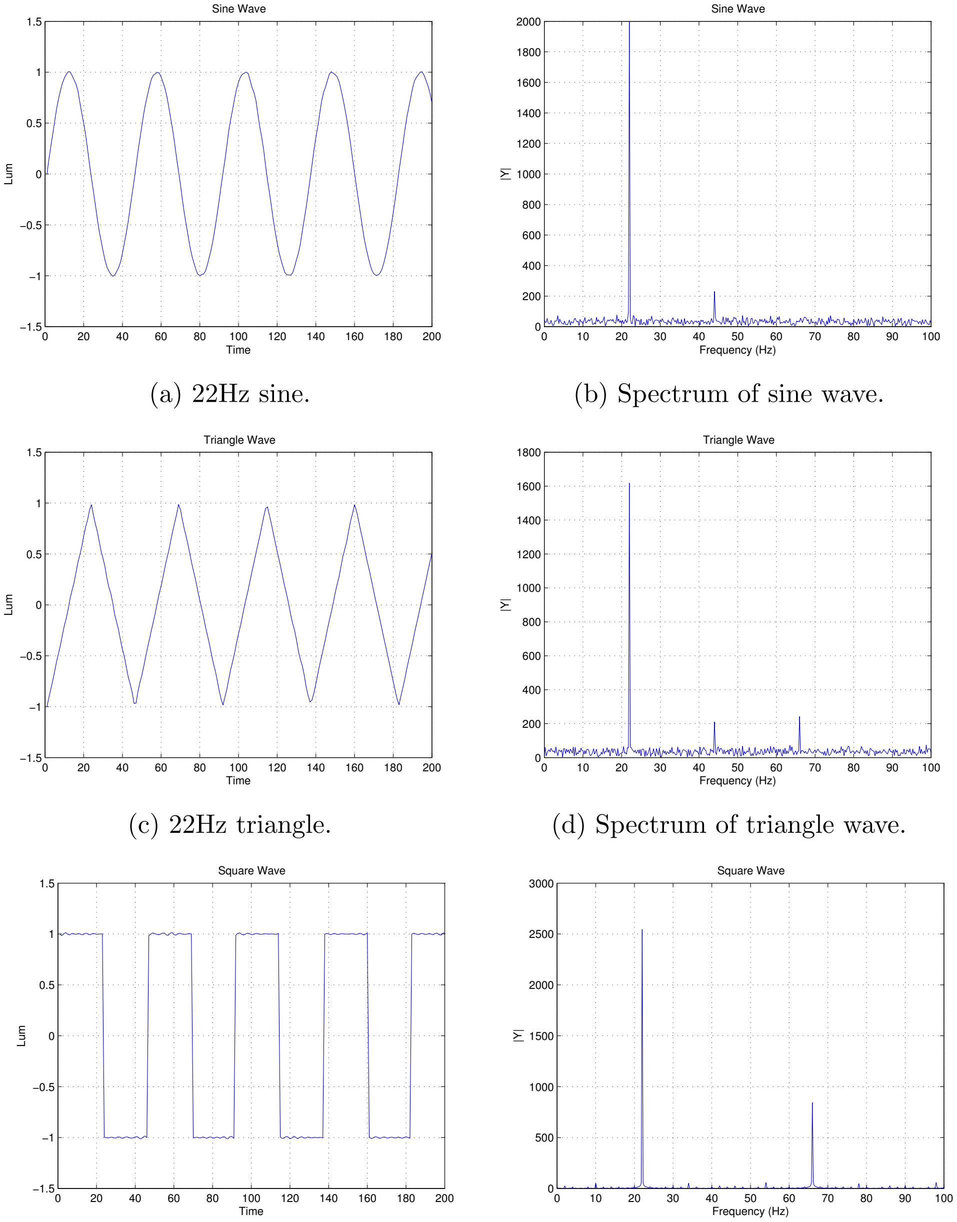
3.1 Methodology

A stimulus is a object flashing at a certain frequency, while the frequency could be delivered as a sine wave, or a square wave. If different stimuli perform differently at evoking SSVEP, among square wave, triangle wave and sine wave light signals, which one has the highest success rate of eliciting SSVEP? Furthermore, from a signal perspective, the commonly used flickering stimulus is a periodic square wave with 50% duty cycle. Its spectrum contains nonzero Fourier components at $\pm(2k-1)f$, $k = 1, 2, \cdots$. Researchers observed that a stimulus flickering at frequency f can elicit SSVEP not only at frequency f, but also the harmonics at $2f$, $3f$, or sometimes even at higher orders (17; 71). Therefore, under a square wave stimulus, the cause of a $3f$

harmonic in SSVEP is unclear, i.e., Are the harmonics in SSVEP elicited by the fundamental frequency, i.e., f, or by the artifacts of the stimulus? We explore the SSVEP responses of three periodic stimuli, square waves with different duty cycles, triangle wave, and a sine wave, to answer the above two questions.

Three types of periodic stimulus were used in the experiments: square wave (with duty cycle $\tau \in (0, 1)$), triangle wave, and sine wave. If we define the relative strength of the k-th harmonic frequency with respect to the fundamental frequency as $r(k) = \left|\frac{G_k}{G_1}\right|$ where G_1 and G_k are the Fourier coefficients for the fundamental frequency and the k-th harmonic frequency, respectively, it is straightforward to show that $r_{\text{sine}}(k) = 1$ for $k = \pm 1$ and 0 otherwise; $r_{\text{triangle}}(k) = \left[\frac{\pi}{2}\text{sinc}\left(\frac{k\pi}{2}\right)\right]^2$; $r_{\text{square}}(k) = \left|\frac{\text{sinc}(k\tau)}{\text{sinc}(\tau)}\right|$. Clearly, in theory there are no harmonic frequencies in a sine wave. In a triangle wave, the harmonic frequencies only exist for odd k. Its magnitude is proportional to $\frac{1}{k^2}$. For a square wave with duty cycle $\tau = 0.5$, there are also no harmonics for even k. The magnitude of odd harmonics is however proportional to $\frac{1}{k}$, i.e., stronger than that of a triangle wave. Note that the magnitude of harmonics of a square wave depends on its duty cycle, e.g., $r_{\text{sine}}(2) > 0$ for $\tau \neq 0.5$.

The above wave forms were rendered using an LED. In order to generate sine and triangle luminance signal, the LED needs to work in its lin-



(a) 22Hz sine.

(b) Spectrum of sine wave.

(c) 22Hz triangle.

(d) Spectrum of triangle wave.

(e) 22Hz 50% duty cycle square.

(f) Spectrum of square wave.

Figure 3.1: (a), (c), and (e) are the luminance figures of a LED measured by a Lutron LX-102 light meter. Their corresponding frequency representations are given in (b), (d), and (f), respectively. The spectrum of the square wave strictly adheres to theory, that is, a peak demonstrated at fundamental frequency f as well as a peak at the $3f$ harmonic. The sine wave and the triangle wave do not. They have weak harmonics that should not exist at $2f$. However, these harmonics should not affect the result since their strength are one tenth that of the fundamental frequency.

ear (or close to linear) operating region. For the LED used in our experiments, a $3.25V$ DC bias was applied. The resulting linear operating region is $[3V, 3.5V]$. The luminance of the LED was converted to an electrical signal using a Lutron LX-102 light meter. The output of the light meter was visualized using an Agilent 54621D oscilloscope. Figure 3.1 shows the luminance signal and its spectrum (in dB) of the three waves on the oscilloscope. Note that the light signals were not perfectly sine, triangle or square waves due to the nonlinearity of the LED. The artifacts on the sine and triangle waves were more significant than on the square wave. For example, $2f$, which should not exist theoretically in sine or triangle waves, appeared in the measured luminance signal. Nevertheless, the amplitude of $2f$ in the measured sine or triangle luminance is around 20dB weaker than the fundamental frequency, i.e., the amplitude is about one order of magnitude smaller.

Five subjects participated in this experiment. The EEG was recorded with one channel over the occipital cortex at a sampling rate of $1k$Hz, then filtered by a 0.15Hz high-pass filter and a 150Hz low-pass filter. The distance between the LED and a subject was 50 cm. We examined stimuli of 11Hz, 13Hz, 15Hz, 18Hz and 22Hz, and recorded the SSVEPs of square, triangle, and sine waves. Square waves were generated with $10\%, 25\%$ and 50% duty cycles. In each recording session, the subject was told to keep looking at the stimulus for 8 seconds and close eyes for a rest period of a random duration from 10 to 20 seconds. The recorded data were discarded when muscle movements artifacts were significant.

3.2 Results and Conclusions

Table 3.1 reports the SSVEP results from all subjects. f is the fundamental frequency of the stimulus. *"Valid trials"* is the number of trials that the magnitude of FFT coefficients of SSVEP at f, $2f$, or $3f$ are 50% greater than the baseline. *"Total trials"* is the number of experiments in which a stimulus is presented to a user, regardless of whether the SSVEP peaks were detected. *"$1f$ occurs, $2f$ occurs*, and *$3f$ occurs"* are the number of observed SSVEP peaks at $1f$, $2f$ and $3f$, respectively.

Theoretically, SSVEP peaks appear at the stimulus frequency $1f$ and its harmonics $2f$, $3f$ etc. An SSVEP system has to use an recognizable $1f$ component to identify which frequency the subject is looking at, while sometimes uses its harmonics to improve the accuracy. Thus, a valid trial without a $1f$ peak may not be acceptable in a real SSVEP system. So we define a trial in which $1f$ occurs as an accurate trial, and the accuracy of a certain type of waveform of a certain frequency is $Accuracy_{wave,frequency} = \frac{1f\ occurs}{Total\ trials}$. Figure 3.2 shows the accuracies of SSVEP trials driven by the three waves above.

Table 3.1: Statistics of harmonics in SSVEP

	1f occurs	2f occurs	3f occurs	Valid trials	Total trials
11Hz sine	20	10	7	22	29
13Hz sine	22	9	2	22	30
15Hz sine	23	8	5	25	33
18Hz sine	23	9	6	25	34
22Hz sine	19	12	1	20	26
11Hz triangle	14	10	4	16	22
13Hz triangle	19	10	0	19	21
15Hz triangle	16	5	5	16	17
18Hz triangle	17	6	2	17	21
22Hz triangle	15	9	3	15	19
11Hz 50% square	20	11	15	20	21
13Hz 50% square	17	5	5	17	19
15Hz 50% square	17	9	8	16	17
18Hz 50% square	18	9	8	19	19
22Hz 50% square	18	9	8	18	19
11Hz 25% square	11	9	5	11	15
13Hz 25% square	17	8	6	18	18
15Hz 25% square	7	7	7	10	15
18Hz 25% square	17	14	10	18	18
22Hz 25% square	15	15	10	18	18
11Hz 10% square	8	9	4	12	17
13Hz 10% square	13	13	6	17	17
15Hz 10% square	15	12	11	19	20
18Hz 10% square	16	9	10	20	21
22Hz 10% square	13	6	6	15	19

We have the following observations.

- *A square waves with* 50% *duty cycle have a significantly higher accuracy than other stimuli in our experiment.*

  As shown in Figure 3.2, the average accuracies $(\frac{\sum_{allfrequencies} number\ of\ accurate\ trials}{\sum_{allfrequencies} total\ number\ of\ trials})$ of sine, triangle, and square waves with duty cycle 50%, 25% and 10% were 70.4%, 81.0%, 94.7%, 79.8%, and 69.1% respectively. Using statistic analysis techniques, we check if the performance of 50% square wave is better than that of triangle wave, which is intuitively the second best

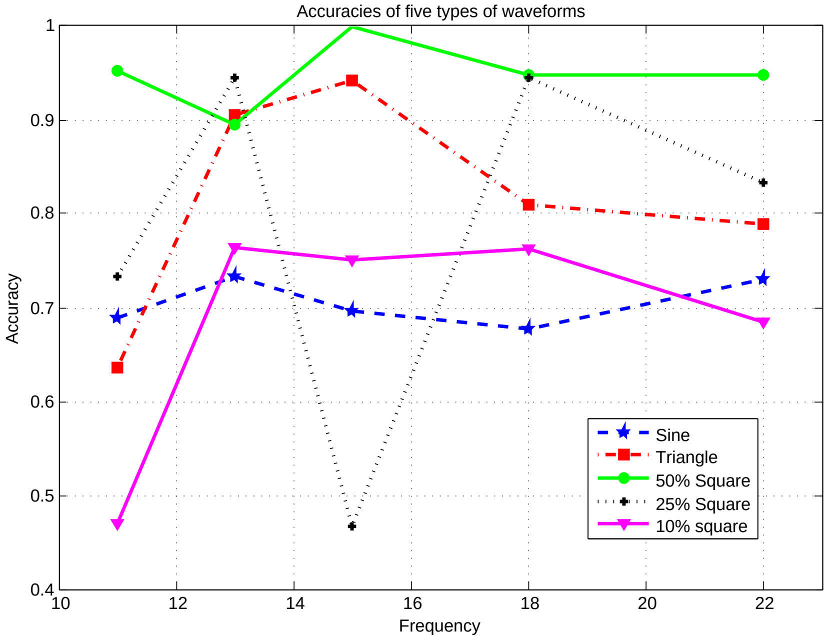


Figure 3.2: 11, 13, 15, 18 and 22Hz were used as the stimulus frequencies. The accuracies of the SSVEP experiments are computed with equation $Accuracy = \frac{1f\ occurs}{Total\ trials}$.

waveform as seen in Figure 3.2, with a significant level less than 0.05. $\frac{90}{95}$ 50% square waves and $\frac{81}{100}$ triangle waves evoked $1f$ SSVEP, thus $Z = \frac{(p_1-p_2)-(\pi_1-\pi_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}} = 1.728$. Since $Z_\alpha = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} = 1.645 < Z$, we conclude that a square waves with 50% duty cycle have a significantly higher accuracy than other stimuli in our experiment.

- *A square wave has a higher success rate than sine or triangle waves in eliciting SSVEPs.*

  In our experiments, the success rates (number of valid trials divided by the total number of trials) for sine, triangle, and square waves were 75.0%, 83.0%, and 90.8%, respectively.

- *All three wave forms elicited* $2f$ *component in SSVEPs.*

  In our experiments, the success rates for $2f$ component in SSVEP were 42.9% for sine waves, 48.2% for triangle waves, and 56.2% for square waves (averaged over all three duty cycles). Among the three duty cycles, 10%, 25%, and 50%, of the square wave, the $2f$ success rates were 43.0%, 70.7%, and 59.0%, respectively.

- *A square wave has a significantly higher success rate than sine or triangle wave in eliciting* $3f$ *component in SSVEPs.*

  In our experiments, the success rates for $3f$ component in SSVEP were 18.4% for sine waves, 14.0% for triangle waves and 48.0% for square waves (averaged over all three duty cycles). Among the three duty cycles, 10%, 25%, and 50%, of the square wave, the $3f$ success rates were 44.6%, 50.7%, and 55.0%, respectively.

Although sine, triangle, and square waves with 50% duty cycle do not contain $2f$ component, they all elicited $2f$ in SSVEP with similar success rates. Square wave with 25% duty cycle contains a strong $2f$ component. Its $2f$ success rate is significantly higher (70.7%). This suggests that: (1) the $2f$ component is primarily elicited by the fundamental frequency; (2) 8 in the stimuli increase the success rate of $2f$ in SSVEP. A similar observation is obtained for $3f$. This seems to suggest that *although the fundamental* *frequency can elicit harmonics (2f and 3f in our experiments) in SSVEP, the success rate of getting harmonics in SSVEPs is positively correlated with the strength of the artifacts in a stimulus.*

We observed that square waves with 50% duty cycle have a significantly higher accuracy than other stimuli in our experiment. As a result, the use of square waves with 50% duty cycle is preferred if high 1f SSVEP eliciting rate is the goal, while sine waves for SSVEP simulation should be chosen if few harmonic artifacts are wanted.

Our results also show that the harmonics associated with SSVEP are elicited both by the fundamental frequency and the artifacts of the stimuli, with the 2f component produced by the fundamental frequency and the 3f by the artifacts of square waves. At the same time, SSVEP elicited with square waves do not always contain all the artifactual frequency components, e.g. 3f, and SSVEP with sine waves may have 3f harmonics, which is not a part of the stimuli artifacts.

Chapter 4

DUAL AND TRI-STIMULI

According to Equation 1.1, it is straightforward that the more the possible selections, the more bits (information) a decision carries. Thus, after an effective stimulus, a stimulation method that provides more distinguishable stimuli is the second aspect enhancing the performance of an SSVEP BCI. In this chapter, we propose dual stimuli as the solution, and claim that 4Hz is the resolution[1] of the dual stimuli.

4.1 Methodology

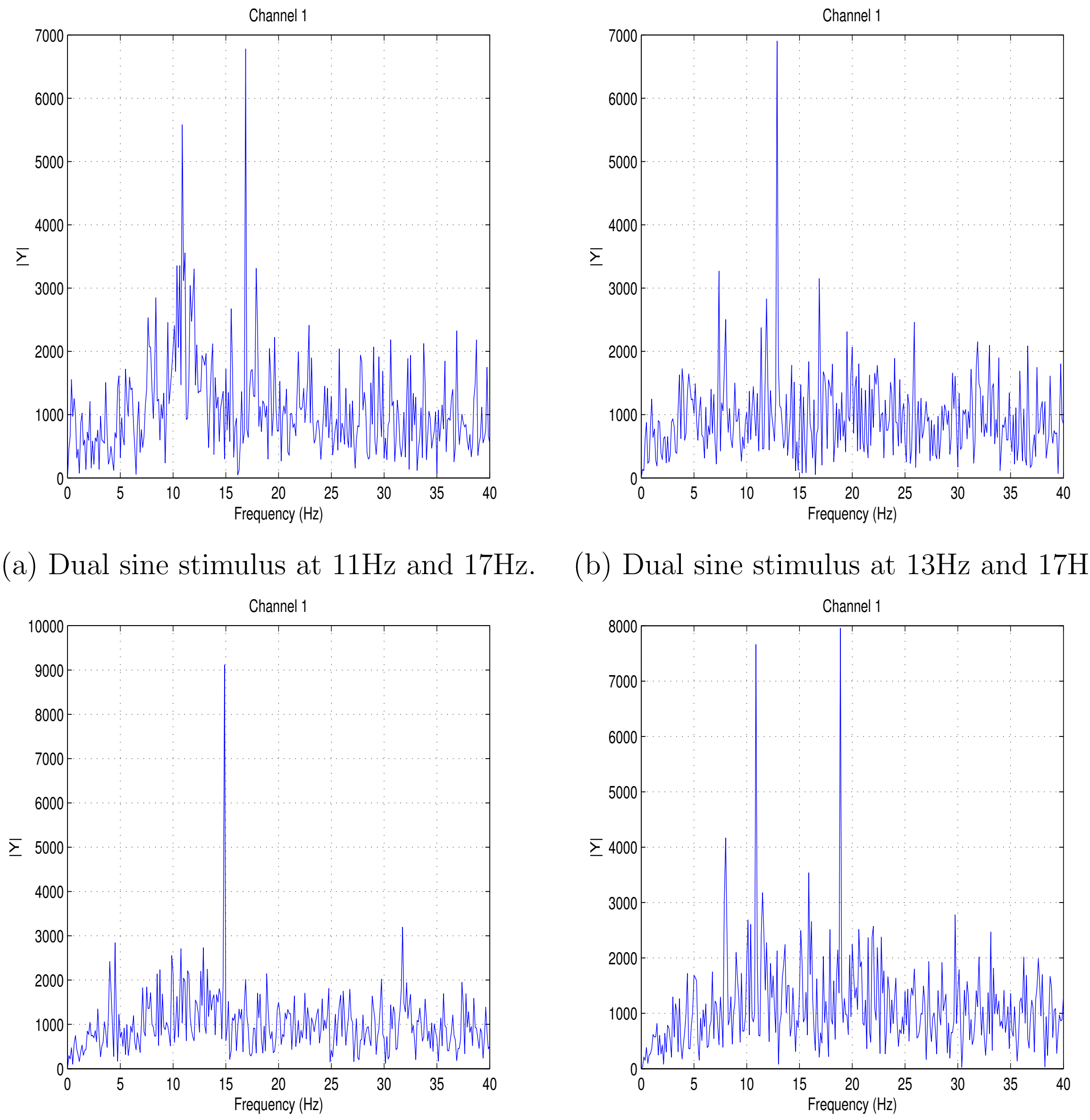
Because the strongest SSVEP responses are observed in the range of 5–20Hz (34; 47; 68), our SSVEP BCI uses 10-20 integer Hz signals as stimuli. Theoretically, SSVEP occurs at exactly the same frequency as a stimulus. However, considering noises from the outside world, an error margin has to be introduced. In our system, we only use integer Hz stimuli between ten to twenty Hz, and round any SSVEP peak (in the frequency domain) between

[1] The resolution is defined as the minimum distance between two frequencies in a dual-stimulus that elicits consistent SSVEP

$[n-0.5, n+0.5)$ Hz, n=10..20, to n. Under this scenario, one round of SSVEP detection can only make one "one out of eleven" choice. In order to increase the information throughput, the use of dual stimuli is proposed. Dual stimuli increase the number of distinct stimuli. For example, the sum of 13Hz and 17Hz sine waves is considered a dual stimuli, while the sum of 13Hz and 18Hz sine waves is considered another dual stimuli. Cheng et al. (34) used multiple color stimuli to deliver two stimuli simultaneously. However, no research has been done regarding the resolution of the dual stimuli. This section identifies the resolution of dual stimuli that provides consistent SSVEP. Stimuli were generated by summation of two sine waves on an LED.

Compared with a sine wave, which has no harmonics, a square wave and a triangle wave contain strong harmonic components as given by their Fourier representation. This suggests that the use of sine waves for SSVEP simulation may be preferred over the other wave forms due to reduced harmonic artifacts. In order to make an LED emit a sine wave light signal, a carefully selected DC bias has to be added to the sine input signal. For the LED used in our experiments, the linear region is 3v to 3.5v with DC bias 3.25v.

We tested the dual and tri-stimuli on one human subject. An LED stimulator was used to elicit an SSVEP response. For the LED used in our experiments, the linear region is 3v to 3.5v with DC bias 3.25v.



(a) Dual sine stimulus at 11Hz and 17Hz.

(b) Dual sine stimulus at 13Hz and 17Hz.

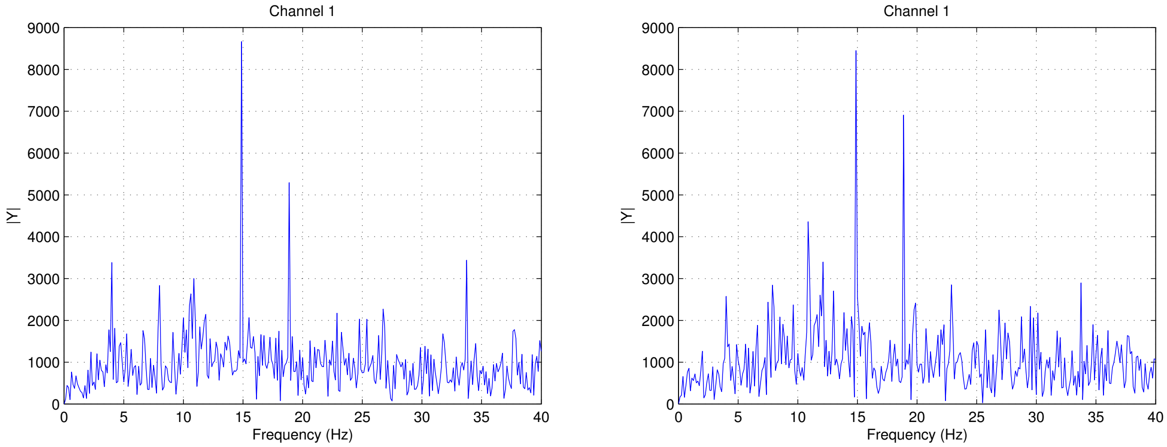
(c) Dual sine stimulus at 15Hz and 17Hz.

(d) Dual sine stimulus at 11Hz and 19Hz.

Figure 4.1: Spectrums of SSVEP for dual stimuli.

4.2 Results and Conclusions

We tested the dual stimulus on one human subject. An LED stimulator was used to elicit SSVEP. Five seconds of EEG signal were recorded in each



(a) Tri sine stimulus at 11Hz, 15Hz, 19Hz, Test 1.

(b) Tri sine stimulus at 11Hz, 15Hz, 19Hz, Test 2.

Figure 4.2: Spectrums of SSVEP for tri stimuli.

test. Figure 4.1 shows the spectrum of SSVEP for dual stimulus tests with frequency combination of 11-17Hz, 13-17Hz, 15-17Hz and 11-19Hz. It was observed when the two frequencies in the stimulus were only 2Hz or less apart, SSVEP can only detect one stimulus frequency (Figure 4.1(c)). In most cases, the detected frequency is the lower frequency in the stimulus. Noticeable dual SSVEP spikes could be seen if two frequencies were 4 Hz apart (Figure 4.1(b)), while in most cases, the amplitude of the higher frequency in SSVEP is lower than that of the lower frequency. Two distinctive spikes can be detected if the frequencies of the stimulus were at least 6 Hz apart (Figure 4.1(a)(b)).

For the tri-stimulus tests, we saw three noticeable SSVEP spikes in only one out of five tests (Figure 4.2(b)). In the other four tests, there were spikes at one or two of the three stimulus frequencies. Figure 4.2 shows the results

Table 4.1: SSVEP at different combinations of stimulus frequencies.

Stimulus Frequencies	Number of Tests	Good Responses	Fair Responses	Failed Responses
11-13Hz	3	0	0	3
11-15Hz	5	3	1	1
11-17Hz	3	3	0	0
11-19Hz	3	3	0	0
13-19Hz	3	3	0	0
15-19Hz	5	1	1	3
11-15-19Hz	5	1	3	1

of two tri stimulus tests with 11-15-19Hz visual stimuli. It is interesting to observe that in all five tests the lowest frequency was lost in the EEG spectrum instead of the largest frequency as in the dual stimulus tests.

The SSVEP results for different dual- and tri-frequency combinations are summarized in Table 4.1. A "Good response" is one in which all stimulus frequencies are distinctive in SSVEP. A "Fair response" is one in which some stimulus frequencies are distinctive in SSVEP. When there is no SSVEP, we call it a "Failure".

Chapter 5

STIMULI AND COLOR-SPACE DECOMPOSITION

A good SSVEP system shall focus not only on the usability and speed, but also the user experience. Because the best stimulation frequency region of an SSVEP BCI is 5–20Hz, which reside in the low frequencies (< 30 Hz) range that may cause safety hazards linked to photo-induced epileptic seizures (46; 47), we explore the design of a visually friendly stimulus from the perspective of color-space decomposition in this chapter. This low-frequency visually friendly stimulus is designed with a fixed luminance component and variations in the other two dimensions in the HSL space, based on the assumption that iso-luminant stimuli may ease the feeling of dizziness.

5.1 Methodology

We designed iso-luminant stimuli in the HSL color space. Because the SSVEP has the same fundamental frequency as the visual stimulus (17), it is important to ensure that the stimulators are exact as the software generator set it; otherwise accurate results may not be achieved. In our experiments, the stimuli were carefully designed to achieve credible results, described below.

- **Accurate Frequencies**

  It is not straightforward to deliver accurate stimuli with computer screens. Jaganathan claimed that the PC hardware and operating system seem to determine the variability of stimulation frequency (62). Sugiarto and Sutoyo claimed that DirectX, OpenGL and Matlab are effective in implementing an accurate stimulus with a computer screen (123; 124). The refresh rate of the monitor also limits the frequency rage of the stimulus. The refresh rate R is the number of times a display's image is repainted or refreshed per second. Intuitively, as at least two points form a cycle, only frequencies lower than R/2 Hz can be used and only the subharmonics of the screen refresh rate can be obtained. Furthermore, the task scheduling that most operating systems perform often affects the rendering of the frequency, which are usually unpredictably delayed, especially when a lot of stimuli were set simultaneously. Thus, we used DirectX and a CRT monitor with a refresh rate of 60Hz and 120Hz to deliver 6Hz and 12Hz stimuli, respectively. And the program only shows one flashing object on the screen at a time.

- **Stimuli**

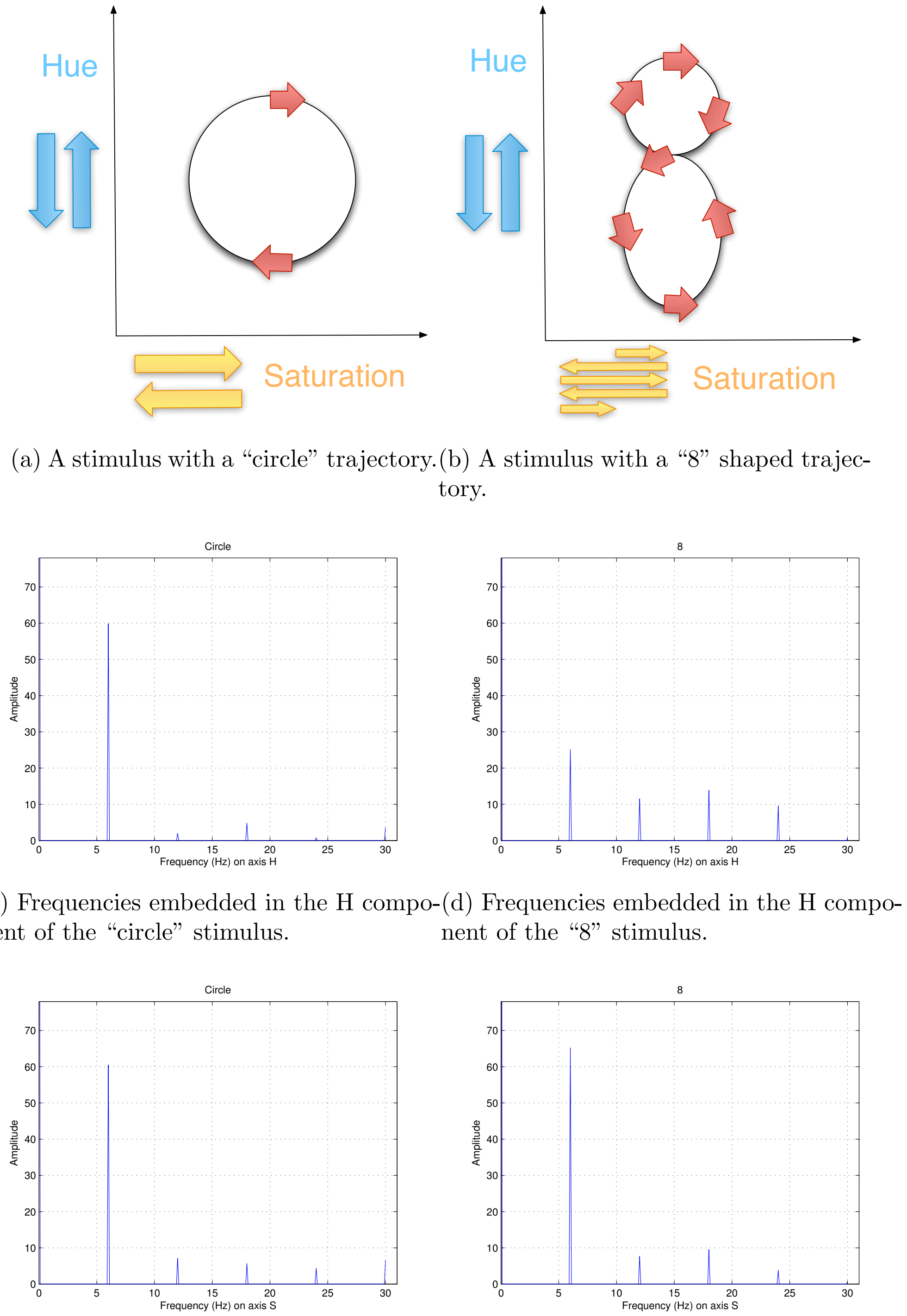
  The HSL stimulus was designed as a flashing square box with changing color, sized 100*100 pixels in a 17inch monitor, with a resolution of

Table 5.1: HSL Space Stimuli HSL and RGB Values in One Cycle

	Two points		Circle		“8” size	
	HSL	RGB	HSL	RGB	HSL	RGB
1	0.12,0.56,0.80	208,200,200	0.82,0.20,0.80	219,190,189	0.86,0.86,0.80	248,161,160
2	0.12,0.56,0.80	208,200,200	0.72,0.23,0.80	216,193,192	0.80,0.95,0.80	252,157,156
3	0.12,0.56,0.80	208,200,200	0.56,0.24,0.80	216,192,192	0.71,0.86,0.80	248,161,160
4	0.12,0.56,0.80	208,200,200	0.50,0.31,0.80	220,188,188	0.79,0.78,0.80	244,165,164
5	0.12,0.56,0.80	208,200,200	0.47,0.41,0.80	225,183,183	0.69,0.70,0.80	240,169,168
6	1.28,0.87,0.80	233,176,175	0.50,0.53,0.80	231,177,177	0.67,0.53,0.80	231,177,177
7	1.28,0.87,0.80	233,176,175	0.58,0.59,0.80	234,174,174	0.77,0.45,0.80	227,181,181
8	1.28,0.87,0.80	233,176,175	0.74,0.60,0.80	235,174,173	0.89,0.53,0.80	231,178,177
9	1.28,0.87,0.80	233,176,175	0.82,0.52,0.80	231,178,177	0.90,0.69,0.80	239,170,169
10	1.28,0.87,0.80	233,176,175	0.85,0.40,0.80	224,184,184	0.79,0.78,0.80	244,165,164

1024*768 pixels. Three typical HSL-space stimuli were tested, one for a cycle formed by two points jumping between each other, one for a circle and one for a size of number eight. Trajectories and frequency analysis of two of them are shown in Figure 5.1. HSL and RGB values (10 sample points per cycle[1]) within one cycle are shown in Table 5.1. They have a fixed luminance component and variations in the other two dimensions in the HSL space. Furthermore, it is noteworthy that any frequency could be embedded in HSL stimuli by adding them to either H or S axis. For example, if 11,15 and 18Hz are wanted, we could use $sin(2\pi * 11 * t) + sin(2\pi * 15 * t)$ as H values and $sin(2\pi * 18 * t)$ as S values.

[1] Refresh rate / Stimulus frequency = Sampling points per cycle : 60Hz/6Hz=120Hz/12Hz=10.



(a) A stimulus with a “circle” trajectory.

(b) A stimulus with a “8” shaped trajectory.

(c) Frequencies embedded in the H component of the “circle” stimulus.

(d) Frequencies embedded in the H component of the “8” stimulus.

(e) Frequencies embedded in the S component of the “circle” stimulus.

(f) Frequencies embedded in the S component of the “8” stimulus.

Figure 5.1: In the HSL color-space, the luminance is fixed, while the hue and the saturation vary along a trajectory. Frequencies are delivered by the change of the Hue and Saturation together (the closed curve), by the change of the Hue only (the H axis, figure(c)(d)) or by the change of Saturation only (the S axis, figure(e)(f)). A cycle begins at a certain point on the curve and ends when the trajectory of the stimulus hits this point again. The changes in the SL space can be continuous or discrete.

5.2 Results and Conclusions

Six subjects participated in this experiment. EEG was recorded with one channel over the occipital cortex at a sampling rate of 1kHz, filtered by a 0.15Hz high-pass filter and a 150Hz low-pass filter. The resistances between the skin and the sensor are all below 10k. The distance between the CRT and a subject was 40 cm. We examined stimuli of 6Hz and 12Hz, and recorded the SSVEPs of "two points", circle, "“8" shaped trajectory and a black-white flashing box as the control stimuli. This test session was repeated for three times. In each recording session, the subject was told to look at the stimulus for 10 seconds and close their eyes for a rest period of a random duration from 10 to 20 seconds. The recorded data were discarded then repeated when muscle movements artifacts were significant. Figure 5.2 shows the SSVEP spectrums of the above four stimuli.

The primary research goals of these experiments are to find out if these stimuli elicit SSVEP, and if this color-space decomposition makes low-frequency stimuli more visually friendly than ordinary luminance stimuli. Table 5.2 reports the SSVEP results of all subjects. f is the fundamental frequency of the stimulus. *"Total trials"* is the number of experiments in which a stimulus is presented to a user. *"*$1f$ *occurs,* $2f$ *occurs"* are the number of observed SSVEP peaks at $1f$ and $2f$.

We have the following observations.

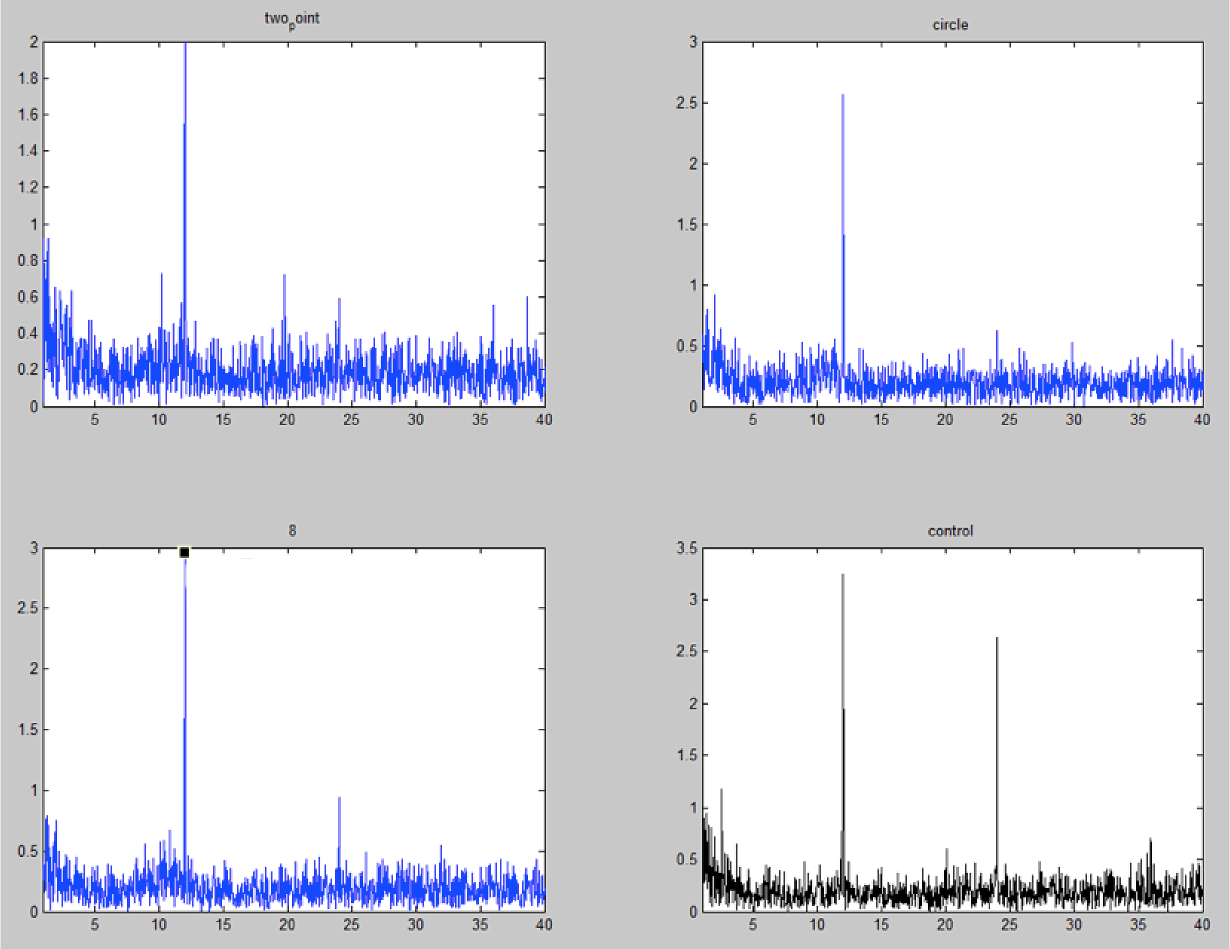


Figure 5.2: Spectrums of SSVEP of three types of stimulation. The stimulus is a 12Hz flashing square on a computer screen.

Table 5.2: Statistic of harmonics in SSVEP

	1f occurs	2f occurs	Total trials
6Hz two pints	18	10	18
6Hz circle	18	11	18
6Hz eight	18	15	18
6Hz control	18	18	18
12Hz two points	18	10	18
12Hz circle	18	12	18
12Hz eight	18	15	18
12Hz control	18	18	18

- *A stimulus with a fixed luminance and variations in the other two dimensions in the HSL space elicits SSVEP.*

  As shown in Table 3.1, all HSL space stimuli elicit SSVEP at their fundamental frequency.

- *The embedded frequencies affect SSVEP.*

  In our experiments, the success rates (number of its occurrence divided by the total number of trials) of "2f occurs" for two points, circle and "8" stimuli were 55.6%, 63.9%, and 83.3%, respectively, which suggests that the embedded $2f$ in "8" stimulus affects the $2f$ harmonic in its SSVEP.

- *All stimuli elicit SSVEP harmonics.*

  All types of stimuli evoke harmonics, though the success rates vary.

- *This color-space decomposition makes low-frequency stimuli more visually friendly than ordinary luminance stimuli.*

  All six subjects felt these fixed luminance stimuli were more comfortable than the control "black-white flashing box" stimulus. However, there is not enough evidence to conclude that this technique decreases the risk of safety hazards.

Chapter 6

POTENTIAL FUNCTION CLASSIFIER

A machine learning approach introduces adaptiveness, accuracy and speed to an SSVEP BCI, and improves BCI performance by learning brain patterns. Considering that a subject's brain signal is non-stationary, e.g., the SSVEP responds may be strong in the morning but weak in the afternoon, a simple threshold may not be a good choice: if it is set too high, it will miss peaks in the afternoon, if it is set too low, it will categorize noises as SSVEP. Consequently, Potential Function Classifier (PFR) is introduced to our SSVEP BCI.

The PFR is motivated by the potential field of static electricity. A binary PFR views each training sample as an electrical charge, positive or negative according to its class label. The resulting potential field divides the feature space into two decision regions based on the polarity of the potential. The basic idea of binary PFRs can be generalized to the multiclass scenario, in which a potential function is defined for each class using the training observations within that class. A new observation is then assigned a label corresponding to the class of the highest potential value. Intuitively, adding new classes does not affect the existing potential functions. Removing or merging classes influence only the potential functions of the classes involved in the operation. In SSVEP-based BCI context, these advantages can be interpreted as: Adding a new stimulus do not affect the existing PFRs. Removing a stimulus that is not currently well responded or merging stimuli that are not clearly separable influence only the PFRs involved in the operation. This good scalability of PFRs makes it suitable to BCI systems.

In this chapter, we first introduce the PFR method from the perspective of a machine learning technique. Then run PFR in offline SSVEP data and compare its bit rate calculated by Eq.(1.1) as the comparison metric.

6.1 Introduction

For thousands of years, various civilizations have observed "static electricity" where pieces of small objects with the same kind of electricity repelled each other and pieces with the opposite kind attracted each other. *potential function rules* were motivated from the underlying property of static electricity to predict the unknown binary nature of an observation, a problem commonly known as *binary classification.* Potential function rules were originally studied by Aizerman, Braverman, Rozonoer, and several other researchers in the 1960's ((2; 3; 12; 27; 28)). In its simplest form, a potential function rule puts a unit of positive electrical charge at every positive observation and a unit of negative electrical charge at every negative observation. The resulting potential field defines an intuitively appealing classifier: a new observation is predicted positive if the potential at that location is positive, and negative if its potential is negative.

Below, we revisit potential function rules (PFRs) in their original form and reveal their connections with other well-known results in the literature. We derive a bound on the generalization performance of potential function classifiers based on the observed margin distribution of the training data. A new model selection criterion using a normalized margin distribution is then proposed to learn "good" potential function classifiers in practice.

6.2 Background

There is an abundance of prior work in the field of pattern recognition and machine learning. It is beyond the scope of this study to supply a complete review of the area (for more comprehensive surveys on various subjects, the reader is referred to Devroye et al. (40), Duda et al. (44), Bishop (20) for patter recognition, to Schölkopf and Smola (116), Shawe-Taylor and Cristianini (119) for kernel methods, to Anthony and Biggs (5), Kearns and Vazirani (66) for computational learning theory, and to Mitchell (87), Hastie et al. (58), Vapnik (131) for machine/statistical learning). Nevertheless, a brief synopsis of some of the main findings will serve to provide a rationale for the proposal of a new machine learning approach used in an SSVEP BCI.

A multiclass classification problem aims at foretelling the unknown nature of an observation. More formally, an observation is a d-dimensional vector of numerical measurements denoted as $\mathbf{x} \in \mathbb{R}^d$. The unknown nature of the observation, z, takes values in a finite set $\mathbb{K} = \{1, 2, \ldots, K\}$, the set of *class labels*. A mapping $f : \mathbb{R}^d \rightarrow \mathbb{K}$, which is named a *classifier*, predicts the class label of an observation.

Does there exist an "optimal" classifier for a given classification task? Under a probabilistic setting, the Bayesian decision theory (13; 15) gives an affirmative answer – the *Bayes decision rule* (called the Bayes classifier). If the pair of observations and their nature, $(\mathbf{x}, z)$, is a random variable with a joint probability distribution $p(\mathbf{x}, z)$, the Bayes classifier, f^*, selects the class label for an observation $\mathbf{x}$ as $f^*(\mathbf{x}) = \text{argmax}_{z \in \mathbb{K}} \Pr(z|\mathbf{x}) = \text{argmax}_{z \in \mathbb{K}} p(\mathbf{x}, z)$. The optimality of f^* is defined by the minimum probability of error, i.e., $\Pr[f^*(\mathbf{x}) \neq z] \leq \Pr[f(\mathbf{x}) \neq z]$ for any $f : \mathbb{R}^d \rightarrow \mathbb{K}$, which is well-known as the *Bayesian probability of error.* This probability measures the 'hardness' of a classification problem. It can theoretically be evaluated if the joint distribution is known, but the calculation may be (and usually is) intractable in practice due to the min operator inside of the integral. Several tight bounds are proposed in the literature for computational approximations of the Bayesian probability of error (38; 57; 7).

The crux of the Bayesian approach is the difficulty of determining the joint distribution. Plug-in decision (40) is a natural way of applying the Bayesian classification in practice, where an approximated Bayes classifier is constructed using an estimated joint distribution. Depending upon the way in which the joint distribution is estimated, plug-in decision rules fall roughly into parametric approaches and nonparametric approaches.

In a parametric approach, the unknown joint distribution is described by a set of parameters based on certain structural assumptions, e.g., conditional independence of attributes within each class (75; 41; 26), mixture of Gaussians (69; 122), and mixture of Bernoullis (122). The values of the parameters are obtained by optimizing a loss function, e.g., a likelihood function. In many applications, a parametric approach presents an efficient means of incorporating prior knowledge about the data. For example, Hofmann et al. (61) used a latent variable model (*aspect model*) to remove the statistical dependence among words in a document for textual data. Barnard et al. (9) explored several generative models to describe statistical relevance between image regions and associated texts. Veeramachaneni and Nagy (133) studied the interpattern dependence, named *style context*, for Optical Character Recognition. Intraclass style (statistical dependence between patterns of the same class in a field) and interclass style (statistical dependence between patterns of different classes in the same field) were formalized to derive style-constrained Bayesian classification.

The performance of a plug-in decision rule is determined by the quality of the estimated joint distribution. Ben-Bassat et al. analyzed the sensitivity of Bayesian classification under multiplicative perturbation on the joint distribution. Devroye (39) presented a more general result showing that if the estimated posterior probability is close to the true posterior probability in L_1-sense, the error probability of the plug-in decision rule is near the Bayesian probability of error. Nevertheless, does the error probability converge to the Bayesian probability of error if more training samples are obtained to approximate an arbitrary joint distribution? This is a question regarding the universal consistency of a classification rule. Loosely speaking, a *universally consistent rule* (40) guarantees us that taking more samples suffices to roughly reconstruct an arbitrary, fixed, but unknown distribution, hence to asymptotically achieve the optimality. While parametric approaches are efficient, in general they are not universally consistent.

In 1977, Stone proved the existence of a universally consistent rule (121). He showed that any k-nearest neighbor classifier is universally consistent if

k is allowed to grow with n, the sample size, at a speed slower than that of n. Since then, several rules have been shown to be universally consistent including histogram rules (53) and kernel rules (40). We put these approaches under the category of nonparametric plug-in decisions because of the underlying nonparametric estimation of joint distributions. Representing all the data with a nonparametric model is sometimes preferred over summarizing it with a parametric model because of the rich detail held by very large data sources (56).

Universal consistency describes the asymptotic behavior of a classifier, i.e., the number of training samples is potentially infinite. For real-life problems, however, the size of a training set is finite and, usually, fixed. This leads to a basic question in classifier design: how do we select a classifier, which performs well on future examples, from a given set of classifiers based on a given finite training set? Two basic principles were investigated in the literature for classifier selection: empirical risk minimization (129) and complexity regularization (80).

In order to achieve good generalization performance, the empirical risk minimization principle seeks for a classifier that minimizes the training error (empirical risk). Vapnik and Chervonenkis presented a theoretical ground for empirical risk minimization (129). It was shown that if the 'capacity' of

$\mathbb{C}$, the set of classifiers to choose from, is sufficiently restricted, minimizing the empirical risk guarantees a classifier whose performance is close to that of the best classifier in $\mathbb{C}$. Here the capacity of $\mathbb{C}$ is defined by the VC-dimension of $\mathbb{C}$, which is defined as the maximum h such that some data point set of cardinality h can be shattered by $\mathbb{C}$ (see Figure 6.1). The above result reveals two competing factors in classifier selection. On one hand, a low capacity model set may not contain any classifier that generalizes well. On the other hand, too much freedom may over fit the data resulting a model behaving like a refined look-up-table: perfect for the training data but poor on generalization.

This suggests that a classifier, built on a finite training set, generalizes the best if the right tradeoff is found between the training accuracy and the capacity of the model set. Complexity regularization applies the above idea to search for a classifier that minimizes the sum of empirical risk and a term penalizing the complexity (130; 10; 80; 11). Amongst various definitions of the penalty term, margin-based approaches received broad attention in the literature. A series of results were obtained that exhibit the intrinsic connection between generalization and different measures of margin distribution (e.g., maximal margin, margin percentile, soft margin) (131; 74; 119; 105; 55). These theoretical results led to the discovery of new learning algorithms (e.g.,

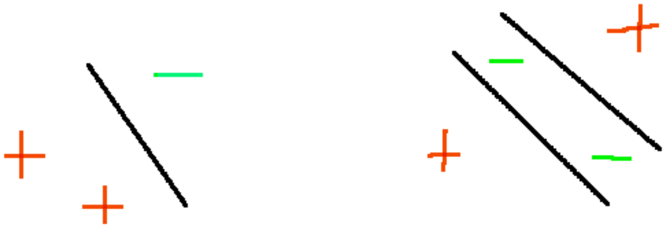


Figure 6.1: Considering a straight line is used as the classifier to separate the "+" points from the "-" points. It is intuitive that any three points that do not fall on a same straight line can be shattered by this model (left), while some set of four points can not be shattered (right). Thus, the VC dimension of this particular classifier is three.

support vector machines (131), margin distribution optimization (52), large margin multiple-instance learning (31), margin trees (128), large margin semi-supervised learning (139), dissimilarity-based learning (99), similarity-based learning (82; 33), large margin nearest neighbor classification (140)) and new interpretations of known learning algorithms (e.g., boosting (114; 110), additive fuzzy systems (30)).

Classifiers derived from complexity regularization are not necessarily consistent. Lugosi and Zeger (80) presented a sufficient condition for universal consistency of a particular method of complexity regularization, structural risk minimization, using Vapnik-Chervonenkis complexity classes (131).

6.3 Potential Function Rules

For thousands of years, various civilizations have observed "static electricity" where pieces of small objects with the same kind of electricity repelled each other and pieces with the opposite kind attracted each other. In pattern recognition and machine learning, *potential function rules* were motivated from the underlying property of static electricity to predict the unknown binary nature of an observation, a problem commonly known as *binary classification.* Potential function rules were originally studied by Aizerman, Braverman, Rozonoer, and several other researchers in the 1960's ((2; 3; 12; 27; 28)). In its simplest form, a potential function rule puts a unit of positive electrical charge at every positive observation and a unit of negative electrical charge at every negative observation. The resulting potential field defines an intuitively appealing classifier: a new observation is predicted positive if the potential at that location is positive, and negative if its potential is negative.

In the following sections, we revisit potential function rules (PFRs) in their original form and reveal their connections with other well-known results in the literature. We derive a bound on the generalization performance of potential function classifiers based on the observed margin distribution of the training data. A new model selection criterion using a normalized margin distribution is then proposed to learn “good” potential function classifiers in practice.

6.4 Potential Function Rules and The Bayes Decision Theory

We start with a brief review of electrostatic potential functions (54). We then introduce the general form of binary potential function classifiers. Finally, we demonstrate connections between PFRs and the Bayes classifiers.

Given a positive point charge at location $\mathbf{y}$, the electrostatic potential at location $\mathbf{x}$ is proportional to $\frac{1}{\|\mathbf{x}-\mathbf{y}\|}$, which is called the electrostatic point potential function. For a ‘cloud’ of positive charges with density ρ_+ over a space $\mathbb{X}$, the electrostatic potential function Φ is,

$$\Phi(\mathbf{x}) = \int_{\mathbb{X}} \frac{\rho_+(\mathbf{y})}{\|\mathbf{x}-\mathbf{y}\|} d\mathbf{y} \ .$$

Therefore, if ρ_+ and ρ_- are respectively the charge density of positive and negative charges over $\mathbb{X}$, the electrostatic potential function Φ is defined as

$$\Phi(\mathbf{x}) = \int_{\mathbb{X}} \frac{\rho_+(\mathbf{y})}{\|\mathbf{x}-\mathbf{y}\|} d\mathbf{y} - \int_{\mathbb{X}} \frac{\rho_-(\mathbf{y})}{\|\mathbf{x}-\mathbf{y}\|} d\mathbf{y} \ .$$

The above electrostatic potential function can be generalized by replacing the electrostatic point potential function with a general *point potential* *function* $\psi : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$:

$$\Phi(\mathbf{x}) = \int_{\mathbb{X}} \rho_+(\mathbf{y})\psi(\mathbf{x},\mathbf{y})d\mathbf{y} - \int_{\mathbb{X}} \rho_-(\mathbf{y})\psi(\mathbf{x},\mathbf{y})d\mathbf{y} \ . \tag{6.1}$$

Note that the electrostatic potential at a location $\mathbf{x}$ is not well defined if $\mathbf{x}$ falls in the support of ρ_+ or ρ_- due to the fact that $\frac{1}{\|\mathbf{x}-\mathbf{y}\|}$ is ∞ when $\mathbf{x} = \mathbf{y}$. This limitation, however, can be avoided by a general potential function (6.1) with a proper choice of the point potential function ψ.

Given ρ_+ and ρ_-, let Q_+ and Q_- be the total positive charge and negative charge, respectively:

$$Q_+ = \int_{\mathbb{X}} \rho_+(\mathbf{x})d\mathbf{x}, \quad Q_- = \int_{\mathbb{X}} \rho_-(\mathbf{x})d\mathbf{x}.$$

We normalize the potential function (6.1) by the sum of the total positive and total negative charges:

$$\frac{\Phi(\mathbf{x})}{Q_+ + Q_-} = \frac{Q_+}{Q_+ + Q_-}\int_{\mathbb{X}} \frac{\rho_+(\mathbf{y})}{Q_+}\psi(\mathbf{x},\mathbf{y})d\mathbf{y} - \frac{Q_-}{Q_+ + Q_-}\int_{\mathbb{X}} \frac{\rho_-(\mathbf{y})}{Q_-}\psi(\mathbf{x},\mathbf{y})d\mathbf{y} \ . \tag{6.2}$$

It is not difficult to check that $\frac{\rho_+(\mathbf{x})}{Q_+}$ and $\frac{\rho_-(\mathbf{x})}{Q_-}$ can be viewed as probability density functions because they are nonnegative over $\mathbb{X}$ and $\int_{\mathbb{X}} \frac{\rho_+(\mathbf{x})}{Q_+}d\mathbf{x} = \int_{\mathbb{X}} \frac{\rho_-(\mathbf{x})}{Q_-}d\mathbf{x} = 1$, i.e., normalized charge densities are probability density functions. Therefore, we define conditional probability density functions as

$$p(\mathbf{x}|+) = \frac{\rho_+(\mathbf{x})}{Q_+}, \quad p(\mathbf{x}|-) = \frac{\rho_-(\mathbf{x})}{Q_-}, \tag{6.3}$$

and prior probability as

$$\Pr(+) = \frac{Q_+}{Q_+ + Q_-}, \quad \Pr(-) = \frac{Q_-}{Q_+ + Q_-}. \tag{6.4}$$

Consequently, the above normalized potential function 6.2 is rewritten in terms of (6.3) and (6.4) as

$$\frac{\Phi(\mathbf{x})}{Q_+ + Q_-} = \Pr(+) \int_{\mathbb{X}} p(\mathbf{y}|+)\psi(\mathbf{x},\mathbf{y})d\mathbf{y} - \Pr(-) \int_{\mathbb{X}} p(\mathbf{y}|-)\psi(\mathbf{x},\mathbf{y})d\mathbf{y} \ .$$

Hence a binary potential function classifier is defined as

$$f(\mathbf{x}) = \operatorname{sign}\left(\Phi(\mathbf{x})\right) = \operatorname{sign}\left(\Pr(+) \int_{\mathbb{X}} p(\mathbf{y}|+)\psi(\mathbf{x},\mathbf{y})d\mathbf{y} - \Pr(-) \int_{\mathbb{X}} p(\mathbf{y}|-)\psi(\mathbf{x},\mathbf{y})d\mathbf{y}\right), \tag{6.5}$$

i.e., the polarity of the potential determines the class label.

Next, we present a Bayesian interpretation of the above potential function classifier. In particular, we show that with a proper choice of ψ, the decision boundary of (6.5) is identical to that of the optimal Bayes classifier. Our first choice of ψ is the Dirac delta function which is zero everywhere except at the origin, where it is infinite,

$$\delta(\mathbf{x}) = \begin{cases} +\infty & \mathbf{x} = \mathbf{0} \\ 0 & \mathbf{x} \neq \mathbf{0} \end{cases}$$

and which also satisfies the identity

$$\int_{-\infty}^{\infty} \delta(\mathbf{x}) d\mathbf{x} = 1.$$

Theorem 1 Let ρ_+ and ρ_- be the charge densities; $p(\mathbf{x}|+)$, $p(\mathbf{x}|-)$, $\Pr(+)$, and $\Pr(-)$ be defined by (6.3) and (6.4). If we choose $\psi(\mathbf{x}, \mathbf{y}) = \delta(\mathbf{x} - \mathbf{y})$, the decision boundary of the potential function classifier (6.5) is equivalent to that of the Bayes classifier for conditional probability distributions $p(\mathbf{x}|+)$ and $p(\mathbf{x}|-)$, and class prior probabilities $\Pr(+)$ and $\Pr(-)$.

A proof of Theorem 1:

Because $\delta(\cdot)$ is a Dirac delta function, it follows that

$$\begin{aligned} \int_{\mathbb{X}} p(\mathbf{y}|+)\delta(\mathbf{x} - \mathbf{y}) d\mathbf{y} &= p(\mathbf{x}|+) \\ \int_{\mathbb{X}} p(\mathbf{y}|-)\delta(\mathbf{x} - \mathbf{y}) d\mathbf{y} &= p(\mathbf{x}|-). \end{aligned}$$

Therefore,

$$\begin{aligned}\Pr(+)\int_{\mathbb{X}} p(\mathbf{y}|+)\delta(\mathbf{x}-\mathbf{y})d\mathbf{y} &\propto \Pr(+|\mathbf{x}) \\ \Pr(-)\int_{\mathbb{X}} p(\mathbf{y}|-)\delta(\mathbf{x}-\mathbf{y})d\mathbf{y} &\propto \Pr(-|\mathbf{x}),\end{aligned}$$

i.e., the potential of the positive (negative) class is proportional to the posterior probability of the positive (negative) class. Hence the decision boundary of (6.5) is identical to that of the Bayes classifier.

We may interpret the above theorem from the perspective of Fourier analysis. Specifically, for a translation invariant point potential function, i.e., $\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}-\mathbf{y})$, the evaluation of $\int_{\mathbb{X}} p(\mathbf{y}|+)\psi(\mathbf{x}-\mathbf{y})d\mathbf{y}$ is essentially the convolution of $p(\mathbf{x}|+)$ and $\psi(\mathbf{x})$, which is equivalent to computing the inverse Fourier transform of the product of the Fourier transforms of $p(\mathbf{x}|+)$ and $\psi(\mathbf{x})$. When ψ is the Dirac delta function, potential function classifiers are equivalent to Bayes classifiers because the Fourier transform of the Dirac delta function is the constant 1.

Theorem 1 holds independent of the specific forms of the charge densities, i.e., it is distribution free. Nevertheless, the unboundedness of the Dirac delta function makes it a poor choice in numerical implementations. Next, by assuming that the Fourier transform of the charge densities have finite support, we extend the conclusion of Theorem 1 to a wider class of translation invariant point potential functions.

Theorem 2 Let ρ_+ and ρ_- be the charge densities; $p(\mathbf{x}|+)$, $p(\mathbf{x}|-)$, $\Pr(+)$, and $\Pr(-)$ be defined by (6.3) and (6.4). Let $\hat{p}_+(\boldsymbol{\omega})$ and $\hat{p}_-(\boldsymbol{\omega})$ be the Fourier transform of $p(\mathbf{x}|+)$ and $p(\mathbf{x}|-)$, respectively, i.e.,

$$\begin{aligned} \hat{p}_+(\boldsymbol{\omega}) &= \int_{\mathbb{X}} p(\mathbf{x}|+)e^{-2\pi i \boldsymbol{\omega}^T \mathbf{x}} d\mathbf{x}, \\ \hat{p}_-(\boldsymbol{\omega}) &= \int_{\mathbb{X}} p(\mathbf{x}|-)e^{-2\pi i \boldsymbol{\omega}^T \mathbf{x}} d\mathbf{x}, \end{aligned}$$

where i is the complex number $\sqrt{-1}$. We assume that $\hat{p}_+$ and $\hat{p}_-$ have finite support, namely, there exist constants s_+ and s_- such that $\hat{p}_+(\boldsymbol{\omega}) = 0$ for $\|\boldsymbol{\omega}\| \geq s_+$ and $\hat{p}_-(\boldsymbol{\omega}) = 0$ for $\|\boldsymbol{\omega}\| \geq s_-$. For any translation invariant point potential function $\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x} - \mathbf{y})$, if its Fourier transform satisfies that $\Psi(\boldsymbol{\omega}) = 1$ for $\|\boldsymbol{\omega}\| < s = \max(s_+, s_-)$, the decision boundary of the potential function classifier (6.5) is identical to that of the Bayes classifier using conditional probability distributions $p(\mathbf{x}|+)$ and $p(\mathbf{x}|-)$, and class prior probabilities $\Pr(+)$ and $\Pr(-)$.

A proof of Theorem 2:

For a translation invariant ψ,

$$\int_{\mathbb{X}} p(\mathbf{y}|+)\psi(\mathbf{x}, \mathbf{y})d\mathbf{y} = \int_{\mathbb{X}} p(\mathbf{y}|+)\psi(\mathbf{x} - \mathbf{y})d\mathbf{y} = \mathcal{F}^{-1}[\hat{p}_+(\boldsymbol{\omega})\Psi(\boldsymbol{\omega})]$$

where $\mathcal{F}^{-1}$ is the inverse Fourier transform. Because $\hat{p}_+(\boldsymbol{\omega}) = 0$ for $\|\boldsymbol{\omega}\| \geq s$

and $\Psi(\boldsymbol{\omega}) = 1$ for $\|\boldsymbol{\omega}\| \leq s$, we have $\hat{p}_+(\boldsymbol{\omega})\Psi(\boldsymbol{\omega}) = \hat{p}_+(\boldsymbol{\omega})$. It follows that

$$\Pr(+) \int_{\mathbb{X}} p(\mathbf{y}|+)\psi(\mathbf{x},\mathbf{y})d\mathbf{y} = \Pr(+)p(\mathbf{x}|+) \propto \Pr(+|\mathbf{x}).$$

Similarly,

$$\Pr(-) \int_{\mathbb{X}} p(\mathbf{y}|-)\psi(\mathbf{x},\mathbf{y})d\mathbf{y} = \Pr(-)p(\mathbf{x}|-) \propto \Pr(-|\mathbf{x}).$$

The potential of the positive (negative) class hence is proportional to the posterior probability of the positive (negative) class. Therefore the decision boundary of (6.5) is identical to that of the Bayes classifier.

The above theorem states that if the charge densities are 'band limited' (i.e., its Fourier transform is zero everywhere outside a hyperball of finite radius s) and the point potential function has value 1 over the support of charge densities in the frequency domain, the potential function conveys the same information as the class conditional density. In the one dimensional case, a possible choice of ψ is a sinc function,

$$\psi(x,y) = \frac{\sin[2\pi s(x-y)]}{\pi(x-y)} = 2s \cdot \mathrm{sinc}[2s(x-y)],$$

whose Fourier transform is a rectangular window function

$$\Psi(\omega) = \begin{cases} 1 & |\omega| \leq s \\ 0 & |\omega| > s \end{cases} = \text{rect}\left(\frac{\omega}{2s}\right) \ .$$

This choice of Ψ can be generalized to higher dimensional spaces: for a hyper-rectangular window function in a d-dimensional frequency domain,

$$\Psi(\boldsymbol{\omega}) = \begin{cases} 1 & |\omega_i| \leq s, \forall i \in [1, d] \\ 0 & |\omega_i| > s, \exists i \in [1, d] \end{cases} = \prod_{i=1}^{d} \text{rect}\left(\frac{\omega_i}{2s}\right) \ ,$$

the corresponding point potential function is

$$\psi(\mathbf{x}, \mathbf{y}) = (2s)^d \prod_{i=1}^{d} \text{sinc}[2s(x_i - y_i)] \ . \tag{6.6}$$

Theorem 2 has implications on the practical design of potential function classifiers using a finite training set. This will be discussed in the next section.

6.5 Potential Function Rules as Plug-in Decision Rules

The main difficulty of using the potential function classifier (6.5) in practice is that charge densities are usually unknown. An approximation method is therefore presented in this section. Next, we first generalize the above binary potential function classifier to multiple classes. All the results discussed in Section 6.4 can be extended to the multi-class scenario. We then present an approximation on PFR and a discussion on its connection with plug-in decision rules.

6.5.1 An Approximation on Multi-class Potential Function classifiers

Let $z \in \mathbb{K} = \{1, \ldots, K\}$ be the class label of observation $\mathbf{x} \in \mathbb{X}$. The observation-label pair $(\mathbf{x}, z)$ is generated by a distribution F, which is a mixture of K unknown distributions $F_1, \ldots, F_K$,

$$F = \sum_{k=1}^{K} P_k F_k,$$

where P_k is the marginal probability of label k, i.e., $P_k = \Pr(z = k)$; F_k is the cumulative distribution function of $\mathbf{x}$ conditioned on $z = k$. Analogous to (6.1), (6.3), and (6.4), we define Φ_k as a class potential function - the potential with respect to $P_k F_k$:

$$\Phi_k(\mathbf{x}) = P_k \int_{\mathbb{X}} \psi(\mathbf{x}, \mathbf{y}) dF_k(\mathbf{y}) \ . \tag{6.7}$$

A multi-class potential classifier is defined as

$$f(\mathbf{x}) = \operatorname*{argmax}_k \Phi_k(\mathbf{x}) \ . \tag{6.8}$$

Note that the class potential (6.7) is the product of P_k and the expectation of the point potential function ψ with respect to F_k, i.e.,

$$\Phi_k(\mathbf{x}) = P_k \mathbb{E}_{\mathbf{y} \sim F_k}[\psi(\mathbf{x}, \mathbf{y})] \ .$$

Although F is unknown in most applications, a training set is usually given. Therefore, we approximate the above expectation by the sample mean. Let $\mathcal{S} = \{(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_\ell, z_\ell)\} \subset \mathbb{X} \times \mathbb{K}$ be the training set, a random i.i.d. sample from F.

Definition 1 (Sample Class Potential Function) Given a point potential function $\psi : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, we define the sample class potential of an observation $\mathbf{x}$ with respect to class k and sample $\mathcal{S}$ as

$$\phi_k(\mathbf{x}, \mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{z_i = k} \psi(\mathbf{x}, \mathbf{x}_i). \tag{6.9}$$

A multi-class sample potential classifier is then defined using sample class potential functions as follows.

Definition 2 (A Multi-class Sample Potential Function Classifier) Given $\mathcal{S}$, a set of i.i.d. training samples generated by an unknown distribution F on $\mathbb{X} \times \mathbb{K}$, we define a potential classifier $f_{\mathcal{S}} : \mathbb{X} \to \mathbb{K}$ as

$$f_{\mathcal{S}}(\mathbf{x}) = \operatorname*{argmax}_k \phi_k(\mathbf{x}, \mathcal{S}). \tag{6.10}$$

Clearly, the sample class potential (6.9) can be written as

$$\phi_k(\mathbf{x}, \mathcal{S}) = \frac{|\mathcal{S}_k|}{|\mathcal{S}|} \frac{1}{|\mathcal{S}_k|} \sum_{z_i = k} \psi(\mathbf{x}, \mathbf{x}_i) \ ,$$

where $\mathcal{S}_k = \{(\mathbf{x}, z) \in \mathcal{S} : z = k\}$. It is not difficult to observe that $\frac{|\mathcal{S}_k|}{|\mathcal{S}|}$ is an estimate of the marginal probability P_k. Furthermore, if we restrict ψ to be a nonnegative translation invariant function and $\int_{\mathbb{X}} \psi(\mathbf{x}) d\mathbf{x} = c < \infty$, it is straightforward to show that $\frac{1}{c|\mathcal{S}_k|} \sum_{z_i = k} \psi(\mathbf{x}, \mathbf{x}_i)$ is an estimate of the probability density of F_k at location $\mathbf{x}$ using the kernel density estimation (ψ is the kernel function). Hence, for any given $\mathbf{x}$, $\phi_k(\mathbf{x}, \mathcal{S})$ is proportional to an estimation of the posterior probability $\Pr(z = k|\mathbf{x})$.

This implies that the family of multi-class potential function classifiers (6.10) includes those plug-in decision Bayes classifiers that use kernel density estimation. Therefore, if ψ is chosen from regular kernels, the universal consistency of PFRs follows from the universal consistency of kernel rules (Devroye et al. (40)). Universal consistency characterizes an asymptotic property of a decision rule - a decision rule converges to the optimal solution as the number of training sample is sufficiently large. For kernel rules, universal consistency requires the 'width' of the kernel to decrease to 0 as the sample size increases to infinity. Next, we show that under the conditions of Theorem 2, for a fixed width of ψ (i.e., $\frac{1}{s}$ in (6.6) is fixed), with high probability the prediction of a sample PFR converges to that of the Bayes classifier for any given input.

6.5.2 The Potential Gap and the Generalization Performance

For a set of numbers $a_1, \ldots, a_K$, the k-th smallest number is denoted by $a_{(k)}$, i.e., $a_{(1)} \leq a_{(2)} \leq \cdots \leq a_{(K)}$. We define the *potential gap* of a multi-class classifier f given in (6.8) on an observation $\mathbf{x}$ by

$$\Gamma(\mathbf{x}) = \Phi_{f(\mathbf{x})}(\mathbf{x}) - \Phi_{(K-1)}(\mathbf{x}), \tag{6.11}$$

which is the difference between the largest class potential and the second largest class potential at $\mathbf{x}$. It should be clear that $\Gamma(\mathbf{x}) \geq 0$. The following theorem demonstrates that under the conditions of Theorem 2 (class conditional densities are band limited), the performance of a sample potential classifier (6.10) is closely related to the potential gap.

Theorem 3 Let $\mathcal{S} = \{(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_\ell, z_\ell)\} \subset \mathbb{R}^d \times \mathbb{K}$ be a random i.i.d. sample from F, a mixture of K distributions $F_1, \ldots, F_K$: $F = \sum_{k=1}^{K} P_k F_k$, where P_k is the marginal probability of class k; F_k, defined by a density function $p_k(\mathbf{x})$, is the distribution of $\mathbf{x}$ for class k. The conditional density functions are band limited, i.e., there exists $s > 0$ such that $\hat{p}_k(\boldsymbol{\omega}) = 0$ when $\|\boldsymbol{\omega}\| \geq s$ for all $k = 1, \ldots, K$, where $\hat{p}_k(\boldsymbol{\omega})$ is the Fourier transform of $p_k(\mathbf{x})$.

For any $\mathbf{x} \in \mathbb{R}^d$ the following inequality holds:

$$\Pr[f_{\mathcal{S}}(\mathbf{x}) \neq f^*(\mathbf{x})] \leq 2Ke^{-\frac{\ell\Gamma(\mathbf{x})^2}{2(2s)^{2d}}}, \tag{6.12}$$

where $f_{\mathcal{S}}(\mathbf{x})$ is the sample potential function classifier given in Definition 2 with (6.6) being the point potential function, $f^*(\mathbf{x})$ the Bayes classifier, and $\Gamma(\mathbf{x})$ the potential gap.

A proof of Theorem 3:

We need the following Lemma to prove Theorem 3.

Lemma 1 For any $a_1, a_2, \ldots, a_K \in \mathbb{R}$ and $b_1, b_2, \ldots, b_K \in \mathbb{R}$, if $|a_k - b_k| \leq \epsilon$ for all $k \in \mathbb{K}$, we have $|a_{(j)} - b_{(j)}| \leq \epsilon$ for all $j \in \mathbb{K}$.

Proof: For any $j \in \mathbb{K}$,

$$a_{(j)} - \epsilon \leq a_{(j+1)} - \epsilon \leq \cdots \leq a_{(K)} - \epsilon.$$

Because $b_k \geq a_k - \epsilon$ for all $k \in \mathbb{K}$, the number of b_k's that are greater than or equal to $a_{(j)} - \epsilon$ is at least $K - j + 1$. Therefore $b_{(j)} \geq a_{(j)} - \epsilon$. Similarly, for any $j \in \mathbb{K}$,

$$a_{(1)} + \epsilon \leq a_{(2)} + \epsilon \leq \cdots \leq a_{(j)} + \epsilon.$$

Because $b_k \leq a_k + \epsilon$ for all $k \in \mathbb{K}$, the number of b_k's that are less than or equal to $a_{(j)} + \epsilon$ is at least j. Therefore $b_{(j)} \leq a_{(j)} + \epsilon$. This completes the proof. □

Proof of Theorem 3: We introduce a new random variable $\mathbf{S}_k = |\{(\mathbf{x}, z) \in \mathcal{S} : z = k\}|$. For any given $\mathbf{x}$,

$$\begin{aligned}\mathbb{E}_F[\phi_k(\mathbf{x}, \mathcal{S})] &= \mathbb{E}_{\mathbf{S}_k}\left\{\mathbb{E}_{F|\mathbf{S}_k}\left[\frac{\mathbf{S}_k}{\ell}\frac{1}{\mathbf{S}_k}\sum_{z_i=k}\psi(\mathbf{x}, \mathbf{x}_i)\right]\right\} \\ &= \mathbb{E}_{\mathbf{S}_k}\left\{\frac{\mathbf{S}_k}{\ell}\mathbb{E}_{\mathbf{y}\sim F_k}[\psi(\mathbf{x}, \mathbf{y})]\right\} = P_k\mathbb{E}_{\mathbf{y}\sim F_k}[\psi(\mathbf{x}, \mathbf{y})] = \Phi_k(\mathbf{x}).\end{aligned}$$

We rewrite $\phi_k(\mathbf{x}, \mathcal{S})$ as $\phi_k(\mathbf{x}, \mathcal{S}) = \frac{1}{\ell}\sum_{i=1}^{\ell} I(z_i = k)\psi(\mathbf{x}, \mathbf{x}_i)$ where the indicator function $I(z_i = k) = 1$ if $z_i = k$, $I(z_i = k) = 0$ otherwise. Because $(\mathbf{x}_i, z_i)$'s are i.i.d., so are $I(z_i = k)\psi(\mathbf{x}, \mathbf{x}_i)$. In addition, from (6.6) it is clear that $|I(z_i = k)\psi(\mathbf{x}, \mathbf{x}_i)| \leq (2s)^d$. It follows from Hoeffding's inequality that for any given $\mathbf{x}$, $\epsilon > 0$, and $k = 1, \dots, K$,

$$\Pr[|\phi_k(\mathbf{x}, \mathcal{S}) - \Phi_k(\mathbf{x})| \geq \epsilon] \leq 2e^{-\frac{2\ell\epsilon^2}{(2s)^{2d}}} \ . \tag{6.13}$$

Because the conditional densities are band limited, it follows from the proof of Theorem 2 that

$$\Phi_k(\mathbf{x}) = P_k\mathbb{E}_{\mathbf{y}\sim F_k}[\psi(\mathbf{x}, \mathbf{y})] \propto \Pr(z = k|\mathbf{x}) \ .$$

Hence we have $f^*(\mathbf{x}) = \operatorname{argmax}_k \Phi_k(\mathbf{x})$. From Lemma 1, we know that if $|\phi_k(\mathbf{x}, \mathcal{S}) - \Phi_k(\mathbf{x})| \leq \frac{\Gamma(\mathbf{x})}{2}$ for $k = 1, \dots, K$, $|\phi_{(K)}(\mathbf{x}, \mathcal{S}) - \Phi_{(K)}(\mathbf{x})| \leq \frac{\Gamma(\mathbf{x})}{2}$.

Combining this with the facts that

$$\phi_{f_{\mathcal{S}}(\mathbf{x})}(\mathbf{x}, \mathcal{S}) = \phi_{(K)}(\mathbf{x}, \mathcal{S}) \text{ and } \Phi_{f^*(\mathbf{x})}(\mathbf{x}) = \Phi_{(K)}(\mathbf{x}) \ ,$$

it is straightforward to derive that $f_{\mathcal{S}}(\mathbf{x}) = f^*(\mathbf{x})$. Therefore,

$$\Pr\left[|\phi_k(\mathbf{x}, \mathcal{S}) - \Phi_k(\mathbf{x})| < \frac{\Gamma(\mathbf{x})}{2}, \ \forall\, k = 1, \ldots, K\right] \leq \Pr[f_{\mathcal{S}}(\mathbf{x}) = f^*(\mathbf{x})] \ . \tag{6.14}$$

Let $\epsilon = \frac{\Gamma(\mathbf{x})}{2}$. Using (6.13), (6.14), and the union bound, we have

$$\Pr[f_{\mathcal{S}}(\mathbf{x}) \neq f^*(\mathbf{x})] \leq \Pr\left[\exists k, \ |\phi_k(\mathbf{x}, \mathcal{S}) - \Phi_k(\mathbf{x})| \geq \frac{\Gamma(\mathbf{x})}{2}\right] \leq 2Ke^{-\frac{\ell\Gamma(\mathbf{x})^2}{2(2s)^{2d}}} \ .$$

This completes the proof.

Theorem 3 suggests that for any given $\mathbf{x}$ and a band limited joint probability density function, the probability that the sample potential function classifier behaves differently from the Bayes classifier depends on two parameters: the potential gap $\Gamma(\mathbf{x})$ and the sample size ℓ. The larger the potential gap and the sample size, the more likely that the sample potential function classifier makes the optimal prediction. In this sense, the generalization performance of $f_{\mathcal{S}}$ depends on the potential gap. Nevertheless, Theorem 3 does not tell us how to pick a sample size ℓ, neither could we compute the right hand side of the inequality (6.12), because the potential gap is unknown in practice. Motivated by the potential gap, we present, in the next section, a probabilistic bound on the generalization performance of a sample potential function classifier based on the margin of $f_{\mathcal{S}}$, which is closely related to the sample version of the potential gap.

6.6 A Generalization Bound for Potential Function Classifiers

As indicated in Definition 1, the sample class potential $\phi_k(\mathbf{x}, \mathcal{S})$ is an estimate of the class potential $\Phi_k(\mathbf{x})$. Analogous to the potential gap, we define the *margin* of $f_{\mathcal{S}}$ on an observation $(\mathbf{x}, z) \in \mathbb{R}^d \times \mathbb{K}$ as

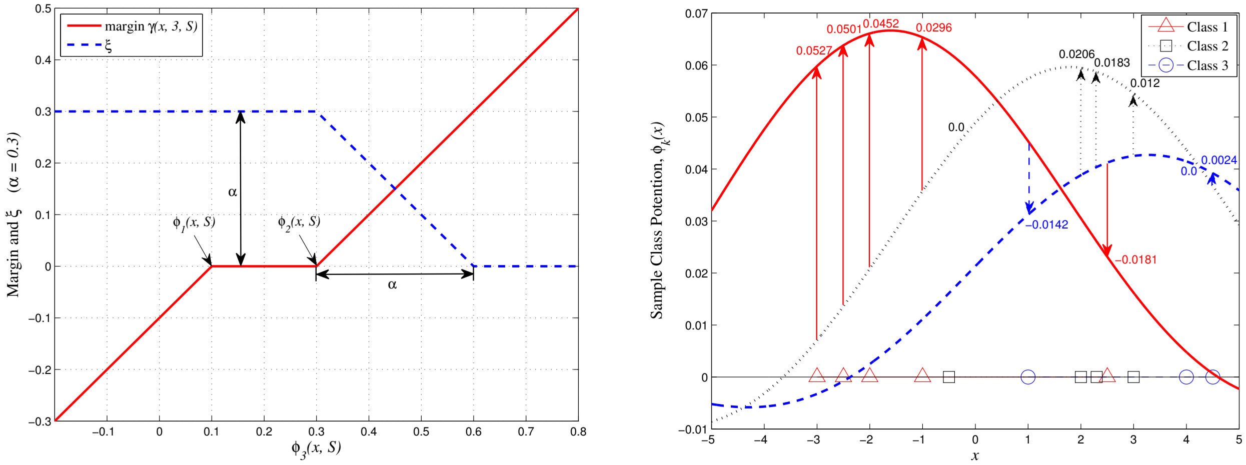
$$\gamma(\mathbf{x}, z, \mathcal{S}) = \phi_z(\mathbf{x}, \mathcal{S}) - \phi_{(K-1)}(\mathbf{x}, \mathcal{S}) \ . \tag{6.15}$$

Given a classifier $f_{\mathcal{S}}$ and a desired margin $\alpha > 0$, we denote by ξ the bounded amount by which $f_{\mathcal{S}}$ fails to achieve the desired margin α on sample $(\mathbf{x}, z)$,

$$\xi = \min\{\alpha, [\alpha - \gamma(\mathbf{x}, z, \mathcal{S})]_+\} \ ,$$

where $[x]_+ = x$ if $x \geq 0$ and 0 otherwise. For an observation $(\mathbf{x}_i, z_i) \in \mathcal{S}$, we define its margin shortage, ξ_i, as

$$\xi_i = \min\{\alpha, [\alpha - \gamma(\mathbf{x}_i, z_i, \mathcal{S}(i))]_+\} \ , \tag{6.16}$$



(a) Margin as a function of class potential.

(b) Sample class potential functions and margins.

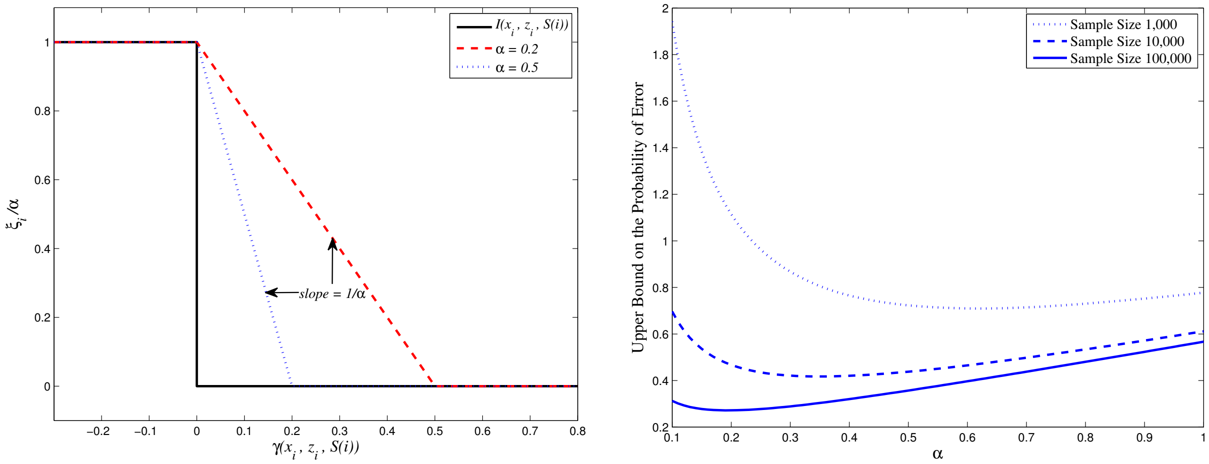
Figure 6.2: Sample class potential functions and margins under a 3-class scenario. (a) The solid curve describes the variation of margin $\gamma(\mathbf{x}, 3, \mathcal{S})$ with respect to the sample class potential $\phi_3(\mathbf{x}, \mathcal{S})$ when the sample class potential $\phi_1(\mathbf{x}, \mathcal{S})$ and $\phi_2(\mathbf{x}, \mathcal{S})$ are fixed. The dashed curve represents ξ, the bounded amount by which the margin is less than $\alpha = 0.3$. (b) The three curves represent sample class potential functions built upon 12 training observations (denoted by the markers on the horizontal axis) using a 1-d sinc point potential function with $s = 0.1$. Each arrow corresponds to a margin, which is computed as the difference between the vertical coordinate of the tip of the arrow and that of the end of the arrow. The numeric value of the margin is given along with the arrow. The arrow is absent if the margin is 0.

where $\mathcal{S}(i) = \mathcal{S} - \{(\mathbf{x}_i, z_i)\}$. Note that both ξ and $\xi_i \in [0, \alpha]$.

We illustrate the concepts of margin and ξ in Figure 6.2 under a 3-class scenario. The solid curve in Figure 6.2(a) shows the variations of the margin for an observation, $(\mathbf{x}, 3)$, as a function of its sample class potential $\phi_3(\mathbf{x}, \mathcal{S})$. The sample class potentials of $\mathbf{x}$ with respect to class 1 and class 2, i.e, $\phi_1(\mathbf{x}, \mathcal{S})$ and $\phi_2(\mathbf{x}, \mathcal{S})$, are fixed. For a desired margin $\alpha = 0.3$, the dashed curve represents the value of ξ: the bounded amount by which the margin is less than α. Figure 6.2(b) shows three sample class potential functions constructed from 12 training observations. Each class is associated with a distinct marker: circle, triangle, or square. The point potential function defined in (6.6) with $s = 0.1$ is used in the evaluation of the sample class potential functions. We visualize the margin for each training observation using an arrow where the margin is computed as the difference between the vertical coordinate of the tip of the arrow ($\phi_z(\mathbf{x}, \mathcal{S})$) and that of the end of the arrow ($\phi_{(2)}(\mathbf{x}, \mathcal{S})$). The numerical value of a margin is also listed along with the arrow. For observations $(-0.5, 2)$ and $(4, 3)$, the arrows are absent because their margins are 0.

It is not difficult to relate margins to classification errors. Positive margins suggest correct classifications. Negative margins imply mis-classifications. There are only two scenarios that result in the 0 margin: $\phi_z(\mathbf{x}, \mathcal{S}) = \phi_{(K)}(\mathbf{x}, \mathcal{S}) = \phi_{(K-1)}(\mathbf{x}, \mathcal{S})$ or $\phi_{(K)}(\mathbf{x}, \mathcal{S}) > \phi_z(\mathbf{x}, \mathcal{S}) = \phi_{(K-1)}(\mathbf{x}, \mathcal{S})$. In the former case, which is rare in practice, the correctness of the classification depends on the tie breaking strategy, which is usually random. The second case is more common, for example the two 0 margins in Figure 6.2(b). It leads to mis-classifications. If we introduce the following indicator function

$$I(\mathbf{x}, z, \mathcal{S}) = \begin{cases} 1 & \gamma(\mathbf{x}, z, \mathcal{S}) \leq 0 \\ 0 & \text{otherwise} \end{cases}, \tag{6.17}$$



(a) Bounded margin shortage.

(b) An upper bound on generalization performance.

Figure 6.3: (a) Plots of ξ_i, the bounded margin shortage, as a function of the margin $\gamma(\mathbf{x}_i, z_i, \mathcal{S}(i))$. When α approaches 0, $\frac{\xi_i}{\alpha}$ converges to the indicator function $I(\mathbf{x}_i, z_i, \mathcal{S}(i))$. (b) Plots of the upper bound on the probability of error in (6.19) as a function of the desired margin α.

$\sum_{i=1}^{\ell} I(\mathbf{x}_i, z_i, \mathcal{S}(i))$ is an upper bound on the number of mis-classified observations in a leave-one-out evaluation.

The connection between the bounded margin shortage ξ_i, which is defined in (6.16), and a classification error is more subtle. If we divide ξ_i by α, we have

$$\frac{\xi_i}{\alpha} = \begin{cases} 1 & \gamma(\mathbf{x}_i, z_i, \mathcal{S}(i)) \le 0 \\ 1 - \frac{\gamma(\mathbf{x}_i, z_i, \mathcal{S}(i))}{\alpha} & 0 < \gamma(\mathbf{x}_i, z_i, \mathcal{S}(i)) \le \alpha \\ 0 & 0 < \gamma(\mathbf{x}_i, z_i, \mathcal{S}(i)) \end{cases} . \tag{6.18}$$

Figure 6.3(a) compares $\frac{\xi_i}{\alpha}$ with $I(\mathbf{x}_i, z_i, \mathcal{S}(i))$ as a function of $\gamma(\mathbf{x}_i, z_i, \mathcal{S}(i))$. It is clear that $\frac{\xi_i}{\alpha}$ is always greater than or equal to $I(\mathbf{x}_i, z_i, \mathcal{S}(i))$. Therefore, $\sum_{i=1}^{\ell} \frac{\xi_i}{\alpha}$ is an upper bound on the number of mis-classified observations in a leave-one-out evaluation. Next, we present a generalization bound based on the desired margin α and the bounded margin shortage ξ_i for any given bounded point potential function ψ. Without loss of generality, we assume that $\psi : \mathbb{X} \times \mathbb{X} \to [0, 1]$.

Theorem 4 Let $\mathcal{S} = \{(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_\ell, z_\ell)\} \subset \mathbb{X} \times \mathbb{K}$ be a random i.i.d. sample from an unknown distribution F, and $f_\mathcal{S} : \mathbb{X} \to \mathbb{K}$ a sample potential function classifier defined according to (6.10) using a given point potential function $\psi : \mathbb{X} \times \mathbb{X} \to [0, 1]$. For a fixed $\delta \in (0, 1)$, a desired margin $\alpha > 0$, and a new random sample $(\mathbf{x}, z)$ generated from F, the following bound holds with probability at least $1 - \delta$ over $\mathcal{S}$:

$$\Pr_F \left[z \neq f_\mathcal{S}(\mathbf{x}) | \mathcal{S} \right] \leq \frac{1}{\ell} \sum_{i=1}^{\ell} \frac{\xi_i}{\alpha} + \frac{2}{\ell\alpha} + \left(1 + \frac{4}{\alpha}\right) \sqrt{\frac{\ln(2/\delta)}{2\ell}} \,. \tag{6.19}$$

where ξ_i is defined in (6.16).

A proof of Theorem 4:

In order to prove the upper bound on generalization of sample potential function classifiers in Theorem 4, we need the following Lemma and an inequality attributed to McDiarmid.

Lemma 2 Let $\mathcal{S}(i) = \mathcal{S} - \{(\mathbf{x}_i, z_i)\}$. For a change of one $(\mathbf{x}_t, z_t)$ to $(\hat{\mathbf{x}}_t, \hat{z}_t)$, denote

$$\hat{\mathcal{S}}_t = \{(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_{t-1}, z_{t-1}), (\hat{\mathbf{x}}_t, \hat{z}_t), (\mathbf{x}_{t+1}, z_{t+1}), \ldots, (\mathbf{x}_\ell, z_\ell)\}.$$

We define $\hat{\mathcal{S}}_t(i) = \hat{\mathcal{S}}_t - \{(\mathbf{x}_i, z_i)\}$, hence $\hat{\mathcal{S}}_t(t) = \mathcal{S}(t)$. Let $\mathbf{x} \in \mathbb{X}$ be any observation in $\mathbb{X}$ and $z \in \mathbb{K}$ a class label. The following inequalities hold for any point potential function $\psi : \mathbb{X} \times \mathbb{X} \to [0, 1]$:

$$\begin{aligned}
|\gamma(\mathbf{x}, z, \mathcal{S}) - \gamma(\mathbf{x}, z, \mathcal{S}(i))| &\leq \frac{2}{\ell}, && (6.20)\\
\left|\gamma(\mathbf{x}, z, \mathcal{S}(i)) - \gamma(\mathbf{x}, z, \hat{\mathcal{S}}_t(i))\right| &\leq \frac{2}{\ell - 1}, && (6.21)\\
\left|\gamma(\mathbf{x}_i, z_i, \mathcal{S}(i)) - \gamma(\mathbf{x}_i, z_i, \hat{\mathcal{S}}_t(i))\right| &\leq \frac{2}{\ell - 1}, \text{ if } i \neq t. && (6.22)
\end{aligned}$$

Proof: It is readily checked that for any $z \in \mathbb{K}$,

$$\begin{aligned}
|\phi_z(\mathbf{x}, \mathcal{S}) - \phi_z(\mathbf{x}, \mathcal{S}(i))| &= \begin{cases} \left|\frac{1}{\ell}\sum_{z_j = z}\psi(\mathbf{x}, \mathbf{x}_j) - \frac{1}{\ell - 1}\sum_{z_j = z, j \neq i}\psi(\mathbf{x}, \mathbf{x}_j)\right| & \text{if } z = z_i \\ \left|\frac{1}{\ell}\sum_{z_j = z}\psi(\mathbf{x}, \mathbf{x}_j) - \frac{1}{\ell - 1}\sum_{z_j = z}\psi(\mathbf{x}, \mathbf{x}_j)\right| & \text{if } z \neq z_i \end{cases}\\
&= \begin{cases} \left|\frac{1}{\ell}\psi(\mathbf{x}, \mathbf{x}_i) - \frac{1}{\ell(\ell - 1)}\sum_{z_j = z, j \neq i}\psi(\mathbf{x}, \mathbf{x}_j)\right| \leq \frac{1}{\ell} & \text{if } z = z_i \\ \left|\frac{1}{\ell(\ell - 1)}\sum_{z_j = z}\psi(\mathbf{x}, \mathbf{x}_j)\right| \leq \frac{1}{\ell} & \text{if } z \neq z_i \end{cases}.
\end{aligned}$$

From (6.15) we have

$$\begin{aligned}
|\gamma(\mathbf{x}, z, \mathcal{S}) - \gamma(\mathbf{x}, z, \mathcal{S}(i))| &= \left|\phi_z(\mathbf{x}, \mathcal{S}) - \phi_{(K-1)}(\mathbf{x}, \mathcal{S}) - \phi_z(\mathbf{x}, \mathcal{S}(i)) + \phi_{(K-1)}(\mathbf{x}, \mathcal{S}(i))\right|\\
&\leq |\phi_z(\mathbf{x}, \mathcal{S}) - \phi_z(\mathbf{x}, \mathcal{S}(i))| + \left|\phi_{(K-1)}(\mathbf{x}, \mathcal{S}) - \phi_{(K-1)}(\mathbf{x}, \mathcal{S}(i))\right|\\
&\leq \frac{1}{\ell} + \left|\phi_{(K-1)}(\mathbf{x}, \mathcal{S}) - \phi_{(K-1)}(\mathbf{x}, \mathcal{S}(i))\right| \leq \frac{2}{\ell},
\end{aligned}$$

where the last step is based on Lemma 1.

It is not difficult to show that for any $z \in \mathbb{K}$, $\left|\phi_z(\mathbf{x}, \mathcal{S}(i)) - \phi_z(\mathbf{x}, \hat{\mathcal{S}}_t(i))\right| = 0$ when $i = t$, otherwise,

$$\left|\phi_z(\mathbf{x}, \mathcal{S}(i)) - \phi_z(\mathbf{x}, \hat{\mathcal{S}}_t(i))\right| = \begin{cases} 0 & \text{if } z_t \neq z,\ \hat{z}_t \neq z \\ \left|\frac{1}{\ell-1}\psi(\mathbf{x}, \hat{\mathbf{x}}_t)\right| & \text{if } z_t \neq z,\ \hat{z}_t = z \\ \left|\frac{1}{\ell-1}\psi(\mathbf{x}, \mathbf{x}_t) - \frac{1}{\ell-1}\psi(\mathbf{x}, \hat{\mathbf{x}}_t)\right| & \text{if } z_t = z,\ \hat{z}_t = z \\ \left|\frac{1}{\ell-1}\psi(\mathbf{x}, \mathbf{x}_t)\right| & \text{if } z_t = z,\ \hat{z}_t \neq z \end{cases} \leq \frac{1}{\ell - 1}\ .$$

Therefore,

$$\begin{aligned} \left|\gamma(\mathbf{x}, z, \mathcal{S}(i)) - \gamma(\mathbf{x}, z, \hat{\mathcal{S}}_t(i))\right| &= \left|\phi_z(\mathbf{x}, \mathcal{S}(i)) - \phi_{(K-1)}(\mathbf{x}, \mathcal{S}(i)) - \phi_z(\mathbf{x}, \hat{\mathcal{S}}_t(i)) + \phi_{(K-1)}(\mathbf{x}, \hat{\mathcal{S}}_t(i))\right| \\ &\leq \left|\phi_z(\mathbf{x}, \mathcal{S}(i)) - \phi_z(\mathbf{x}, \hat{\mathcal{S}}_t(i))\right| + \left|\phi_{(K-1)}(\mathbf{x}, \mathcal{S}(i)) - \phi_{(K-1)}(\mathbf{x}, \hat{\mathcal{S}}_t(i))\right| \\ &\leq \frac{2}{\ell - 1}. \end{aligned}$$

Finally, for $i \neq t$ and any $z_i \in \mathbb{K}$,

$$\left|\phi_{z_i}(\mathbf{x}_i, \mathcal{S}(i)) - \phi_{z_i}(\mathbf{x}_i, \hat{\mathcal{S}}_t(i))\right| = \begin{cases} 0 & \text{if } z_t \neq z_i,\ \hat{z}_t \neq z_i \\ \left|\frac{1}{\ell-1}\psi(\mathbf{x}_i, \hat{\mathbf{x}}_t)\right| & \text{if } z_t \neq z_i,\ \hat{z}_t = z_i \\ \left|\frac{1}{\ell-1}\psi(\mathbf{x}_i, \mathbf{x}_t) - \frac{1}{\ell-1}\psi(\mathbf{x}_i, \hat{\mathbf{x}}_t)\right| & \text{if } z_t = z_i,\ \hat{z}_t = z_i \\ \left|\frac{1}{\ell-1}\psi(\mathbf{x}_i, \mathbf{x}_t)\right| & \text{if } z_t = z_i,\ \hat{z}_t \neq z_i \end{cases} \leq \frac{1}{\ell - 1}\ .$$

Therefore,

$$\begin{aligned}
&\left|\gamma(\mathbf{x}_i, z_i, \mathcal{S}(i)) - \gamma(\mathbf{x}_i, z_i, \hat{\mathcal{S}}_t(i))\right| \\
&= \left|\phi_{z_i}(\mathbf{x}_i, \mathcal{S}(i)) - \phi_{(K-1)}(\mathbf{x}_i, \mathcal{S}(i)) - \phi_{z_i}(\mathbf{x}_i, \hat{\mathcal{S}}_t(i)) + \phi_{(K-1)}(\mathbf{x}_i, \hat{\mathcal{S}}_t(i))\right| \\
&\leq \left|\phi_{z_i}(\mathbf{x}_i, \mathcal{S}(i)) - \phi_{z_i}(\mathbf{x}_i, \hat{\mathcal{S}}_t(i))\right| + \left|\phi_{(K-1)}(\mathbf{x}_i, \mathcal{S}(i)) - \phi_{(K-1)}(\mathbf{x}_i, \hat{\mathcal{S}}_t(i))\right| \leq \frac{2}{\ell-1}.
\end{aligned}$$

This completes the proof. □

Lemma 3 (McDiarmid's Inequality) *Let $X_1, X_2, \ldots, X_n$ be independent random variables taking values in a set $\mathbb{X}$. Suppose that $f : \mathbb{X}^n \to \mathbb{R}$ satisfies*

$$\sup_{\mathbf{x}_1,\ldots,\mathbf{x}_n,\hat{\mathbf{x}}_j \in \mathbb{X}} |f(\mathbf{x}_1, \ldots, \mathbf{x}_n) - f(\mathbf{x}_1, \ldots, \hat{\mathbf{x}}_j, \ldots, \mathbf{x}_n)| \leq c_j$$

for constants $c_j, 1 \leq j \leq n$. Then for every $\epsilon > 0$,

$$\Pr[f(X_1, \ldots, X_n) - \mathbb{E}f \geq \epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{j=1}^n c_j^2}\right).$$

Proof of Theorem 4: Consider the loss function

$$g(\mathbf{x}, z, \mathcal{S}) = \begin{cases} 1, & \text{if } \gamma(\mathbf{x}, z, \mathcal{S}) \leq 0, \\ \frac{\alpha - \gamma(\mathbf{x},z,\mathcal{S})}{\alpha}, & \text{if } 0 < \gamma(\mathbf{x}, z, \mathcal{S}) \leq \alpha, \\ 0, & \text{otherwise.} \end{cases}$$

It is not difficult to show that

$$\Pr_F\left[z \neq f_{\mathcal{S}}(\mathbf{x}) | \mathcal{S}\right] \leq \mathbb{E}_{F|\mathcal{S}}\left[g(\mathbf{x}, z, \mathcal{S})\right] \ ,$$

where the equality holds when $\alpha = 0$. Hence it suffices to show that $\mathbb{E}_{F|\mathcal{S}}\left[g(\mathbf{x}, z, \mathcal{S})\right]$ is bounded by the right side of (6.19).

We break $\mathbb{E}_{F|\mathcal{S}}[g(\mathbf{x}, z, \mathcal{S})] - \frac{1}{\ell\alpha}\sum_{i=1}^{\ell}\xi_i = \mathbb{E}_{F|\mathcal{S}}[g(\mathbf{x}, z, \mathcal{S})] - \frac{1}{\ell}\sum_{i=1}^{\ell} g(\mathbf{x}_i, z_i, \mathcal{S}(i))$ into $A + B + C$:

$$\begin{aligned} A &= \mathbb{E}_{F|\mathcal{S}}[g(\mathbf{x}, z, \mathcal{S})] - \mathbb{E}_{F|\mathcal{S}}\left[\frac{1}{\ell}\sum_{i=1}^{\ell} g(\mathbf{x}, z, \mathcal{S}(i))\right], \\ B &= \mathbb{E}_{F|\mathcal{S}}\left[\frac{1}{\ell}\sum_{i=1}^{\ell} g(\mathbf{x}, z, \mathcal{S}(i))\right] - \mathbb{E}_{F}[g(\mathbf{x}_j, z_j, \mathcal{S}(j))], \\ C &= \mathbb{E}_{F}[g(\mathbf{x}_j, z_j, \mathcal{S}(j))] - \frac{1}{\ell}\sum_{i=1}^{\ell} g(\mathbf{x}_i, z_i, \mathcal{S}(i)), \end{aligned}$$

where $(\mathbf{x}_j, z_j)$ is any fixed sample in $\mathcal{S}$.

We first look at A. It is straightforward to show that

$$|g(\mathbf{x}, z, \mathcal{S}) - g(\mathbf{x}, z, \mathcal{S}(i))| \le \frac{1}{\alpha}|\gamma(\mathbf{x}, z, \mathcal{S}) - \gamma(\mathbf{x}, z, \mathcal{S}(i))| \le \frac{2}{\ell\alpha},$$

where the last inequality is based on (6.20). Therefore

$$A = \mathbb{E}_{F|\mathcal{S}}\left\{\frac{1}{\ell}\sum_{i=1}^{\ell}[g(\mathbf{x}, z, \mathcal{S}) - g(\mathbf{x}, z, \mathcal{S}(i))]\right\} \le \mathbb{E}_{F|\mathcal{S}}\left|\frac{1}{\ell}\sum_{i=1}^{\ell}[g(\mathbf{x}, z, \mathcal{S}) - g(\mathbf{x}, z, \mathcal{S}(i))]\right| \le \frac{2}{\ell\alpha}. \tag{6.23}$$

Next, we look at B. It is not difficult to verify that

$$\mathbb{E}_{F}\left\{\mathbb{E}_{F|\mathcal{S}}\left[\frac{1}{\ell}\sum_{i=1}^{\ell} g(\mathbf{x}, z, \mathcal{S}(i))\right]\right\} = \mathbb{E}_{F}[g(\mathbf{x}_j, z_j, \mathcal{S}(j))].$$

For a change of one $(\mathbf{x}_t, z_t)$ to $(\hat{\mathbf{x}}_t, \hat{z}_t)$, we denote

$$\hat{\mathcal{S}}_t = \{(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_{t-1}, z_{t-1}), (\hat{\mathbf{x}}_t, \hat{z}_t), (\mathbf{x}_{t+1}, z_{t+1}), \ldots, (\mathbf{x}_\ell, z_\ell)\}.$$

From (6.21) we have for any $z \in \mathbb{K}$

$$\left| g(\mathbf{x}, z, \mathcal{S}(i)) - g(\mathbf{x}, z, \hat{\mathcal{S}}_t(i)) \right| \leq \frac{1}{\alpha} \left| \gamma(\mathbf{x}, z, \mathcal{S}(i)) - \gamma(\mathbf{x}, z, \hat{\mathcal{S}}_t(i)) \right| \leq \frac{2}{\alpha(\ell - 1)}.$$

Therefore,

$$\begin{aligned}
& \sup_{(\mathbf{x}_1,z_1),\ldots,(\mathbf{x}_\ell,z_\ell),(\hat{\mathbf{x}}_t,\hat{z}_t)} \left| \mathbb{E}_{F|\mathcal{S}} \left[\frac{1}{\ell} \sum_{i=1}^{\ell} g(\mathbf{x}, z, \mathcal{S}(i)) \right] - \mathbb{E}_{F|\hat{\mathcal{S}}_t} \left[\frac{1}{\ell} \sum_{i=1}^{\ell} g(\mathbf{x}, z, \hat{\mathcal{S}}_t(i)) \right] \right| \\
= & \sup_{(\mathbf{x}_1,z_1),\ldots,(\mathbf{x}_\ell,z_\ell),(\hat{\mathbf{x}}_t,\hat{z}_t)} \frac{1}{\ell} \left| \sum_{i=1}^{\ell} \mathbb{E}_{F|\mathcal{S},\hat{\mathcal{S}}_t} \left[g(\mathbf{x}, z, \mathcal{S}(i)) - g(\mathbf{x}, z, \hat{\mathcal{S}}_t(i)) \right] \right| \\
\leq & \sup_{(\mathbf{x}_1,z_1),\ldots,(\mathbf{x}_\ell,z_\ell),(\hat{\mathbf{x}}_t,\hat{z}_t)} \frac{1}{\ell} \sum_{i=1}^{\ell} \mathbb{E}_{F|\mathcal{S},\hat{\mathcal{S}}_t} \left| g(\mathbf{x}, z, \mathcal{S}(i)) - g(\mathbf{x}, z, \hat{\mathcal{S}}_t(i)) \right| \\
= & \sup_{(\mathbf{x}_1,z_1),\ldots,(\mathbf{x}_\ell,z_\ell),(\hat{\mathbf{x}}_t,\hat{z}_t)} \frac{1}{\ell} \sum_{i=1, i\neq t}^{\ell} \mathbb{E}_{F|\mathcal{S},\hat{\mathcal{S}}_t} \left| g(\mathbf{x}, z, \mathcal{S}(i)) - g(\mathbf{x}, z, \hat{\mathcal{S}}_t(i)) \right| \leq \frac{2}{\alpha \ell}. \qquad (6.24)
\end{aligned}$$

By (6.24), we apply the McDiamid's inequality to get

$$\Pr(B > \epsilon_1) \leq \exp\left(\frac{-\alpha^2 \ell \epsilon_1^2}{2} \right). \qquad (6.25)$$

Next, we look at C. It is clear that

$$\mathbb{E}_F \left[\frac{1}{\ell} \sum_{i=1}^{\ell} g(\mathbf{x}_i, z_i, \mathcal{S}(i)) \right] = \mathbb{E}_F[g(\mathbf{x}_j, z_j, \mathcal{S}(j))].$$

Let $\bar{g}(\mathcal{S}) = \frac{1}{\ell}\sum_{i=1}^{\ell} g(\mathbf{x}_i, z_i, \mathcal{S}(i))$. For a change of one $(\mathbf{x}_t, z_t)$ to $(\hat{\mathbf{x}}_t, \hat{z}_t)$, denote

$$\hat{\mathcal{S}}_t = \{(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_{t-1}, z_{t-1}), (\hat{\mathbf{x}}_t, \hat{z}_t), (\mathbf{x}_{t+1}, z_{t+1}), \ldots, (\mathbf{x}_\ell, z_\ell)\}.$$

For any $i \neq t$, it follows from (6.22) that for any $z_i \in \mathbb{K}$

$$\left| g(\mathbf{x}_i, z_i, \mathcal{S}(i)) - g(\mathbf{x}_i, z_i, \hat{\mathcal{S}}_t(i)) \right| \leq \frac{1}{\alpha} \left| \gamma(\mathbf{x}_i, z_i, \mathcal{S}(i)) - \gamma(\mathbf{x}_i, z_i, \hat{\mathcal{S}}_t(i)) \right| \leq \frac{2}{\alpha(\ell - 1)}.$$

Therefore,

$$\begin{aligned}
&\sup_{(\mathbf{x}_1,z_1),\ldots,(\mathbf{x}_\ell,z_\ell),(\hat{\mathbf{x}}_t,\hat{z}_t)} \left| \bar{g}(\mathcal{S}) - \bar{g}(\hat{\mathcal{S}}_t) \right| \\
&= \sup_{(\mathbf{x}_1,z_1),\ldots,(\mathbf{x}_\ell,z_\ell),(\hat{\mathbf{x}}_t,\hat{z}_t)} \left| \frac{1}{\ell}\sum_{i=1}^{\ell} g(\mathbf{x}_i, z_i, \mathcal{S}(i)) - \frac{1}{\ell}\left[g(\hat{\mathbf{x}}_t, \hat{z}_t, \hat{\mathcal{S}}_t(t)) + \sum_{i=1, i\neq t}^{\ell} g(\mathbf{x}_i, z_i, \hat{\mathcal{S}}_t(i)) \right] \right| \\
&\leq \frac{1}{\ell} + \sup_{(\mathbf{x}_1,z_1),\ldots,(\mathbf{x}_\ell,z_\ell),(\hat{\mathbf{x}}_t,\hat{z}_t)} \frac{1}{\ell} \left| \sum_{i=1, i\neq t}^{\ell} \left[g(\mathbf{x}_i, z_i, \mathcal{S}(i)) - g(\mathbf{x}_i, z_i, \hat{\mathcal{S}}_t(i)) \right] \right| \leq \frac{1}{\ell} + \frac{2}{\alpha\ell}. \qquad (6.26)
\end{aligned}$$

By (6.26), we apply the McDiarmid's inequality to get

$$\Pr(C > \epsilon_2) \leq \exp\left(\frac{-2\ell\epsilon_2^2}{\left(1 + \frac{2}{\alpha}\right)^2} \right). \qquad (6.27)$$

Finally, setting

$$\exp\left(\frac{-\gamma^2 \ell \epsilon_1^2}{2} \right) = \exp\left(\frac{-2\ell\epsilon_2^2}{\left(1 + \frac{2}{\alpha}\right)^2} \right) = \frac{\delta}{2}$$

and solving for ϵ_1 and ϵ_2, we obtain

$$\epsilon_1 = \frac{1}{\alpha}\sqrt{\frac{2\ln(2/\delta)}{\ell}}, \ \ \epsilon_2 = \left(\frac{1}{2}+\frac{1}{\alpha}\right)\sqrt{\frac{2\ln(2/\delta)}{\ell}} \ .$$

Because $B + C > \epsilon_1 + \epsilon_2$ implies $B > \epsilon_1$ or $C > \epsilon_2$,

$$\Pr(B + C > \epsilon_1 + \epsilon_2) \le \Pr(B > \epsilon_1 \text{ or } C > \epsilon_2) \le \Pr(B > \epsilon_1) + \Pr(C > \epsilon_2) \le \delta.$$

So, with probability at least $1-\delta$, $B + C \le \epsilon_1 + \epsilon_2$. Because $A \le \frac{2}{\ell\alpha}$, $B + C \le \epsilon_1 + \epsilon_2$ implies that $A + B + C \le \epsilon_1 + \epsilon_2 + \frac{2}{\ell\alpha}$. Therefore, with probability at least $1-\delta$, $A + B + C \le \epsilon_1 + \epsilon_2 + \frac{2}{\ell\alpha}$, i.e.

$$\mathbb{E}_{F|\mathcal{S}}[g(\mathbf{x}, z, \mathcal{S})] \le \frac{1}{\ell}\sum_{i=1}^{\ell} g(\mathbf{x}_i, z_i, \mathcal{S}(i)) + \frac{2}{\ell\alpha} + \left(1+\frac{4}{\alpha}\right)\sqrt{\frac{\ln(2/\delta)}{2\ell}} \ .$$

It is easy to verify that $\frac{1}{\ell}\sum_{i=1}^{\ell} g(\mathbf{x}_i, z_i, \mathcal{S}(i)) = \frac{1}{\ell}\sum_{i=1}^{\ell}\frac{\xi_i}{\alpha}$. Therefore, with probability at least $1-\delta$,

$$\mathbb{E}_{F|\mathcal{S}}[g(\mathbf{x}, z, \mathcal{S})] \le \frac{1}{\ell}\sum_{i=1}^{\ell}\frac{\xi_i}{\alpha} + \frac{2}{\ell\alpha} + \left(1+\frac{4}{\alpha}\right)\sqrt{\frac{\ln(2/\delta)}{2\ell}} \ .$$

This completes the proof.

It is worthwhile to note that there are two sources of randomness in the above inequality: the random sample $\mathcal{S}$ and the random observation $(\mathbf{x}, z)$. For a specific $\mathcal{S}$, the above bound is either true of false, i.e., it is not random.

For a random sample $\mathcal{S}$, the probability that the bound is true is at least $1-\delta$. The inequality shows that the error probability, $\Pr_F\left[z \neq f_{\mathcal{S}}(\mathbf{x})|\mathcal{S}\right]$, of a sample potential function classifier depends on three terms. The first term, $\frac{1}{\ell}\sum_{i=1}^{\ell}\frac{\xi_i}{\alpha}$, is an upper bound on the leave-one-out training error. The second and the third terms are determined by the training sample size ℓ, the desired margin α, and the confidence parameter δ. In general, for fixed ℓ and δ, the generalization performance of $f_{\mathcal{S}}$ a trade-off between training error and the desired margin α. On one hand, a smaller α produces a tighter bound on the training error, but larger values for the second and the third term. On the other hand, a larger α can reduce the values of the second and the third term, but makes the first term a looser bound on the training error. This is illustrated in Figure 6.3(b) using margins generated from a uniform distribution on $[-0.1, 1]$. The values of the upper bound are shown as a function of the desired margin α. In the next section, we discuss classifier selection methods motivated by the above bound on the generalization performance.

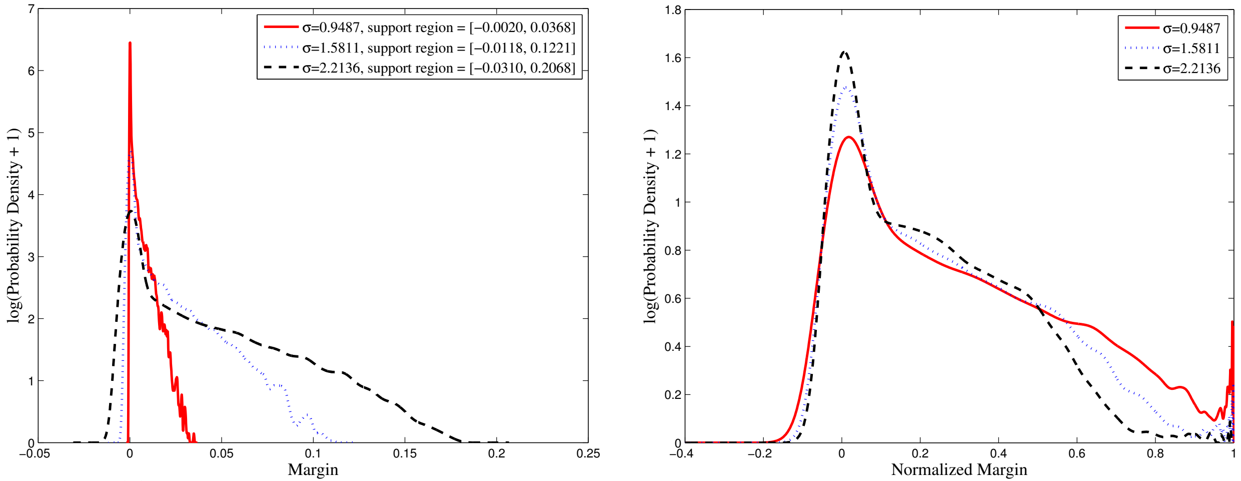
6.7 Margin Distributions and Classifier Selection

The learning of a potential function classifier is essentially the selection of a point potential function (or its parameters). Figure 6.3(b) shows that given ℓ and δ, the upper bound on the probability of error has a minimum. Hence it is tempting to choose a classifier that minimizes the upper bound in (6.19). Unfortunately, this is not an effective approach in practice because the bound is usually loose even for large training sets with $50,000$–$100,000$ observations.

As we discussed in Section 6.6, the desired margin α plays a key role in estimating the generalization performance. If we define $i^* = \operatorname{argmin}_{i=1,\ldots,\ell,\gamma(\mathbf{x}_i,z_i,\mathcal{S}(i))>0} \gamma(\mathbf{x}_i, z_i, \mathcal{S}(i))$, it is clear from Figure 6.3(a) that $\frac{1}{\ell}\sum_{i=1}^{\ell}\frac{\xi_i}{\alpha}$ achieves the minimum (which is equal to the training error) when $0 < \alpha \leq \gamma(\mathbf{x}_{i^*}, z_{i^*}, \mathcal{S}(i^*))$ [1]. Although a larger value of α decreases the values of the last two terms in (6.19), it also increases the value of $\frac{1}{\ell}\sum_{i=1}^{\ell}\frac{\xi_i}{\alpha}$. However, for a fixed value of α, the bound is tigher if the margins are concentrated more towards the positive end than towards the negative end. This suggest that we may select classifiers based on the distribution of margins.

However, a direct comparison of margin distributions may not be meaningful because the support region of a margin distribution largely depends on the selected point potential function ψ and its parameters. For example, Figure 6.4(a) shows the probability distributions of margins under a Gaussian point potential function (i.e., $\psi(\mathbf{x}, \mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{\sigma^2}}$) using the MAGIC dataset from UCI Machine Learning Repository (details of the dataset is given in Section 6.8). The support region of the margin distribution varies signifi-

[1] This is because there will be no observations whose margin falls into the sloped region. Hence $\frac{1}{\ell}\sum_{i=1}^{\ell}\frac{\xi_i}{\alpha} = \frac{1}{\ell}\sum_{i=1}^{\ell} I(\mathbf{x}_i, z_i, \mathcal{S}(i))$, which is the leave-one-out training error.



(a) Probability density of margin.

(b) Probability density of normalization margin.

Figure 6.4: Distributions of margin and normalized margin under a Gaussian point potential function $e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{\sigma^2}}$ with different values of σ.

cantly with the values of σ: when $\sigma = 0.9487$, the support region is the interval $[-0.0020, 0.0368]$; when $\sigma = 1.5811$, the support region is the interval $[-0.0118, 0.1221]$; when $\sigma = 2.2136$, the support region is the interval $[-0.0310, 0.2068]$. Therefore, a margin with value 0.03 is on the high end for $\sigma = 0.9487$, but is on the low end for $\sigma = 2.2136$.

To make margins comparable under different point potential functions or different parameter values, we propose the following normalization procedure. For any given $\mathbf{x} \in \mathbb{R}^d$, we define a normalized sample class potential, $\hat{\phi}_k(\mathbf{x}, \mathcal{S})$, as

$$\hat{\phi}_k(\mathbf{x}, \mathcal{S}) = \frac{\phi_k(\mathbf{x}, \mathcal{S})}{\sum_{i=1}^{K} |\phi_i(\mathbf{x}, \mathcal{S})|}.$$

Clearly, the above normalization does not change the order of sample class potentials, hence the classification decisions. The normalized margin of $f_{\mathcal{S}}$

on an observation $(\mathbf{x}, z) \in \mathbb{R}^d \times \mathbb{K}$ is then defined as

$$\hat{\gamma}(\mathbf{x}, z, \mathcal{S}) = \hat{\phi}_z(\mathbf{x}, \mathcal{S}) - \hat{\phi}_{(K-1)}(\mathbf{x}, \mathcal{S}) \ .$$

Figure 6.4(b) shows the probability densities of the margins after normalization. In both figures, the densities are shown under a log transformation. As we discussed in Section 6.5.1, if ψ is a nonnegative translation invariant function that is integrable over $\mathbb{X}$, $\phi_k(\mathbf{x}, \mathcal{S})$ is *proportional to* an estimation of the posterior probability $\Pr(z = k|\mathbf{x})$. The normalized class potential, $\hat{\phi}_k(\mathbf{x}, \mathcal{S})$, is an estimate of the posterior probability $\Pr(z = k|\mathbf{x})$. Hence the normalized margin $\hat{\gamma}$ can be viewed as an estimation on the posterior probability gap.

In classifier selection, we would like to choose a classifier whose margins concentrate towards the positive end. In terms of normalized margin, this suggests that $\hat{\gamma}$ should concentrate towards 1. We propose the following metric:

$$h(f_{\mathcal{S}}) = \text{var}_{\hat{\gamma}} - \text{mean}_{\hat{\gamma}} \qquad (6.28)$$

where $\text{mean}_{\hat{\gamma}} = \frac{1}{\ell} \sum_{i=1}^{\ell} \hat{\gamma}(\mathbf{x}_i, z_i, \mathcal{S}(i))$ and $\text{var}_{\hat{\gamma}} = \frac{1}{\ell-1} \sum_{i=1}^{\ell} [\hat{\gamma}(\mathbf{x}_i, z_i, \mathcal{S}(i)) - \text{mean}_{\hat{\gamma}}]^2$. Clearly, the desired normalized margins should have large mean and small variance, i.e., we select a classifier that minimizes h.

6.8 Experimental Results of the Model Selection Method

We present systematic evaluations of potential function classifiers and the proposed classifier selection method. The multi-class sample potential function classifier, $f_{\mathcal{S}}$, is compared with the Bayes classifier on a synthetic dataset to empirically illustrate the connection between the potential gap (6.11) and the performance of $f_{\mathcal{S}}$. We then compare the proposed model selection method using normalized margin distribution with a traditional approach using the leave-one-out training error on twenty real life data sets.

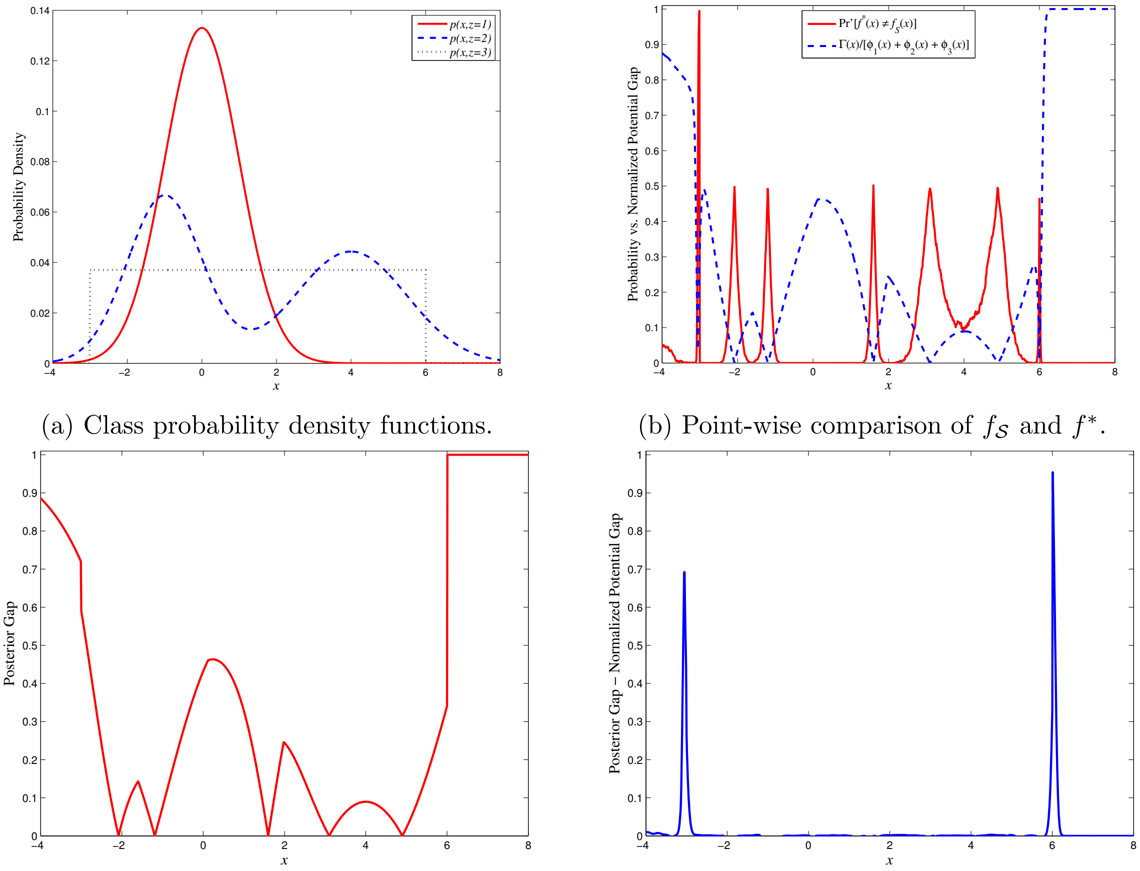
6.8.1 Synthetic Data

We consider a synthetic data set generated by the following distribution:

$$p(x,z) = \Pr(z=1)p(x|z=1) + \Pr(z=2)p(x|z=2) + \Pr(z=3)p(x|z=3)$$

where $\Pr(z=1) = \Pr(z=2) = \Pr(z=3) = \frac{1}{3}$; $p(x|z=1)$ is a normal distribution with 0 mean and unit variance; $p(x|z=2) = \frac{1}{2}F_1 + \frac{1}{2}F_2$ is a mixture of two normal distributions; F_1 has mean -1 and unit variance; F_2 has mean 4 and variance 2.25; $p(x|z=3)$ is a uniform distribution on $[-3,6]$. The class probability density functions are shown in Figure 6.5(a).

For this synthetic data, the Bayesian decisions can be evaluated from the known joint probability distribution of (x,z). Therefore we first compare the performance of a sample potential function classifier $f_{\mathcal{S}}$ with that of the



(a) Class probability density functions.

(b) Point-wise comparison of $f_{\mathcal{S}}$ and f^*.

(c) Posterior gap.

(d) The posterior gap and normalized potential gap.

Figure 6.5: Comparing the sample potential function classifier with the Bayes classifier on a synthetic data set. (a) Joint probability density functions for each category. (b) A point-wise comparison of the probability that $f_{\mathcal{S}}$ is different from f^* with the normalized potential gap. (c) The posterior gap of the synthetic data. (d) The difference between the posterior gap and the normalized potential gap.

Bayes classifier f^*. The point potential function is chosen to be a Gaussian function $\psi(x, y) = e^{-\frac{(x-y)^2}{\sigma^2}}$ with $\sigma = 0.1$. For a point-wise comparison of $f_{\mathcal{S}}$ and f^*, we select $\Pr[f_{\mathcal{S}}(x) \neq f^*(x)]$ as the metric for any fixed x. Note that this probability is defined with respect to a randomly generated training set $\mathcal{S}$ with fixed size. In this experiment, the size of a training set is chosen to be $10,000$.

Theorem 3 suggests that, under certain conditions, for a fixed sample size, the probability that $f_{\mathcal{S}}$ behaves differently from f^* depends on the potential gap (6.11). When the potential gap at x is large, it is more likely that $f_{\mathcal{S}}(x)$ makes the optimal prediction. Next, we illustrate this relationship using the above synthetic data. The computation of $\Pr[f_{\mathcal{S}}(x) \neq f^*(x)]$ is however difficult even with the knowledge of the joint probability distribution. So we estimate this probability using

$$\Pr'[f_{\mathcal{S}}(x) \neq f^*(x)] = \frac{m(x)}{n}$$

where n is the number of independently generated training sets (i.e., n experiments); $m(x)$ is the number of times that $f_{\mathcal{S}}(x)$ does not agree with $f^*(x)$ in all n experiments. Using Hoeffding's inequality, it can be derived that with probability at least $1 - \delta$ over independently generated training sets $\mathcal{S}_1, \ldots, \mathcal{S}_n$,

$$|\Pr[f_{\mathcal{S}}(x) \neq f^*(x)] - \Pr'[f_{\mathcal{S}}(x) \neq f^*(x)]| \leq \sqrt{\frac{\ln \frac{2}{\delta}}{2n}}$$

for any given x. In the experiment, $\Pr[f_{\mathcal{S}}(x) \neq f^*(x)]$ was estimated using $1,000$ independent runs, i.e., $n = 1000$. For $\delta = 0.05$, the above inequality implies that at each x, $|\Pr[f_{\mathcal{S}}(x) \neq f^*(x)] - \Pr'[f_{\mathcal{S}}(x) \neq f^*(x)]| \leq 0.0387$ with probability at least 0.95.

In general, the potential gap can be several orders of magnitude smaller than $\Pr'$, which makes it difficult to visually compare the potential gap with $\Pr'$. To overcome this difficulty, we normalize the potential gap by dividing it with the total class potential:

$$\hat{\Gamma}(\mathbf{x}) = \frac{\Gamma(\mathbf{x})}{\sum_{k=1,\ldots,K} \Phi_k(\mathbf{x})} \ ,$$

which is essentially the gap of normalized class potential. It is not difficult to show that $\hat{\Gamma}(\mathbf{x}) \in [0, 1]$. In Figure 6.5(b), $\Pr'[f_{\mathcal{S}}(x) \neq f^*(x)]$ (solid curve) is compared against the normalized potential gap (dashed curve). We observed that overall, $\Pr'[f_{\mathcal{S}}(x) \neq f^*(x)]$ is small (large) when $\hat{\Gamma}(x)$ is large (small). This is in line with the conclusion of Theorem 3.

As shown in Figure 6.5(b), $\hat{\Gamma}(x)$ has a total of 7 local minimums occurring at $x = -3$, -2.08, -1.2, 1.6, 3.1, 4.9, and 6, respectively. Local minimums of $\hat{\Gamma}(x)$ correspond to local maximums of $\Pr'$. It turns out that $\hat{\Gamma}$ is closely related to the *posterior gap*, which is defined as the difference between the largest and the second largest posterior probabilities. The posterior gap of the synthetic data is shown in Figure 6.5(c). A closer examination of Figure 6.5(b) and (c) reveals that some of the locations of the local minimums of $\hat{\Gamma}$ coincide with that of the posterior gap, i.e., at $x = -2.08$, -1.2, 1.6, 3.1, and 4.9. In addition, the posterior gap is 0 at these locations. We call the local potential minimums at these locations Type I minimums. From Figure 6.5(a), we can verify that the largest posterior probability is identical to the second largest posterior probability at Type I locations. Hence the Bayes classifier picks one of the two classes with equal probability, i.e. a random decision. This implies that $\Pr[f_{\mathcal{S}}(x) \neq f^*(x)] = 0.5$ at type I locations. Our estimates ($\Pr'[f_{\mathcal{S}}(x) \neq f^*(x)]$) reflect this fact very well.

The other two locations, i.e., $x = -2.08$ and $x = 6$, correspond to local minimums of the normalized potential gap. But they are not local minimums of the posterior gap. We call locations as such Type II minimums. Unlike a Type I minimum, the posterior gaps at a Type II minimum is significantly greater than 0. Hence the Bayesian decision is not random. Figure 6.5(d) shows the difference between the posterior gap and the normalized potential gap. It is interesting to observe that the normalized potential gap follows closely the posterior gap, except at Type II locations, where the difference is significantly larger. One may recall that a condition of Theorem 3 is a band limited joint distribution. This condition is not satisfied on this synthetic data set as illustrated by the posterior gap: the posterior gap is not continuous at the two Type II locations $x = -2.08$ and $x = 6$. From a Fourier analysis perspective, a sharp change in function values (in this example, the changes in posterior gap at Type II locations) cannot be reconstructed using only frequencies of finite values, hence not band limited. A point potential function with a given value of σ is not capable of of capturing all the high frequency information. The missing high frequency information contributes to the sharp spikes and the large values of $\Pr'[f_{\mathcal{S}}(x) \neq f^*(x)]$ at the two Type II locations.

From this synthetic data, we observed the connection between the potential gap and the performance of $f_{\mathcal{S}}$. However, the potential gap in general cannot be computed without the knowledge of the joint distribution, hence provides little information in classifier selection in practice. As discussed in Section 6.6, margin, which is analogous to the potential gap, can be evaluated from a given training set. Next, we present the results of the proposed margin based classifier selection method using real life data sets.

6.8.2 Comparison with Leave-one-out Classifier Selection

The experiments were conducted on 20 datasets, namely Balancescale, Bloodtransfusion, Breastcancer, Ecoli, Glass, Imgseg, Ionosphere, Letter, Liver, Magic, Multi-Feature1, Multi-Feature2, Multi-Feature3, Satimage, Sonar, Spectfheart, Survival, Vehicle, Vowel, Winequality, from UCI Machine Learning Repository. Each dataset is randomly divided into a training set and a test set. We built a potential function classifier with Gaussian point potential function for each dataset. The bandwidth parameter σ of the point potential function is determined from 20 different values (0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 2.0, 3.0, and 4.0), using two strategies: (1) minimizing the leave-one-out training error; (2) minimizing the margin distribution metric defined in (6.28). The above procedure was repeated for 50 runs. In each run, test errors were recorded. In Table 6.1, we list the names of the datasets, the sizes of training and test sets, the dimension of the feature space, the number of categories, and the number of runs in which the proposed model selection method outperformed (better), tied with (equal), and underperformed (worse) the leave-one-out approach. Among the 20 datasets, the proposed method outperformed the leave-one-out model selection on 15 datasets, which are highlighted in Table 6.1. The two approaches tied on 1 dataset. This suggests a very competitive performance of the proposed method.

6.8.3 Conclusions

The contributions of PFRs are given as follows:

- *Connections of PFRs with the Bayes decision theory.* Given charge density functions a priori, we present conditions under which a PFR is essentially optimal under the framework of the Bayesian decision the-

Table 6.1: The comparison results of model selection using leave-one-out error and the margin distribution metric defined in (6.28).

Dataset	**Size of training set**	**Size of test set**	**Feature dimension**	**Number of classes**	**Better**	**Equal**	**Worse**
Balancescale	**570**	**55**	**4**	**2**	**7**	**37**	**6**
Bloodtransfusion	**600**	**148**	**4**	**2**	**29**	**13**	**8**
Breastcancer	**600**	**83**	**9**	**2**	**10**	**39**	**1**
Ecoli	**200**	**136**	**7**	**8**	**15**	**22**	**13**
Glass	**150**	**64**	**9**	**6**	**22**	**21**	**7**
Imageseg	2100	210	19	7	15	22	13
Ionosphere	**320**	**31**	**34**	**2**	**5**	**29**	**16**
Letter	**18000**	**2000**	**16**	**26**	**32**	**7**	**11**
Liver	**300**	**45**	**6**	**2**	**23**	**14**	**13**
Magic	**10000**	**9020**	**10**	**2**	**44**	**4**	**2**
Multi-Feature1	**1800**	**200**	**216**	**10**	**9**	**34**	**7**
Multi-Feature2	**1800**	**200**	**64**	**10**	**7**	**43**	**0**
Multi-Feature3	**1800**	**200**	**240**	**10**	**9**	**39**	**2**
Satimage	**5835**	**600**	**36**	**6**	**16**	**23**	**11**
Sonar	150	58	60	2	15	20	15
Spectfheart	**200**	**67**	**44**	**2**	**3**	**37**	**0**
Survival	**206**	**100**	**3**	**2**	**22**	**8**	**20**
Vehicle	800	46	18	4	8	31	11
Vowel	890	100	10	11	3	33	14
Winequality	6000	497	11	7	17	0	33

ory. We then look into a more practical scenario where a PFR is built from a given set of training observations with unknown but fixed charge density functions. We show that a PFR is, in this case, equivalent to a plug-in decision rule using kernel density estimation, hence universally consistent.

- *A new generalization bound for PFRs.* We discuss the classifier selection for PFRs using complexity regularization. An upper bound on the generalization performance for PFRs are derived using a margin distribution.

- *A simple classifier selection method for PFRs.* Motivated by the above generalization bound, we propose a simple kernel selection method using a normalized margin distribution. Extensive experimental results on artificial data and real applications demonstrate the competitive performance of the proposed framework.

6.9 Experimental Results over SSVEP Data

Because the bit rate calculated by Eq.(1.1) takes into account the accuracy, the number of possible selections and the time to make a decision, it should be an convincing parameter to evaluate the proposed BCI.

We used offline SSVEP data to test the PFRs classifier. The dataset contains ten classes, that are SSVEP responses to ten different frequencies (10Hz to 19Hz) delivered by HSL space stimuli described in Section 5. Each class has 20 samples. Each sample are 5 seconds of SSVEP. In our experiments, the size of the training set varies, while the size of the test set is always five. The performance of the PFRs classifier is shown in Table 6.2[2]. In this table, *"Training Samples"* is the number of samples used as the training data.

Another experiment was conducted to show the good scalability of PFR as shown in Table 6.3. The"good scalability" is defined as PFRs do not need to be re-trained when a new class is added (conversely to SVM), thus take

[2] The time-domain SSVEP data is pre-processed into frequency domain by FFT

Table 6.2: Statistic of PFRs over Offline SSVEP Data

Training Samples	Computation Time	Average Accuracy of 50 runs
5	2.45 seconds	99.2%
6	3.18 seconds	99.4%
7	3.71 seconds	99.45%
8	4.16 seconds	99.78%
9	4.66 seconds	99.49%
10	5.14 seconds	99.70%
11	5.65 seconds	99.84%
12	5.58 seconds	99.75%
13	5.68 seconds	99.80%
14	5.62 seconds	99.87%
15	5.79 seconds	99.84%

less time to "reboot" in a realtime BCI system. In this experiment, training samples are always set to ten, while the number of classes is 5 at the first, then is progressively added to 10.

Table 6.3: Statistic of PFRs over Offline SSVEP Data 2

Number of Classes	Computation Time	Average Accuracy of 50 runs
5	2.53 seconds	99.68%
6	3.26 seconds	99.70%
7	3.59 seconds	99.65%
8	4.07 seconds	99.73%
9	4.65 seconds	99.72%
10	5.17 seconds	99.70%

Conclusively, if 10 (classes) is the number of possible selections, 15 is the number of training samples, the bits per decision will be:

$B = \log_2 10 + P \log_2 0.99 + (1 - 0.99) \log_2 \frac{1-0.99}{10-1} = 3.2bits.$

Under this scenario, as FFT takes 0.3 seconds and PFRs take 5.8 seconds, each decision takes 6.1 seconds. Thus the bits rate per minute is $B * \frac{60}{5+6.1} =$

$17.3 bits/min$. Similarly, we calculate the bits rates of PFRs over different numbers of classes and list the results in Table 6.4. The best performance 18.51 bits/min was achieved with 8 classes.

Table 6.4: Statistic of PFRs over Offline SSVEP Data 3

Number of Classes	Time per Decision	Bits Rate
5	7.83 seconds	16.85 bits/min
6	8.56 seconds	17.38 bits/min
7	8.89 seconds	18.22 bits/min
8	9.37 seconds	18.51 bits/min
9	9.95 seconds	18.45 bits/min
10	10.47 seconds	18.39 bits/min

Chapter 7
FUTURE WORK

So far, from the perspective of the stimulation, we have focused on finding an effective stimulus, dual stimuli that increase the number of possible selections, and a visually friendly stimulus. This suggests a direction for future work: an HSL space stimulus shall be designed with two frequencies delivered by 50% duty cycle square waves along its H and S axis.

Another improvement may come from the machine learning technique. The good scalability of PFRs can help a user find her/his optimality based on the assumption that people react differently to different frequency stimuli[1]. The optimality can be interpreted as: when presented with some randomly selected stimuli, a user expects the set of stimuli provided to be pure – containing only frequencies he responds well – and complete – containing all frequencies he responds well, thus enhances his accuracy and speed.

Purity and completeness are analogous, respectively, to the criteria of precision and recall from information retrieval (IR). IR systems are often

[1] This is true over experiments in this dissertation. For example, in the HSL space stimuli experiment, subject1's response to 6Hz is stronger than subject2's.

evaluated in terms of their precision and recall with respect to a labeled data collection; human judges decide which objects match a particular query, and the system is rated on how closely its results accord with the human judges. This suggests hand labeling of good stimulation frequencies – an user tries all stimuli, for each stimulus, we would have to judge if the user responds well. Instead of fulfilling both purity and completeness, our next goal is to build an SSVEP BCI that accepts user's feedback of removing a certain stimulus, to make his stimuli pure, since it is obvious that purity is much easier to realize.

There are two ways to make the notion of a "bad" stimulus. First, if the user notices PFRs usually misclassify a stimulus, then removing it is straightforward. Alternatively, if the PFRs realize a set of training data is too close to an existing good training set, then we know it is not distinguishable and should be removed. This scenario could be done by viewing the existing good training set as positive, and new training data as negative, and apply a noise-tolerant symbolic learning technique to output a decision.

Finally, our work has focused on separated applications for convenience. We plan to test our approach on a real SSVEP BCI system in the future.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Allison B, Sugiarto I, Graimann, B and Graser A "A Display Optimization in SSVEP BCIs," '*Computer-Human Interaction* 2008

[2] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer, "Theoretical Fundations of the Potential Function Method in Pattern Recognition Learning," *Automation and Remote Control*, vol. 25, no. 6, pp. 917–936, 1964.

[3] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer, "The Probability Problem of Pattern Recognition Learning and The Method of Potential Functions," *Automation and Remote Control*, vol. 25, no. 9, pp. 1307–1323, 1964.

[4] C. W. Anderson and Z. Sijerycic, "Classification of EEG signals from four subjects during five mental tasks," *Solving Engineering Problems with Neural Networks: Proceedings of the Conference on Engineering Applications in Neural Networks*, pp. 407-414, 1996.

[5] M. Anthony and N. Biggs, *Computational Learning Theory*, Cambridge University Press, 1992.

[6] Kenji Arakawa, Shozo Tobimatsu, Hiroyuki Tomoda, Jun-ichi Kira, Motohiro Kato, *The Effect of Spatial Frequency on Chromatic and Achromatic Steady-state Visual Evoked Potentials*, newblock *Clinical Neurophysiology*, vol. 110, Issue 11, pp. 1959-1964, 1999.

[7] H. Avi-Itzhak and T. Diep, "Arbitrarily Tight Upper and Lower Bounds on the Bayesian Probability of Error," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 1, pp. 89–91, 1996.

[8] F. Babiloni, F. Cincotti, L. Lazzarini, and M. G. Marciani, "Linear classification of low-resolution EEG patterns produced by imagined hand movements," *IEEE Trans. Rehab. Eng.*, vol. 8, no. 2, pp. 186–188, 2000.

[9] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, M. I. Jordan, "Matching Words and Pictures," *Journal of Machine Learning Research*, vol. 3, pp. 1107–1135, 2003.

[10] A. Barron, "Complexity Regularization with Application to Artificial Neural Networks," *Nonparametric Functional Estimation and Related Topics*, pp. 561–576, Kluwer Academic Publisher, 1991.

[11] P. L. Bartlett, "For Valid Generalization, the Size of the Weights is More Important Than the Size of the Network," *Advances in Neural Information Processing Systems 9*, pp. 134–140, 1997.

[12] O. A. Bashkirov, E. M. Braverman, and I. B. Muchnik, "Potential Function Algorithms for Pattern Recognition Learning Machines," *Automation and Remote Control*, vol. 25, no. 5, pp. 692–695, 1964.

[13] T. Bayes, "An Essay Towards Solving a Problem in the Doctrine of Chances," *The Philosophical Transactions*, vol. 53, pp. 370–418, 1763.

[14] M. Ben-Bassat, K. L. Klove, and M. H. Weil, "Sensitivity Analysis in Bayesian Classification Models: Multiplicative Deviations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2, no. 3, pp. 261–266, 1980.

[15] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, second edition, Springer, 1985.

[16] Beverina F, Palmas G, Silvoni S, Piccione F and Giove S, "User adaptive BCIs: SSVEP and P300 based interfaces," *PsychNology Journal*, vol. 1, pp. 331-354, 2003.

[17] Bin G, Gao X, Yan Z, Hong B and Gao S, "An online multi-channel SSVEP-based brain-computer interface using a canonical correlation analysis method" J. Neural Eng., vol. 6, no. 4, p. 046002 (6pp), 2000.

[18] Birbaumer N, Kubler A, Ghanayim N, Hinterberger T, Perelmouter J, Kaiser J, Iversen I, Kotchoubey B,Neumann N and Flor H, “The thought translation device (TTD) for completely paralyzed patients” *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 2, pp. 190-193, 2000.

[19] Birbaumer N, Ghanayim N, Hinterberger T, Iversen I, Kotchoubey B, Kabler A, Perelmouter J, Taub E and Flor H, “ A spelling device for the paralyzed” *Nature*, vol. 398, pp.297-298,1999.

[20] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[21] B. Blankertz, B. Curio and G. Muller, “Classifying Single Trial EEG: Towards Brain Computer Interfacing,” *Advances in Neural Information Processing Systems*, vol. 1, pp. 157–164, 2002.

[22] B. Blankertz, G. Dornhege, C. Schafer, R. Krepki, J. Kohlmorgen, K.-R. Muller, V. Kunzmann, F. Losch and G. Curio, “Boosting bit rates and error detection for the classification of fast-paced motor commands based on single-trial EEG analysis,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, pp. 127–131, 2003.

[23] Blankertz, B. Dornhege, G. Krauledat, M. Muller, K.-R. Kunzmann, V. Losch and F. Curio, G. “The Berlin Brain-Computer Interface: EEG-Based Communication Without Subject Training,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 14, no.2, pp. 147–152, 2006.

[24] Blankertz, B., Muller, K.R., Krusienski, D.J., Schalk, G., Wolpaw, J.R., Schlogl, A., Pfurtscheller, G., Millan, Jd.R., Schroder, M. and Birbaumer, N., “The BCI competition III: validating alternative approaches to actual BCI problems,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 14, no.2, pp. 153–159, 2006.

[25] Benjamin Blankertz, Guido Dornhege, Matthias Krauledat, Klaus-Robert Muller and Gabriel Curio, “The non-invasive Berlin Brain-

Computer Interface: Fast acquisition of effective performance in untrained subjects," *Neuro Image*, vol. 37, no.2, pp. 539–550, 2007.

[26] M. Boulle, "Compression-Based Averaging of Selective Naive Bayes Classifiers," *Journal of Machine Learning Research*, vol. 8, pp. 1659–1685, 2007.

[27] E. M. Braverman, "On the Method of Potential Functions," *Automation and Remote Control*, vol. 26, no. 12, pp. 2205–2213, 1965.

[28] E. M. Braverman and E. S. Pyatnitskii, "Estimation of the Rate of Convergence of Algorithms Based on the Potential Functions Method," *Automation and Remote Control*, vol. 27, no. 1, pp. 95–112, 1966.

[29] Cecotti, H., Volosyak, I. and Graser, A., "Evaluation of an SSVEP based Brain-Computer Interface on the command and application levels," *NER '09*, pp. 474–477, 2009.

[30] Y. Chen and J. Z. Wang, "Support Vector Learning for Fuzzy Rule-Based Classification Systems," *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 6, pp. 716–728, 2003.

[31] Y. Chen, J. Bi, and J. Z. Wang, "MILES: Multiple-Instance Learning via Embedded Instance Selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 1931–1947, 2006.

[32] Y. Chen, H. L. Bart, X. Dang, and H. Peng, "Depth-Based Novelty Detection and its Application to Taxonomic Research," *Proc. of The Seventh IEEE International Conference on Data Mining*, pp. 113–122, 2007.

[33] Y. Chen, E. K. Garcia, M. R. Gupta, A. Rahimi, and L. Cazzanti, "Similarity-based Classification: Concepts and Algorithms," *Journal of Machine Learning Research*, vol. 10, pp. 747–776, 2009.

[34] Cheng M, Gao X and Gao S, " Design and implementation of a brain-computer interface with high transfer rates" *IEEE Trans. Biomed. Eng.* vol. 49, pp. 1181-1186, 2002.

[35] M. Cheng, X. Gao, S. Gao and D. Xu, "Multiple color stimulus induced steady state visual evoked potentials" *Proceedings of the 23rd Annual IEEE International Conference of Engineering in Medicine and Biology Society*, pp.1012-1014, 2001.

[36] Chiappa H K "Evoked Potentials in Clinical Medicine" New York: Revan Press, 1983

[37] F. Cincotti, D. Mattia, C. Babiloni, F. Carducci, S. Salinari, L. Bianchi, M. G. Marciani, and F. Babiloni, "The use of EEG modifications due to motor imagery for brain-computer interfaces," *IEEE Trans. Neural. Syst. Rehab. Eng.*, vol. 11, no. 2, pp. 131–133, 2003.

[38] P. A. Devijver, "On a New Class of Bounds on Bayes Risk in Multi-Hypothesis Pattern Recognition," *IEEE Transactions on Computers*, vol. 23, no. 1, pp. 70–80, 1974.

[39] L. Devroye, "On the Asymptotic Probability of Error in Nonparametric Discrimination," *The Annals of Statistics*, vol. 9, no. 6, pp. 1320–1327, 1981.

[40] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag New York, 1996.

[41] P. Domingos and M. J. Pazzani, "On the Optimality of the Simple Bayesian Classifier Under Zero-one Loss," *Machine Learning*, vol. 29, no. 2-3, pp. 103–130, 1997.

[42] E. Donchin, K. M. Spencer, and R.Wijesinghe, "The mental prosthesis: Assessing the speed of a P300-based brain-computer interface," *IEEE Trans. Rehab. Eng.*, vol. 8, no. 2, pp. 174–179, 2000.

[43] G. Dornhege, B. Blankertz, G. Curio, and K.Muller, "Boosting bit rates in non-invasive EEG single-trial classifications by feature combination and multi-class paradigms," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 993-1002, 2004.

[44] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, Second Edition John Wiley & Sons, Inc., 2001.

[45] M. Fatourechi, A. Bashashati, R.Ward, and G. Birch, "A hybrid genetic algorithm approach for improving the performance of the LF-ASD brain computer interface," *Proc. Int. Conf. Acoust. Speech Signal Process.*, vol. 5, pp. 345-348, 2004.

[46] R. S. Fisher, G. Harding, G. Erba, G.L. Barkley, and A. Wilkins. "Photic and Pattern - induced Seizures: A Review for the Epilepsy Foundation of America Working Group," *Epilepsia* vol. 46(9), pp. 1426-1441, 2005.

[47] Friman O, Volosyak I and Graser A, "Multiple channel detection of steady-state visual evoked potentials for brain-computer interfaces," *IEEE Trans. Biomed. Eng.* vol. 54, pp. 742-750, 2007.

[48] G. Gage, K. Ludwig, K. Otto, E. Ionides, and D. Kipke, "Naive co-adaptive cortical control," *J. Neural Eng.* vol. 2, pp. 52-63, 2005.

[49] X. Gao, D. Xu, M. Cheng, and S. Gao, "A BCI-based environmental controller for the motion-disabled," *IEEE Transactions on Neural Systems and Rehabilitation Engineering* vol. 11, no. 2, pp. 137140, 2003.

[50] Garcia G, Ibanez D, Mihajlovic V and Chestakov D "Detection of High Frequency Steady State Visual Evoked Potentials for Brain-Computer Interfaces," *17th European Signal Processing Conference*, 2009.

[51] Garcia G, "Detection of High-Frequency Steady State Visual Evoked Potentials Using Phase Rectified Reconstruction," *16th European Signal Processing Conference*, 2008.

[52] A. Garg and D. Roth, "Margin Distribution and Learning Algorithms," *Proc. of Twentieth International Conf. on Machine Learning*, pp. 210–217, 2003.

[53] L. Gordon and R. A. Olshen, "Asymptotically Efficient Solutions to the Classification Problem," *The Annals of Statistics*, vol. 6, no. 3, pp. 515–533, 1978.

[54] D. J. Griffiths, *Introduction to Electrodynamics*, Third Edition, Prentice Hall, 1998.

[55] Y. Guermeur, "VC Theory of Large Margin Multi-Category Classifiers," *Journal of Machine Learning Research*, vol. 8, pp. 2551–2594, 2007.

[56] A. Halevy, P. Norvig, and F. Pereira, "The Unreasonable Effectiveness of Data," *IEEE Intellegent Systems*, vol. 24, no. 2, pp. 8–12, 2009.

[57] W. A. Hashlamoun, P. K. Varshney, and V. N. S. Samarasooriya, "A Tight Upper Bound on the Bayesian Probability of Error," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 2, pp. 220–224, 1994.

[58] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2001.

[59] Herrmann C "Human EEG responses to 1-100 Hz flicker: resonance phenomena in visual cortex and their potential correlation to cognitive phenomena" *Exp. Brain Res.*, vol. 137, no. 3-4, pp. 346-353, 2001.

[60] Hinterberger T, Kubler A, Kaiser J, Neumann N and Birbaumer N, "A brain-computer-interface (BCI) for the locked-in: comparison of different EEG classifications for the thought translation device" *Clin Neurophysiol.*, vol 114, pp.416-425, 2003.

[61] T. Hofmann, J. Puzicha, and M. I. Jordan, "Unsupervised Learning from Dyadic Data," *Advances in Neural Information Processing Systems 11*, pp. 466-472, 1999.

[62] Jaganathan V, Mukesh T and Reddy M "Design and implementation of high performance visual stimulator for brain computer interfaces," *Proceedings of the 27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 5381-5383, 2005.

[63] Jia, Chuan and Xu, Honglai and Hong, Bo and Gao, Xiaorong and Zhang, Zhiguang and Gao, Shangkai, "A Human Computer Interface

Using SSVEP-Based BCI Technology," *Proceedings of the 27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 4565, pp. 113-119, 2007.

[64] Kaper M, Meinicke P, Grossekathoefer U, Lingner T, Ritter H., "BCI Competition 2003–Data set IIb: support vector machines for the P300 speller paradigm," *IEEE Trans Biomed Eng.*, vol. 51, no. 6, pp. 1073-1076, 2004.

[65] B. Kamousi, Z. Liu, and B. He, "Classification of motor imagery tasks for brain-computer interface applications by means of two equivalent dipoles analysis" *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 13, no. 2, pp. 166-171, 2005.

[66] M. J. Kearns and U. V. Vazirani, *An Introduction to Computational Learning Theory*, The MIT Press, 1994.

[67] Kelly S, Labor E, Finucane C, McDarby G and Reilly R, "Visual spatial attention control in an independent brain-computer interface" *IEEE Trans. Biomed. Eng.*, vol. 52, pp. 1588-1596, 2005.

[68] Kelly S, Lalor E, Reilly R and Foxe J, "Visual spatial attention tracking using high-density SSVEP data for independent brain-computer communication" *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 13, no. 2, pp. 172-178, 2005.

[69] H.-C. Kim and Z. Ghahramani, "Bayesian Gaussian Process Classification with the EM-EP Algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 1948–1959, 2006.

[70] M. Kirby and C. Anderson, "Geometric analysis for the characterization of nonstationary time-series," *in Springer Applied Mathematical Sciences Series Celebratory Volume for the Occasion of the 70th Birthday of, L. Sirovich, E. Kaplan, J. Marsden, and K. R. K. Sreenivasan,*, Eds. New York: Springer-Verlag, ch. 8, pp. 263-292, 2003.

[71] Krusienski D and Allison B, "Harmonic coupling of steady-state visual evoked potentials" *EMBS 2008, 30th Annual International Conference of the IEEE*, pp. 5037-5040, 2008.

[72] Lalor E, Kelly S, Finucane C, Burke R, Smith R, Reilly R and McDarby G, "Steady-state VEP-based brain-computer interface control in an immersive 3D gaming environment" *EURASIP J. Appl. Signal Process.*, vol. 2005, no. 19, pp. 3156-3164, 2005.

[73] E. Lalor, S. P. Kelly, C. Finucane, R. Burke, R. B. Reilly, and G. McDarby, "Brain computer interface based on the steady-state VEP for immersive gaming control" *Biomed. Tech.*, vol. 49, no. 1, pp. 63-64, 2004.

[74] J. Langford and J. Shawe-Taylor, "PAC-Bayes and Margins," Advances in Neural Information Processing Systems 15, pp. 439–446, 2002.

[75] P. Langley, W. Iba, and K. Thompson, "An Analysis of Bayesian Classifiers," *Proc. of the Tenth National Conf. on Artificial Intelligence*, pp. 223-228, 1992.

[76] Lee H, Cichocki A and Choi S, "Nonnegative matrix factorization for motor imagery EEG classification" *Proceedings of the 16th International Conference on Artificial Neural Networks*, vol. 4132 of Lecture Notes in Computer Science, pp. 250-259, 2006.

[77] S. Lemm, B. Blankertz, G. Curio, and K.-R. Muller, "Spatio-spectral filters for improved classification of single trial EEG," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 9, pp. 1541-1548, 2005.

[78] Y. Li, X. Gao, and S. Gao, "Classification of single-trial electroencephalogram during finger movement," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1019-1025, 2004.

[79] Lou B, Hong B, Gao X and Gao S, "Bipolar electrode selection for a motor imagery based brain-computer interface" *J. Neural Eng.*, vol. 5, pp. 342-349, 2008.

[80] G. Lugosi and K. Zeger, “Concept Learning Using Complexity Regularization,” *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 48–54, 1996.

[81] P. Martinez, H. Bakardjian, and A. Cichocki, “Fully online multicommand brain-computer interface with visual neurofeedback using SSVEP paradigm,” *Computational Intelligence and Neuroscience*, vol. 2007, Article ID 94561, 9 pages, 2007.

[82] A. Maurer, “Learning Similarity with Operator-valued Large-margin Classifiers,” *Journal of Machine Learning Research*, vol. 9, pp. 1049–1082, 2008.

[83] D. J. McFarland and J. R.Wolpaw, “Sensorimotor rhythm-based brain-computer interface (BCI): Feature selection by regression improves performance,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 13, no. 3, pp. 372-379, 2005.

[84] Middendorf M, McMillan G,Calhoun G and Jones K, “Brian-computer interfaces based on the steady-state visual-evoked response,” *IEEE Trans. Rehabil. Eng.*, vol. 8, pp. 211-214, 2000.

[85] J. del R. Millan, F. Renkens, J. Mourino, and W. Gerstner, “Brain actuated interaction,” *Artif. Intel.*, vol. 159, no.1-2, pp. 241-259, 2004.

[86] J. del R. Millan and J. Mourino., “Asynchronous BCI and local neural classifiers: An overview of the adaptive brain interface project,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, no. 2, pp. 159-161, 2003.

[87] T. M. Mitchell, *Machine Learning*, McGraw-Hill Companies, Inc., 1997.

[88] Morgan S, Hansen J and Hillyard S, “Selective attention to stimulus location modulates the steady-state visual evoked potential” *Neurobiology*, vol. 93, pp. 4770-4774, 1996.

[89] Mukesh T, Jaganathan V and Reddy M, “A novel multiple frequency stimulation method for steady state VEP based brain computer interfaces” *Physiol. Meas.*, vol. 27, no. 1, pp. 61-71, 2006.

[90] Muller M and Hillyard S, “Effects of spatial selective attention on the steady- state visual evoked potential in the 20-28 hz range” *Cognitive Brain Research*, vol. 6, pp. 249-261, 1997.

[91] M. M. Muller, P. Malinowski, T. Gruber, and S. A. Hillyard, “Sustained division of the attentional spotlight” *Nature*, vol. 424, no. 6946, pp. 309312, 2003.

[92] Muller-Putz G, Scherer R, Brauneis C and Pfurtscheller G, “Steady-state visual evoked potential (SSVEP)-based communication: impact of harmonic frequency components” *Neural Eng.*, vol. 2, pp. 123-130, 2005.

[93] C. Neuper,A. Schlogl, and G. Pfurtscheller, “Enhancement of left-right sensorimotor EEG differences during feedback-regulated motor imagery,” *J. Clin. Neurophysiol.*, vol. 16, no. 4, pp. 373-382, 1999.

[94] Nielsen K, Cabrera A and Nascimento O, “EEG based BCI - towards a better control. Brain-computer interface research at Aalborg university” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 14, no. 2, pp. 202-204, 2006.

[95] Nijholt, A. Tan, D., “Brain-Computer Interfacing for Intelligent Systems,” *Intelligent Systems, IEEE*, vol. 23, no. 3, pp. 72-79, 2008.

[96] OpenEEG Open source project *http://openeeg.sourceforge.net/doc/*.

[97] Sergio Parini, Luca Maggi, Anna C. Turconi, and Giuseppe Andreoni, “A Robust and Self-Paced BCI System Based on a Four Class SSVEP Paradigm: Algorithms and Protocols for a High-Transfer-Rate Direct Brain Communication,” *Computational Intelligence and Neuroscience*, 2009.

[98] Pastor M, Artieda J, Arbizu J, Valencia M and Masdeu J, "Human cerebral activation during steady-state visual-evoked responses" *J. Neurosci.*, vol. 23, no. 37, pp. 11621-11627, 2003.

[99] E. Pekalska, P. Paclik and R. P.W. Duin, "A Generalized Kernel Approach to Dissimilarity-based Classification," *Journal of Machine Learning Research*, vol. 2, pp. 175–211, 2001.

[100] Piccini L, Parini S, Maggi L and Andreoni G, "A wearable home BCI system: preliminary results with SSVEP protocol" *Proceedings of the 27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS 2005)*, pp. 5384-5387, 2005.

[101] Pfurtscheller G, Neuper C, Guger C, Harkam W, Ramoser H, Schlagl A, Obermaier B and Pregenzer M, "Current trends in Graz brain-computer interface (BCI) research" *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 2, pp. 216-219, 2000.

[102] G. Pfurtscheller, C. Neuper, D. Flotzinger, and M. Pregenzer, "EEG-based discrimination between imagination of right and left hand movement," *Electroencephalogr. Clin. Neurophysiol.*, vol. 103, no. 2, pp. 642-651, 1997.

[103] Pfurtscheller, G., Neuper, C., "Motor imagery and direct brain-computer communication," *Proceedings of the IEEE*, vol. 89, no. 7, pp. 1123-1134, 2002.

[104] G. Pfurtscheller, C. Neuper, G. R. Muller, B. Obermaier, G. Krausz, A. Schlogl, R. Scherer, B. Graimann, C. Keinrath, D. Skliris, M. Wortz, G. Supp, and C. Schrank, "Graz-BCI: State of the art and clinical applications," *IEEE Trans. Neural. Syst. Rehab. Eng.*, vol. 11, no. 2, pp. 177-180, 2003.

[105] G. Rätsch and M. K. Warmuth, "Efficient Margin Maximizing with Boosting," *Journal of Machine Learning Research*, vol. 6, pp. 2131–2152, 2005.

[106] Regan D, “Human Brain Electrophysiology: Evoked Potentials and Evoked Magnetic Fields in Science and Medicine” New York: Elsevier, 1989

[107] Regan D, 1966, “An effect of stimulus colour on average steady-state potentials evoked in man” *Nature*, vol. 210, no. 5040, pp. 10561057, 1966.

[108] Regan D, “An effect of stimulus colour on average steady-state potentials evoked in man,” *Science*, vol. 210, pp. 1056–1057, 1966.

[109] Christopher Reichley, Aik Min Choong , Fei Teng, Dwight Waddell, Pamela Lawhead, Scott Gustafson and Yixin Chen, “EEG processing and robotic bio-feedback interfaces ,” *BCI Meeting 2010*, 2010.

[110] S. Rosset, J. Zhu, and T. Hastie, “Boosting as a Regularized Path to a Maximum Margin Classifier,” *Journal of Machine Learning Research*, vol. 5, pp. 941–973, 2004.

[111] T. Rosipal, L. Trejo, and B. Matthews, “Kernel PLS-SVC for linear and nonlinear classification,” *Proc. 20th Int. Conf. Machine Learning*, pp. 640–647, 2003.

[112] F. di Russo, W. A. Teder-Salejarvi, and S. A. Hillyard, “The Cognitive Electrophysiology of Mind and Brain,” *Elsevier*, Amsterdam, The Netherlands, 2002.

[113] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, “Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods,” *The Annals of Statistics*, vol. 26, no. 5, pp. 1651–1686, 1998.

[114] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, “Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods,” *The Annals of Statistics*, vol. 26, no. 5, pp. 1651–1686, 1998.

[115] A. Schlogl, K. Lugger, and G. Pfurtscheller, “Using adaptive autoregressive parameters for a brain-computer-interface experiment,” *Proc.*

19th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., vol. 19, pp. 1533-1535, 1997.

[116] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, The MIT Press, 2002.

[117] Hilit Serby, Elad Yom-Tov, and Gideon F. Inbar, "An Improved P300-Based Brain-Computer Interface," *Clin. Neurophysiol.*, vol. 117(3), pp. 538-548, 2006.

[118] Sellers E and Donchin E, "A P300-based brain-computer interface: Initial tests by ALS patients," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 13, no. 1, pp. 89-98, 2005.

[119] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.

[120] R. Stein, D.Weber, Y. Aoyagi, A. Prochazka, J.Wagenaar, S. Shoham, and R. Normann, "Coding of position by simultaneously recorded sensory neurons in the cat dorsal root ganglion," *J. Physiol.*, vol. 560, pp. 883-896, 2004.

[121] C. J. Stone, "Consistent Nonparametric Regression," *The Annals of Statistics*, vol. 5, no. 4, pp. 595–620, 1977.

[122] J. Sung, Z. Ghahramani, and S.-Y. Bang, "Latent-Space Variational Bayes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2236–2242, 2008.

[123] Indar Sugiarto, Brendan Allison, Axel Graser "Optimization Strategy for SSVEP-Based BCI in Spelling Program Application," *International Conference on Computer Engineering and Technology*, vol. 1, pp. 223-226, 2009.

[124] David Sutoyo, Ramesh Srinivasan "Nonlinear SSVEP responses are sensitive to the perceptual binding of visual hemifields during conven-

tional ‘eye’ rivalry and interocular ‘percept’ rivalry,” *Brain Research*, vol. 1251, pp. 245-255, 2009.

[125] Sutter E, “The brain response interface: communication through visually-induced electrical brain response” *Microcomput. Appl.*, vol. 15, pp. 31-45, 1992.

[126] Teng F, Choong A, Gustafson S, Waddell D, Lawhead P and Chen Y “Steady State Visual Evoked Potentials by Dual Sine Waves,” *Proc. of the ACM Southeast Conference (ACMSE)*, 2010.

[127] Townsend, G., LaPallo, B.K., Boulay, C., Krusienski, D.J., Frye, G.E., Hauser, C.K., Schwartz, N.E., Vaughan, T.M., Wolpaw, J.R., Sellers, E.W. “A novel P300-based brain-computer interface stimulus presentation paradigm: moving beyond rows and columns,” *Clinical Neurophysiology*, vol. 121, pp. 1109-1120, 2010.

[128] R. Tibshirani and T. Hastie, “Margin Trees for High-dimensional Classification,” *Journal of Machine Learning Research*, vol. 8, pp. 637–652, 2007.

[129] V. N. Vapnik and A. Ya. Chervonenkis, “On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities,” *Theory of Probabilities and Its Applications*, vol. 16, no. 2, pp. 264–280, 1971.

[130] V. N. Vapnik, *Estimation of Dependencies Based on Empirical Data*, Springer-Verlag, 1982.

[131] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, Inc., 1998.

[132] P. M. Vasquez, H. Bakardjian, M. Vallverdu, and A. Cichocki, “Fast multi-command SSVEP brain machine interface without training,” *Proceedings of the 18th International Conference on Artificial Neural Networks*, vol. 113, no. 6, pp. 300-307, 2008.

[133] S. Veeramachaneni and G. Nagy, "Analytical Results on Style-Constrained Bayesian Classification of Pattern Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 29, no. 7, pp. 1280-1285, 2007.

[134] J. J. Vidal, "Toward direct brain-computer communication," *Annu. Rev. Biophys,* vol. 2, pp. 157-180, 1973.

[135] J. J. Vidal, "Real-time detection of brain events in EEG," *IEEE Proc,* vol. 65, pp. 633-664, 1977.

[136] Vidaurre, C., Schlogl, A., Cabeza, R., Scherer, R. and Pfurtscheller, G., "A fully on-line adaptive BCI," *IEEE Transactions on Biomedical Engineering,* vol. 53, no. 6, pp. 1214-1219, 2006.

[137] Yijun Wang, Zhiguang Zhang, Xiaorong Gao and Shangkai Gao, "Lead selection for SSVEP-based brain-computer interface," *IEMBS '04*, pp. 4507–4510, 2004.

[138] Yijun Wang, Ruiping Wang, Xiaorong Gao, Bo Hong and Shangkai Gao, "A practical VEP-based brain-computer interface," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 14, no. 2, pp. 234–240, 2006.

[139] J. Wang and X. Shen, "Large Margin Semi-supervised Learning," *Journal of Machine Learning Research*, vol. 8, pp. 1867–1891, 2007.

[140] K. Q. Weinberger and L. K. Saul, "Distance Metric Learning for Large Margin Nearest Neighbor Classification," *Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[141] J. R. Wolpaw, D. J. McFarland, G. W. Neat, and C. A. Forneris, "An EEG-based brain-computer interface for cursor control," *Electroencephalogr. Clin. Neurophysiol.*, vol. 78, pp. 252-259, 1991.

[142] J. R. Wolpaw, N. Birbaumer, W. J. Heetderks, D. J. McFarland, P. H., Peckham, G. Schalk, E. Donchin, L. A. Quatrano, C. J. Robinson, and

T. M. Vaughan, “Brain-computer interface technology: A review of the first international meeting,” *IEEE Trans. Rehab. Eng.*, vol. 8, no. 2, pp. 164-173, 2000.

[143] Wolpaw J, Birbaumer N, McFarland D, Pfurtscheller G and Vaughan T, “Brain-computer interfaces for communication and control” *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767-791, 2002.

[144] J. R. Wolpaw and D. J. McFarland, “Control of a two-dimensional movement signal by a non-invasive brain-computer interface in humans,” *Proc. Nat. Acad. Sci.*, vol. 101, pp. 17849-17854, 2004.

[145] Wu Z, Lai Y, Xia Y, Wu D and Yao D, “Stimulator selection in SSVEP-based BCI” *Med. Eng. Phys.*, vol. 30, no. 8, pp. 1079-88, 2008.

[146] Jake Young, “Ocular Dominance Columns in Humans and the Limits of fMRI” $http : //scienceblogs.com/purepedantry/2007/10/ocular_{d}ominance_{c}olumns_{a}nd_{t}.php.$

VITA

Fei Teng received B.S. degree in computer science from Beijing University of Posts and Telecommunications and M.S. degree in computer science from University of New Orleans. His research interests include pattern recognition and brian-computer interface.