CAPTURING SPIKE TRAIN SIMILARITY STRUCTURE:
A POINT PROCESS DIVERGENCE APPROACH

By
IL “MEMMING” PARK

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
UNIVERSITY OF FLORIDA
2010
ACKNOWLEDGMENTS

It would be only appropriate to thank my advisor Dr. José Carlos Santos Carvalho Príncipe first for his guidance and lessons not only for research but for life in general. A lot of people helped me get through my journey of graduate school, and perhaps my attempt to properly thank them all will fail miserably, but I have to try. Dr. Thomas B. DeMarse helped me enormously especially by letting me perform experiments, and he has been emotionally supporting my research as well. I owe my deepest gratitude to Dr. Murali Rao for bringing mathematical rigor to my clumsy ideas. My committee members Dr. Arunava Banerjee, Dr. Bruce Wheeler and Dr. Justin Sanchez supported me and kept me motivated. Dr. John Harris’s kind support allowed me to make friends and connections around the world. Dr. Purvis Bedenbaugh brought me a special Christmas gift of auditory spiking data in 2009.

I am indebted to many of my colleagues; without their support this dissertation would not have been possible. António Rafael da Costa Paiva has been a great friend and colleague for developing spike train based signal processing algorithms. Jianwu Xu and Weifeng Liu gave me great intuitions for reproducing kernel Hilbert spaces. Dongming Xu enlightened me on dynamical systems. Brain storming with Karl Dockendorf was always a pleasure. I learned so much from the discussions with Steven Van Vaerenbergh and Luis Sanchez. Among all the most fruitful collaboration was with Sohan Seth. He has been a great friend, and brought joy to my work.

I greatly appreciate all the support my friends gave me in a number of ways. I only mention a few of them here: Pingping Zhu the operator operator, Jason Winters the creative, Aysegul Gunduz the brilliant, Vaibhav Garg the good, Sachin Talathi the synchrony, Abhishek Singh the mango shake, Alexander Singh-Alvarado the happy omniscient, Savyasachi Singh the expert hermit, Rajasimhan Rajagovindan the attention measurer, Manu Rastogi the cheerful, Sungho Oh the sosa, Ashish Myles the councilor, Lin Li the strong willed, Florian Kuehnel the adventurer, Yujin Kim the great scientist,
Jennifer Jackson the kind, Dong-uk Hwang the wise, Erion Hasanbelliu the complainer, Katherine Feng the questioner, Austin Brockmeier the smart, and Yuriy Bobkov the polite electrophysiologist. Special thanks goes to Francesca Spedalieri. Without her support, I would have been in a completely different emotional state during the writing.

Last but not least, I would like to thank my family for all their love and support.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABSTRACT</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>1</td>
<td>INTRODUCTION</td>
<td>13</td>
</tr>
<tr>
<td>1.1</td>
<td>Neuroscience Applications</td>
<td>17</td>
</tr>
<tr>
<td>1.1.1</td>
<td>Change Detection</td>
<td>17</td>
</tr>
<tr>
<td>1.1.2</td>
<td>Estimation and Goodness-of-fit</td>
<td>18</td>
</tr>
<tr>
<td>1.1.3</td>
<td>Neural Assembly Identification</td>
<td>18</td>
</tr>
<tr>
<td>1.1.4</td>
<td>Coding Capacity</td>
<td>19</td>
</tr>
<tr>
<td>1.1.5</td>
<td>Functional Connectivity via Dependence Detection</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>SPIKE TRAIN SPACES</td>
<td>20</td>
</tr>
<tr>
<td>2.1</td>
<td>Representations of Spike Trains</td>
<td>20</td>
</tr>
<tr>
<td>2.2</td>
<td>Distance and Dissimilarity Measures</td>
<td>22</td>
</tr>
<tr>
<td>2.3</td>
<td>Reproducing Kernel Hilbert Space for Spike Trains</td>
<td>25</td>
</tr>
<tr>
<td>2.4</td>
<td>Kernel Design Problem</td>
<td>28</td>
</tr>
<tr>
<td>3</td>
<td>POINT PROCESS SPACES</td>
<td>30</td>
</tr>
<tr>
<td>3.1</td>
<td>Definition of Point Process</td>
<td>31</td>
</tr>
<tr>
<td>3.1.1</td>
<td>Random Counting Measure</td>
<td>31</td>
</tr>
<tr>
<td>3.1.2</td>
<td>Conditional Intensity Function</td>
<td>32</td>
</tr>
<tr>
<td>3.1.3</td>
<td>Probability Measure on Spike Train Space</td>
<td>33</td>
</tr>
<tr>
<td>3.2</td>
<td>Precisely Timed Spike Trains</td>
<td>34</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Precisely Timed Action Potential Model</td>
<td>36</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Precisely Timed Spike Train Model</td>
<td>38</td>
</tr>
<tr>
<td>3.3</td>
<td>Point Process Divergence and Dissimilarity</td>
<td>42</td>
</tr>
<tr>
<td>4</td>
<td>CUMULATIVE BASED DIVERGENCE</td>
<td>45</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>45</td>
</tr>
<tr>
<td>4.2</td>
<td>Extended Kolmogorov–Smirnov Divergence</td>
<td>46</td>
</tr>
<tr>
<td>4.3</td>
<td>Extended Cramér–von-Mises Divergence</td>
<td>48</td>
</tr>
<tr>
<td>4.4</td>
<td>Simulation Results</td>
<td>50</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Poisson Process</td>
<td>50</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Stationary Renewal Processes</td>
<td>51</td>
</tr>
<tr>
<td>4.4.3</td>
<td>Precisely Timed Spike Trains</td>
<td>53</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>4.4.4 Neuron Model</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>4.4.5 Serial Correlation Model</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>4.5 Optimal Stimulation Parameter Selection</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>4.5.1 Problem</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>4.5.2 Method</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>4.6 Conclusion</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>5 ( \phi )-DIVERGENCE AND HILBERTIAN METRIC</td>
<td>61</td>
<td></td>
</tr>
<tr>
<td>5.1 Hilbertian Metric</td>
<td>61</td>
<td></td>
</tr>
<tr>
<td>5.2 Radon-Nikodym Derivative</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>5.3 Hellinger Divergence</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>5.3.1 Estimator</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>5.3.2 Illustrative Example</td>
<td>68</td>
<td></td>
</tr>
<tr>
<td>5.3.3 Detecting Learning</td>
<td>68</td>
<td></td>
</tr>
<tr>
<td>5.3.4 Artificial Spike Trains</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td>5.3.5 Non-stationarity Detection</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>5.3.6 Kernel Size</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>5.4 Symmetric ( \chi^2 )-divergence</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>5.4.1 Point Process Representation</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>5.4.2 Estimation of Symmetric ( \chi^2 )-divergence</td>
<td>81</td>
<td></td>
</tr>
<tr>
<td>5.4.2.1 Stratification approach</td>
<td>81</td>
<td></td>
</tr>
<tr>
<td>5.4.2.2 Smoothed spike train approach</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>5.4.3 Results</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>5.4.3.1 Two action potentials</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>5.4.3.2 Inhomogeneous Poisson process</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>5.4.3.3 Stationary renewal processes</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>5.4.4 Auditory Neuron Clustering</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>5.4.5 Discussion</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>6 KERNEL BASED DIVERGENCE</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>6.1 Introduction</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>6.2 Kernel Based Divergence and Dissimilarity</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>6.3 Strictly Positive Definite Kernels on ( \mathbb{R}^n ) and ( L_2 )</td>
<td>93</td>
<td></td>
</tr>
<tr>
<td>6.3.1 Composite Kernels on ( \mathbb{R}^n )</td>
<td>93</td>
<td></td>
</tr>
<tr>
<td>6.3.2 Schoenberg Kernels (or radial basis function) on ( L_2 )</td>
<td>94</td>
<td></td>
</tr>
<tr>
<td>6.4 Representation of Spike Trains and Point Process Spaces</td>
<td>94</td>
<td></td>
</tr>
<tr>
<td>6.4.1 Smoothed Spike Train Space</td>
<td>94</td>
<td></td>
</tr>
<tr>
<td>6.5 Stratified Spike Train Kernels</td>
<td>94</td>
<td></td>
</tr>
<tr>
<td>6.6 Kernels on Smoothed Spike Trains</td>
<td>96</td>
<td></td>
</tr>
<tr>
<td>6.7 Simulation Results</td>
<td>97</td>
<td></td>
</tr>
<tr>
<td>6.7.1 Kernel PCA</td>
<td>97</td>
<td></td>
</tr>
<tr>
<td>6.7.2 Statistical Power</td>
<td>99</td>
<td></td>
</tr>
<tr>
<td>6.8 Discussion</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>6.9 Proof of Theorem 14</td>
<td>101</td>
<td></td>
</tr>
<tr>
<td>Table</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>-------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>5-1</td>
<td>Comparison of integration methods for Hellinger divergence</td>
<td>71</td>
</tr>
<tr>
<td>5-2</td>
<td>Statistical power comparison of dissimilarities and Hellinger divergence on a set of experiments</td>
<td>73</td>
</tr>
<tr>
<td>6-1</td>
<td>Statistical power of different strictly positive definite kernels induced divergences</td>
<td>97</td>
</tr>
<tr>
<td>6-2</td>
<td>List of kernels for spike trains of special interest and their corresponding time complexity</td>
<td>102</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>Illustration of spike train representations</td>
<td>22</td>
</tr>
<tr>
<td>2-2</td>
<td>Illustration of kernel methods and reproducing kernel Hilbert space</td>
<td>26</td>
</tr>
<tr>
<td>3-1</td>
<td>Non-stationary trial-to-trial variability</td>
<td>30</td>
</tr>
<tr>
<td>3-2</td>
<td>Stratified spike train space</td>
<td>34</td>
</tr>
<tr>
<td>3-3</td>
<td>Precisely timed spike train</td>
<td>35</td>
</tr>
<tr>
<td>3-4</td>
<td>Illustration of precisely timed spike train model with realizations</td>
<td>40</td>
</tr>
<tr>
<td>3-5</td>
<td>Precisely timed spike train modeling of neuronal culture data</td>
<td>41</td>
</tr>
<tr>
<td>4-1</td>
<td>Statistical power of cumulative based divergences on Poisson processes</td>
<td>51</td>
</tr>
<tr>
<td>4-2</td>
<td>Statistical power of cumulative based divergence on renewal processes</td>
<td>52</td>
</tr>
<tr>
<td>4-3</td>
<td>Example precisely timed spike trains</td>
<td>52</td>
</tr>
<tr>
<td>4-4</td>
<td>Statistical power of cumulative based divergence on precisely timed spike trains</td>
<td>53</td>
</tr>
<tr>
<td>4-5</td>
<td>Statistical power of cumulative based divergences on Izhikevich neuron model</td>
<td>54</td>
</tr>
<tr>
<td>4-6</td>
<td>Discrimination of serially correlation with cumulative based divergence</td>
<td>55</td>
</tr>
<tr>
<td>4-7</td>
<td>Spike trains from sensory cortex in response to natural and electrical stimulus</td>
<td>57</td>
</tr>
<tr>
<td>4-8</td>
<td>Divergence of artificial response to the natural</td>
<td>58</td>
</tr>
<tr>
<td>4-9</td>
<td>Detailed raster of natural stimulus and selected response</td>
<td>58</td>
</tr>
<tr>
<td>5-1</td>
<td>Toy example raster plots and estimated point process illustration</td>
<td>69</td>
</tr>
<tr>
<td>5-2</td>
<td>Convergence of Hellinger divergence estimator</td>
<td>70</td>
</tr>
<tr>
<td>5-3</td>
<td>Empirical distribution of Hellinger divergence</td>
<td>70</td>
</tr>
<tr>
<td>5-4</td>
<td>Significance of divergence based on Hellinger distance values before and after learning protocol</td>
<td>72</td>
</tr>
<tr>
<td>5-5</td>
<td>Performance of test with Hellinger divergence depends on the number of samples</td>
<td>73</td>
</tr>
<tr>
<td>5-6</td>
<td>Nonstationarity detection from culture stimulus</td>
<td>76</td>
</tr>
<tr>
<td>5-7</td>
<td>Non-stationarity detection with dissimilarities compared to Hellinger</td>
<td>77</td>
</tr>
<tr>
<td>5-8</td>
<td>Kernel size effect on Hellinger divergence estimator</td>
<td>78</td>
</tr>
</tbody>
</table>
5-9 Two action potential example to test $\chi^2$-divergence ........................................ 85
5-10 Performance of $\chi^2$-divergence on Poisson process ............................................... 86
5-11 Performance of $\chi^2$-divergence on renewal process ................................................. 87
5-12 Point process clustering of auditory neurons with $\chi^2$ divergence ........................... 88
6-1 Kernel principal component analysis of spike trains with strictly positive definite kernels ............................................................................................................. 98
6-2 Statistical power of kernel based divergences ............................................................... 99
7-1 Comparison of typical execution time of different divergence estimators ................. 106
7-2 Comparison of statistical power among proposed methods ........................................... 107
A-1 Reproducing kernel Hilbert space interpretation of ITL .............................................. 118
Abstract of Dissertation Presented to the Graduate School of the University of Florida in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

CAPTURING SPIKE TRAIN SIMILARITY STRUCTURE: A POINT PROCESS DIVERGENCE APPROACH

By

Il “Memming” Park

August 2010

Chair: José C. Príncipe
Major: Biomedical Engineering

Neurons mostly communicate via stereotypical events called action potentials (or spikes for short) giving rise to a time series called the neural spike train. Spike trains are random in nature, and hence one needs to deal with the probability law over the spike trains space (point process). Rich statistical descriptors are a prerequisite for statistical learning in the spike train domain; this provides necessary analysis tools for neural decoding, change detection, and neuron model fitting. The first and second order statistics prevalently used in neuroscience – such as mean firing rate function, and correlation function – do not fully describe the randomness, thus are partial statistics. However, restricting the study to these basic statistics implicitly limit what can be discovered. We propose three families of statistical divergences that enable non-Poisson, and more over, distribution-free spike train analysis. We extend the Kolmogorov-Smirnov test, φ-divergence, and kernel based divergence to point processes. This is possible through the development of novel mathematical foundations for point process representations.

Compared to the similarity or distance measures for spike trains that assumes predefined stochasticity and hence are not flexible, divergences applied to sets of spike trains capture the underlying probability law and measures the statistical similarity. Therefore, divergences are more robust, and assumption free. We apply the methodology on real data from neuronal cultures as well as anesthetized animals for neuroscience and neuroengineering applications posed as statistical inferences to evaluate their usefulness.
CHAPTER 1
INTRODUCTION

Imagine you have access to the firing pattern of all neurons in your brain. You will observe seemingly random signals [112]. How will you analyze the data to discover the internal representation and mechanisms for information processing? The brain is spontaneously active and it is perhaps impossible to control its state within the natural operating regime. So instead of the whole brain, you decide to study the sensory or motor subsystem where it is relatively easy to control the internal states and repeat experiments. You will still find a lot of variability in the spatio-temporal neural firing patterns [46, 103, 118], because the brain operates with noisy components. Although the propagation of an action potential within a neuron is quite reliable, the synaptic mechanism and subthreshold neuronal dynamics is stochastic [3, 4, 69, 70, 74]. Therefore, the resulting action potential sequences are stochastic as well. The brain, in a sense, knows this intrinsic stochasticity at the action potential level and builds a reliable system.

Hence, to analyze spike trains – temporal patterns of action potential firing – either an appropriate model for their stochasticity is required or tools that can quantify arbitrary stochasticity. Assumptions of the first kind have been widely deployed in neuroscience in the form of Poisson or renewal process models [13, 97, 98, 120]. These assumptions enable summarizing the complicated spike train observations to a few intuitive statistical quantities and enable building computational tools often with elegant analytic solutions. In fact, the Poisson assumption for spike train analysis is analogous to the Gaussian assumption for conventional (real-valued) stochastic processes; while it allows powerful and simple tools to be built, it often breaks down when the assumptions are violated [23, 116].

The stochastic structure of spike train observations is directly linked to neural codes. If one assumes a Poisson statistic, all the information is available in the rate function, since the rate function completely describes the process. Hence, neurons only need to decode the rate function; this fully supports the notion of rate code [112]. In a similar
way, a renewal process with a fixed interval distribution shape is also fully described by its mean rate function.

However, recent evidence supports the notion of a *temporal code* – a neural coding scheme that assumes that there is more information than what is contained in the mean rate function. The first set of evidences came from the early sensory systems where highly repeatable spatio-temporal pattern at the millisecond time scale were observed. These precisely timed spike trains are abundantly reported both *in vitro* and *in vivo* preparations [26, 28, 54, 66, 74, 102, 103]. Precise phase locking to local field potential oscillation is also widely observed *in vivo*, and it is considered to create population wide precise activity patterns [115]. Synchrony is another temporal coding scheme where neurons fire with highly time locked manner compared to the coincidence probability predicted from the rate function modulations [41]. Modulation of Fano factor (a measure of spike count variability) without co-modulation of the mean firing rate was observed in various tasks and areas in monkey which indicates Poisson statistics is not enough [19]. Although it remains to be shown that these non-Poisson statistics are indeed used by the brain for decoding (rather than being epiphenomena), these evidences suggest that the Poisson/renewal assumption is not always appropriate for analysis.

This dissertation was inspired by the information theoretic learning (ITL) framework [101] for non-Gaussian, non-linear signal processing, and generalizes the ITL approach to the spike trains in several ways. In ITL, information theoretic quantities – entropy, and mutual information – are utilized in developing signal processing and machine learning algorithms. Instead of the traditional mean squared error (MSE) as a central cost function which is optimal in the maximum likelihood estimation (MLE) sense for the white Gaussian noise, information theoretic quantities are used as cost functions in ITL. While the MSE only considers up to second order statistics, information theoretic quantities reflect all aspects of the underlying probability law. The estimation of the information theoretic quantities can be done efficiently, and in fact the success of ITL in part lies in its
simple distribution-free estimators that are derived from Parzen windowing. The estimator provides interesting connection to kernel methods, via the theory of reproducing kernel Hilbert space (RKHS). We present the detailed interpretation and connections in the appendix A.1.

In the spirit of ITL, our goal is to develop distribution-free statistical divergences for point processes (stochastic spike train observations). A statistical divergence (or simply divergence) is a statistic that quantifies the difference between the probability laws – it takes the value of zero if and only if the probability laws are identical, and otherwise it takes a positive real value which indicates how different they are. Many divergences are generalization of information theoretic quantities; for example the Kullback-Leibler divergence has a central role in Shannon’s information theory, the Rényi’s $\alpha$-divergence, and $\phi$-divergence families have their own information theoretic interpretations [22, 67, 89]. In fact, an entropy can often be derived from the divergence against a Lebesgue measure (which is not a probability measure) [20].

Divergences for point processes have been explored in the literature for parametric point processes. Kullback-Leibler divergence, Hellinger distance, and total variation, for example, are often used under Poisson assumption [104]. For general point processes, the Prohorov distance and variation distance can be defined, however, they only have theoretical value since no estimator is provided [58]. Estimation of Kullback-Leibler divergence on binned spike train has been proposed [55], however, due to the large data requirement, it is difficult to apply in practice.

We approach the problem of designing point process divergences while keeping in mind practical estimation. The most straightforward approach is to use a distribution-free point process estimation and plug it into known divergences such as Hellinger distance.\footnote{\int (f(x) - g(x))^2 \, dx$ where $f$ and $g$ are probability densities} We develop distribution-free point process estimators by extending the ideas of empirical
cumulative distribution function (CDF), and kernel density estimation (also known as Parzen windowing) from Euclidean space to spike train space. The extended empirical CDF leads to the generalization of Kolmogorov–Smirnov (K-S) and Cramér–von-Mises (C-M) tests, while the kernel density estimation allows the estimation of a family of divergences known as Hilbertian metrics.

Another generic approach is to use the likelihood ratio based divergences – if the likelihood ratio deviates from 1, the two point processes are different. $\phi$-divergence is based on information theoretic extensions of likelihood deviation. Thus, we develop likelihood ratio estimators for point processes. It is trivial to utilize the probability density estimator to evaluate the likelihood ratio, however, we can skip the complicated estimation of two densities and directly estimate the ratio using kernel regression. Kernel regression requires a symmetric positive definite kernel on the spike trains, which we have developed in previous studies [80, 90]. We show that such kernels are powerful enough to represent any $L_2$ function from the spike trains to reals; thus a good approximation of the likelihood ratio can be obtained and plugged in to estimate $\phi$-divergence.

In fact, kernels provide not only the ability to approximate functions, but also the ability to represent the spike trains in a Hilbert space (a vector space with inner products). Under certain conditions,\textsuperscript{2} we can show that each point process is also uniquely represented in the same Hilbert space. Since the Hilbert space is equipped with a metric, the distance between two point processes in the Hilbert space is also a point process divergence. As a matter of fact, the unique representation corresponds to the expectation of the spike train realizations in the Hilbert space of the point process, and can be easily estimated by empirical mean.\textsuperscript{3} We also show this formulation of divergence and its estimation holds without the symmetry requirement for the kernel (hence not

\textsuperscript{2} Hermitian strictly positive definite kernel.

\textsuperscript{3} convergence is guaranteed by the law of large numbers
having explicit Hilbert space embedding). This family of divergence is a generalization of many divergences and dissimilarities in quadratic form such as $D_{ED}$ of ITL, C-M test statistic on count distribution, and square integrated distance of mean rate functions.

Once we have statistical divergences, we can perform statistical inference on point process. Two major components of statistical inference are estimation and hypothesis testing [58]. Estimation theory requires an objective function or a loss function which captures the optimality. A divergence can provide a proper similarity of the estimate to the data. For hypothesis testing, one builds a hypothesis of the form $P = Q$ where $P$ and $Q$ are two probability laws underlying spike train models or data sets. A divergence can be used to measure the deviation from $P = Q$ with a single number. Unlike partial summarizing statistics such as count distribution, or parametric testing, divergences are fundamentally well suited for hypothesis testing, since given a consistent estimator it will always correctly reject in the asymptotic large data limit.

### 1.1 Neuroscience Applications

A partial list of applications of point process divergences and statistical inferences in the context of neuroscience is presented in this section.

#### 1.1.1 Change Detection

One of the fundamental questions that is constantly encountered in a variety of neuroscience problems is whether two sets of spike train observations follow the same probability law or not. Detecting change is crucial in neuroscience experiments where stable conditions must be ensured across trials, and finding which protocol induces significant change to the neural system is often part of the research goal. From a statistical perspective, this question can be framed as a hypothesis testing problem using a suitable divergence measure; the statistical rejection of null-hypothesis of $P = Q$ would support that there is a change indeed.

The problem of detecting plasticity and non-stationarity of a neural system from spike train observation has been a challenge for neuroscience, and it hindered the plasticity
research using external electrode arrays [91, 92]. Researchers would report changes in the firing rate profile which indicate structural change in the network, but they did not have a sophisticated tool to test for subtle changes other than the rate. We will apply the developed methods to such plasticity experiments on *in vitro* settings without any assumption on the system in section 5.3.5.

### 1.1.2 Estimation and Goodness-of-fit

How good is a model that generates spike trains with respect to the target spike train data set? The standard MSE is not appropriate for measuring the goodness-of-fit of spiking models. For example, the problem of searching for a set of parameters that produces spike trains closest to the specified probability law can be tackled with divergences. We demonstrate the power of divergences for the optimal stimulation parameter search problem in section 4.5.

Using a divergence as the loss function, one can also build *minimum divergence estimators* [89]. For example, in neuroscience Sharpee proposed the use of mutual information (Kullback-Leibler divergence), and Paninski proposed the use of $\phi$-divergence as a tool to estimate the linear portion of the linear-nonlinear spiking neuron model [86, 113]. While the widely used first order and second order based methods – spike triggered averaging (or reverse correlation), and spike triggered covariance techniques – are shown to be inconsistent estimators for this problem, the corresponding divergence based estimators are consistent [86].

### 1.1.3 Neural Assembly Identification

A neural assembly is a group of neurons that work together transiently for a particular task [78]. When electrode array is used to record many neurons in *in vivo* experiments, one of the interesting tasks is to find neural assemblies. It is widely assumed that those neurons fire in synchrony [8, 39, 41], however, there is no biophysical restriction to elect synchrony as the common property. If we can find any statistical similarities between the neuronal population firing pattern, it would be fair to assume that the probed
neurons form a *statistical neural assembly*. In a sense, a statistical neural assembly can be used to perform ensemble averages within a single trial. To assist the identification of statistical neural assemblies, we need to identify neurons with similar statistics – a task that can be solved by clustering algorithm on the divergence values. We demonstrate neuronal response clustering in section 5.4.4.

1.1.4 Coding Capacity

Mutual information is often used to quantify the efficiency of neural code of a spiking neural system [55, 88, 123]. Neural decoding can be viewed as a classification problem when the input is categorical. The classification error limit is due to the overlap of the probability laws given each category. Divergences can often strongly bound the classification errors [67, 117]. Hence, we can analyze the coding capacity by measuring the divergences.

1.1.5 Functional Connectivity via Dependence Detection

A divergence measure can be extended to a (in)dependence measure when the joint and product of marginal probability measures are applied. The nonparametric estimator can also be used to estimate the joint distribution of a pair of spike trains, hence it is possible to extend the method to point process dependence measure.

In the same stimulation paradigm where the input is kept constant, the dependence between point processes implies dependence between neurons. From a signal processing point of view, cross correlation and Granger causality have been applied to binned spike trains to estimate functional connectivity [76, 93]. However, they implicitly assume that the first and second order statistics carry the information. The proposed approach can be a valuable addition for functional connectivity analysis without assumptions. The functional connectivity can also be applied to identifying neural assemblies.
CHAPTER 2
SPIKE TRAIN SPACES

Neurons communicate through a sequence of stereotypical action potentials. Since the amplitude of each action potential is virtually constant, the only information is in the timing of each event. This biological signal abstraction as a sequence of event times is called the spike train. Spike trains are different from conventional amplitude based signals (either in continuous time or in discrete time) in the sense that they cannot be represented in a linear space such as $L_2$ or $l_2$. In fact, spike trains lack a natural algebraic structure: a major roadblock. Neither additive nor multiplicative structures is naturally defined for spike trains.

In this chapter we discuss how spike trains can be mathematically represented, and review two ways of introducing basic structures to the spike train space. The first structure is the concept of distance, and the second is a kernel that induces a Hilbert space. We discuss the advantages and disadvantages of these approaches and pose the problem of capturing the underlying probability structure to better quantify similarity.

2.1 Representations of Spike Trains

Spike trains can be represented as several different types of mathematical objects. These different representations lead to different types of divergences in later chapters.

**Definition 1** (Sequence representation of a spike train). Let $\omega = (t_1, t_2, \ldots)$ be an enumerable sequence of real numbers in strictly increasing order where each element $t_i$ denotes the time of i-th action potential observation. We call $\omega$ the sequence representation of the spike train.

This representation is straightforward and frequently used to store spike trains in devices such as digital computers. Practically for a finite observation interval, $\omega$ has a finite length. Note that even though each sequence is of finite length, their corresponding lengths are different in general. Hence, this representation does not allow a vector space embedding unless an explicit padding rule makes them equal length. Padding
makes undesirable assumptions about correspondence between \(i\)-th spike timings that is unnatural, and not very robust to insertion or deletion of action potentials.

A more technical representation used in the mathematics community is to represent spike trains as counting measures. A measure is a function, hence operations on functions can be extended to spike trains. Since we will be using measure theory frequently in this thesis, here we briefly present the basic definitions (extensive discussion can be found in Daley and Vere-Jones [25], Halmos [42], Karr [58]).

Let \(\Omega\) be a set. A set of subsets \(\mathcal{F}\) of \(\Omega\) is a \(\sigma\)-algebra if \(\Omega \in \mathcal{F}\) and it is closed under set union and complement. The pair \((\Omega, \mathcal{F})\) is called a measurable space and the elements of \(\mathcal{F}\) are called measurable sets. A measure is a function \(\mu : \mathcal{F} \rightarrow \mathbb{R}^+\) such that it is countably additive, that is, if \(A_1, A_2, \ldots\) is a sequence of disjoint measurable sets, \(\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i} \mu(A_i)\). A counting measure is a measure such that \(\mu(A) \in \mathbb{N}\) for any measurable set \(A\). A simple counting (point) measure is a counting measure such that \(\mu(\{x\}) \leq 1\) for \(x \in \Omega\). The fundamental simple point measure is the Dirac measure \(\delta_x(A) = 1_{x \in A}\) where \(1\) is the indicator function.

Since we are dealing with time as the primary space for events, the spike train is represented as a simple counting measure on the real line. Specifically, we use the natural topological structure of \(\mathbb{R}\) and use the Borel algebra \(\mathcal{B}(\mathbb{R})\) as the \(\sigma\)-algebra. A Borel algebra of a topological space is the smallest \(\sigma\)-algebra that contains all the open sets. For a sequence of action potential timings \(t_1, t_2, \ldots, t_n\), the corresponding counting measure is the sum of Dirac measures:

\[
N = \sum_{i=1}^{n} \delta_{t_i}
\]

By the virtue of the real line and extension theorems [42], the counting measures on \((\mathbb{R}, \mathcal{B}(\mathbb{R}))\) can be fully specified by an integer valued function on \(\mathbb{R}\), \(N(t) := N((−\infty, t])\) (see figure 2-1) that is right continuous with left limits (càdlàg). We denote the space of all simple counting measures as \(\mathcal{M}_s\).
An alternative representation scheme is to transform each spike trains to a unique function in $L_2$. Many smoothing schemes fall into this category. A stable causal linear filter $h(t) \in L_2$ applied on the sequence representation, $\sum_{i=1}^{n} h(t - t_i)$, is widely used to create functions in $L_2$. Other non-linear and non-causal schemes are also possible [44]. The binning approach with a small bin size can also be thought of as a crude approximation of smoothing where the uniqueness holds up to timing jitter within each bin. This representation is directly linked to the kernel based representation (section 2.3).

These representations – time sequence, simple counting measure, and smoothed spike trains – are equivalent in the sense that they provide a scheme that uniquely represents each spike train. However, different set of mathematical tools can be more easily applied to each of them.

### 2.2 Distance and Dissimilarity Measures

Similarity and dissimilarity can be imposed on objects without other algebraic structure such as the linear structure, and still allow some pattern recognition, and visualization [96]. In a metric space, algorithms such as $k$-nearest neighbor classifier can be implemented. In this section, we discuss various distance / dissimilarity / similarity measures proposed in the literature [85].

Through decades of neuroscience research on neural codes, we learned that the firing rate and precise timing of action potentials are important for encoding information, and hence similarity in the mean rate function or spike timing should imply similarity or
closeness of spike trains. Given the same or similar experimental conditions, the resulting observed spike trains should be similar and as the condition deviates, the distance should increase. In a probabilistic point of view, ideally, a distance between objects should be inversely related to the posterior probability distribution [96]. The intuitive notion of rate similarity and timing similarity have inspired the two important spike train metrics: Victor-Purpura distance [123], and van Rossum distance [121].

Victor and Purpura suggested to extend the ideas of edit distance that had been successfully used in natural language processing and bioinformatics to a pair of spike trains [123]. Victor-Purpura (VP) distance is the first binless distance measure proposed in the neuroscience literature. Since the trial-to-trial variability of the spike trains suggest the jitter of spike timing and missing spikes to be considered as noise [118], their approach of assigning cost to shifting an action potential in time and adding or removing an action potential is motivated by neurophysiological ideas that emphasizes coincidence.

VP distance between spike trains is based on the cost in transforming one spike train into the other. Three elementary operations in terms of single spikes are established: moving one spike to perfectly synchronize with the other, deleting a spike, and inserting a spike. Once a sequence of operations is set, the distance is given as the sum of the cost of each operation. While the cost of deleting or inserting a spike is set to one, the cost of aligning a spike at $t$ to $t + \Delta t$ is $q|\Delta t|$, where $q$ is a scaling parameter for relative temporal sensitivity. Because a higher $q$ implies that the distance increases more when a spike needs to be moved, the distance as a function of $q$ controls the precision of the spike times of interest.

Recall the axioms of a metric; for spike trains $\omega_i$, $\omega_j$ and $\omega_k$:

(i) Symmetry: $d(\omega_i, \omega_j) = d(\omega_j, \omega_i)$

(ii) Positiveness: $d(\omega_i, \omega_j) \geq 0$, with equality holding if and only if $\omega_i = \omega_j$

(iii) Triangle inequality: $d(\omega_i, \omega_j) \leq d(\omega_i, \omega_k) + d(\omega_k, \omega_j)$. 

23
To ensure the triangle inequality and uniqueness of the distance between any two spike trains, the sequence which yields the minimum cost in terms of the operations is used. Therefore, the VP distance between spike trains \( \omega_i \) and \( \omega_j \) is defined as

\[
D_{\text{VP}}(\omega_i, \omega_j) = \min_{C(\omega_i \leftrightarrow \omega_j)} \sum_l d_q(t_{c_i[l]}^i, t_{c_j[l]}^j),
\]

where \( C(\omega_i \leftrightarrow \omega_j) \) is the set of all possible sequences of elementary operations that transform \( \omega_i \) to \( \omega_j \), and \( c_i[l] \in C(\omega_i \leftrightarrow \omega_j) \) where \( c_i[l] \) denotes the index of the spike time of \( \omega_i \) manipulated in the \( l \)-th step of a sequence. \( d_q(t_{c_i[l]}^i, t_{c_j[l]}^j) \) is the cost associated with the step of mapping the \( c_i[l] \)-th spike of \( \omega_i \) at \( t_{c_i[l]}^i \) to \( t_{c_j[l]}^j \), corresponding to the \( c_j[l] \)-th spike of \( \omega_j \). Note that VP distance is non-Euclidean.

\( d_q \) is a distance metric between two spike times. Suppose two spike trains with only one spike each, the mapping between the two spike trains is achieved through the three above mentioned operations and the distance is given by

\[
d_q(t, s) = \min \{ q|t - s|, 2 \}
= \begin{cases} 
q|t - s|, & |t - s| < 2/q \\
2, & \text{otherwise}
\end{cases}
\]

(2-2)

This means that if the difference between the two spike times is smaller than \( 2/q \) the cost is linearly proportional to their time difference. However, if the spikes are further apart, it is less costly to simply delete and insert appropriately.

Later, Houghton and Sen [45] showed that VP distance yields in practice almost identical results to \( L_1 \) distance of smoothed spike trains with a rectangular kernel. VP metric has been used by the neuroscience community for analyzing precision and potential of temporal coding [99, 103].

van Rossum [121] suggested the use of \( L_2 \) distance between smoothed spike trains. Because the smoothing effectively transforms the spike train to real valued signals, \( L_2 \) distance is a natural choice. This is also a natural extension of the traditionally used time
binning method where the smoothing kernel is rectangular and the smoothed function is sampled on a discrete grid on time. van Rossum distance also has biological relevance, since the smoothing can be considered as a first order approximation of the synaptic signal transduction process [121].

Houghton [44] recently extended this idea to incorporate physiological dynamics at the synapse. As in van Rossum [121], $L_2$ distance is defined between the post-synaptic potentials evoked according to dynamical models of synapses. In spike trains from zebra finch recordings, Houghton [44] showed an improvement in classification performance which was particularly significant when accounting for the depletion of receptors, suggesting the history dependence in the interaction of action potential timings. We showed that the van Rossum and Houghton’s distance measures can be derived from the perspective of kernels [82, 84].

2.3 Reproducing Kernel Hilbert Space for Spike Trains

The main advantage of the smoothed spike train approach is that the spike trains can be treated as regular continuous amplitude, continuous time signals. $L_2$ is a Hilbert space (linear structure with inner product and norm), therefore it allows traditional signal processing tools to be directly applied to spike trains. Linear filtering, cross-correlation analysis and most supervised learning algorithms can be applied to mention a few.

The kernel method is a powerful machine learning technique where linear algorithms can be easily extended to nonlinear ones by the theory of reproducing kernel Hilbert space (RKHS). Given a bivariate symmetric positive definite kernel defined on the input space, the objects in the input space can be mapped to a Hilbert space where the inner product is defined by the evaluation of the kernel [2] (See figure 2-2). The success of support vector machine (SVM) brought the development of a variety of kernel methods such as kernel regression [109], kernel principal component analysis (KPCA) [110], kernel least mean squares (KLMS) [68], and kernel fisher discriminant [109] to name just a few.
Figure 2-2. Illustration of reproducing kernel Hilbert space induced by symmetric positive definite kernel. Elements in the unstructured space $\mathcal{X}$ are projected to an Hilbert space $\mathcal{H}$ via the kernel $K$. Linear structure and inner product are indicated with dotted lines.

The kernel method is particularly useful to data types that do not have natural representation in Euclidean space. Many specialized kernels were designed for strings, bags of objects, images, shapes [109], and transducers [21, 62]. However, they rely on the prior knowledge and insight of the designer.

We have independently developed a kernel method that induces an RKHS that can isometrically embed the van Rossum distance [82, 84, 94]. In addition, the Cauchy-Schwarz divergence which measures the angular distance in the RKHS corresponds to the similarity measure used by Schreiber et al. [111] to perform clustering of in vitro spike train observations.

Using a kernel smoothing representation of the spike train, define the smoothed spike trains as,

$$\hat{\lambda}_i(t) = \sum_{m=1}^{N_i} \kappa_{pdf}(t - t^i_m)$$ (2–3)

where $t^i_m$ is the $m$-th spike timing of spike train indexed by $i$. We define the memoryless cross intensity (mCI) kernel on spike trains as,

$$I(\omega_i, \omega_j) = \int_T \hat{\lambda}_{\omega_i}(t)\hat{\lambda}_{\omega_j}(t)dt.$$ (2–4)
The mCI kernel is symmetric positive (semi-)definite and hence induces an RKHS [84]. This kernel is closely related to other single trial rate estimation processes such as various binning methods. Hence, using this kernel resembles many linear methods that have been widely used in neuroscience.

One important property of the mCI kernel is that when \( \kappa_{pdf} \) is an exponential decay\(^1\), (2–4) can be evaluated efficiently as,

\[
I(\omega_i, \omega_j) = \frac{1}{N_i N_j} \sum_{l=1}^{N_i} \sum_{m=1}^{N_j} K(t_i^{(j)} - t_m^{(j)})
\]  
(2–5)

where \( K(\cdot) \) is the double exponential function (Laplacian kernel) [84]. Also, in this case the van Rossum distance coincides with the distance in this RKHS.

To overcome the linear properties of the kernel inherited from \( L_2 \) space, we can extend it with additional nonlinearity. We have proposed two such nonlinear kernels called nCI\(_1\) and nCI\(_2\). The nCI\(_1\) is defined as,

\[
K_1(\omega_i, \omega_j) = \exp \left[ -\frac{\|\lambda_i(t) - \lambda_j(t)\|^2}{\alpha} \right],
\]  
(2–6)

while nCI\(_2\) is defined as,

\[
K_2(\omega_i, \omega_j) = \int \exp \left[ -\frac{\lambda_i(t) - \lambda_j(t)}{\alpha} \right] \, dt.
\]  
(2–7)

It is a nonlinear transformation of the structure that is created by the mCI RKHS. We have successfully applied this technique for continuous-time cross correlation [93] and spectral clustering [82, 83]. We will discuss the implication of the nCI kernel in the light of strictly positive definite functions in chapter 6.

\(^1\) This corresponds to a first order IIR low-pass filter
2.4 Kernel Design Problem

What is the optimal kernel for a specific problem? Choice of a kernel determines the structure of the space where the data is linearly represented [81, 109]. In general, the kernel shall capture the similarity of the data, and ideally the distance in the space should correspond to the posterior distribution for classification [96]. The problem of finding a kernel that best approximates this condition is known as the “kernel design problem”.

All of the previously mentioned distance or similarity measures are based on the insight that the exact spike timing (or synchronized instantaneous increase of firing rate) is an important feature. Although the measures have a few free parameters related to time scale that can be optimized, they cannot cope with the general statistical structure of the observations. In other words, these measures impose a rigid structure to the spike train space, as parametric models do. In fact, the structures are closely related to the smoothing process; injective mapping of the spike trains to the $L_1$ or $L_2$ space.

There are a set of kernel design methods that utilize a generative modeling of the data such as the Fisher kernel [50], product probability kernel [51], Hilbertian metrics [43], and non-extensive entropic kernel [71] in Euclidean spaces. The Fisher kernel is one of the earliest kernels proposed for generative probability models [49, 50]. Given a parametric family of probabilistic models, the two objects are mapped using the score function (gradient of the log likelihood) and inner product is taken with respect to the Riemannian manifold. The idea is valid when the probability law is in the exponential family with fixed number of parameters, such that the representation becomes the space of sufficient statistics.

In information theoretic learning [101], based on the optimization principles for Rényi’s quadratic entropy, a space of square-integrable probability distributions is used, and it is an RKHS based on the cross information potential kernel\(^2\) as the inner product

\[^2\] \[ K(X, Y) = \int f_X(z)f_Y(z)dz \] where $f_X$ is the probability density of $X$. 

28
between random variables (section A). This kernel can also be viewed as a conditional expectation of the probability density with respect to the other random variable. Hence, Jebara et al. [51] calls it an expected likelihood kernel, and generalizes it to a broader family of kernels of the form $K(p, q) = \int p^r(x)q^r(x)dx$ where $p$ and $q$ are two probability densities, and $r > 0$ (product probability kernel). Given a nonparametric estimator for the probability densities, these kernels can be readily implemented. When a single trial estimation using Parzen windowing is used, it forms a subset of regular kernel methods. Also, when $r = \frac{1}{2}$, it is the Bhattacharyya kernel which is related to the Hellinger distance.

Hellinger distance has a nice property that it is independent on the reference measure for integration. Hein and Bousquet [43] proposed a family of such metrics for probability measures which can be embedded isometrically in a Hilbert space, and presented the kernel that provides inner product between probability measures. This family of Hilbertian metrics provides a foundation for designing kernels from generative models with non-Euclidean representation. The family of RKHS include the classical Hellinger, Shannon-Jensen, $\chi^2$-divergence, and total variation. Martins et al. [71] also suggested a family of kernels on probability measures based on Tsallis-entropy. We will explore the Hilbertian metrics applied for point processes in chapter 5.
Trial-to-trial variability studies clearly show that spike train observations are noisy. Perhaps due to the diversity and heterogeneity of neural systems, this variability structure is highly heterogeneous for neurons or given different experimental conditions; while one neuron may show Poisson process like variability, another neuron may show precisely timed action potentials, or the same neuron might show precise time locking behavior for strong stimulus and increase the time jitter as the stimulus gets weaker. (See figure 3-1 for a trial-to-trial variability example)

Figure 3-1. Recording from ventral auditory thalamus of a rat. Three different frozen noise stimulus was presented: dynamic ripple noise (DRN), filtered gaussian noise, and speech babble from International Collegium of Rehabilitative Audiology (ICRA). Each box contains 56 trials, and the gray line is the smoothed PSTH.

The trial-to-trial variability should be considered as noise in the input for a downstream neuron. To overcome the noise, there are several possible strategies – post-synaptic smoothing and pattern detection, both at the single spike train train level as well as in the population activity. An ideal neuron would encode the information that would survive the spike train noise, and optimally decode given the knowledge of the noise statistics. The
information needs to be conveyed in the maximally “orthogonal dimension” for the noise to be efficiently transferred; signals should be separated in a way the most dissimilar to the noise. However, the relevant “coding dimension” is by no means unique, and will be relative to the tuning of the down stream neuron.

Hence, in the context of optimal decoding, capturing the noise structure is critical. The wide dynamic range of stochasticity of spike train observations are best captured by point process models; A point process is a probabilistic description of spike train generation. In this chapter, we rigorously define point processes based on the representation of spike trains we discussed in the previous chapter. We introduce a point process model of precisely timed spike trains as a benchmark for trial-to-trial variability. In later chapters, we will develop point process divergences as the key statistic for distribution-free statistical inference.

3.1 Definition of Point Process

A point process is a probability law that governs the generation of spike trains. There are multiple formal frameworks to rigorously define point processes.

3.1.1 Random Counting Measure

Analogous to random variables that describe probability laws in Euclidean space \( \mathbb{R}^d \), a random counting measure is a function from the sample space to the space of counting measures. Let \((\Omega, \mathcal{F}, P)\) be a probability space where \(\Omega\) is the sample space, \(\mathcal{F}\) is the \(\sigma\)-algebra, and \(P\) is the probability measure. A measurable function \(N : \Omega \to \mathcal{M}_s\) is a random counting measure that defines a point process.

Since \(\mu \in \mathcal{M}_s\) is a function of the form \(\mu : \mathcal{B}(\mathbb{R}) \to \mathbb{N}\), we can see that \(N : \Omega \to \mathcal{B}(\mathbb{R}) \to \mathbb{N}\), and it can also be represented as \(N : \mathcal{B}(\mathbb{R}) \to \Omega \to \mathbb{N}\). Thus \(N(A)\) is an integer-valued random variable given a Borel set \(A\). Specifying a consistent joint probability law of collection of random variables \(N(A_1), N(A_2), \ldots, N(A_n)\) for any \(\{A_i \in \mathcal{B}(\mathbb{R})\}\) also is equivalent to defining a point process.
Let us define the (inhomogeneous) Poisson process as an example\(^1\). Given an intensity measure \(\Lambda : \mathcal{B}(\mathbb{R}) \rightarrow \mathbb{R}^+\), the random counting process of a Poisson process is specified by independent integer valued random variables,

\[
\Pr[N(A) = k] = \frac{\Lambda(A)^k}{k!}e^{-\Lambda(A)}
\]  

(3–1)

for any \(A \in \mathcal{B}(\mathbb{R})\) where \(k \in \mathbb{N}\) [116].

### 3.1.2 Conditional Intensity Function

Recall that \(\mu \in \mathcal{M}_s\) in time can also be described by non-decreasing staircase functions \(\mu : \mathbb{R} \rightarrow \mathbb{N}\). Again, we can see that \(N : \Omega \rightarrow \mathbb{R} \rightarrow \mathbb{N}\) as well as \(N : \mathbb{R} \rightarrow \Omega \rightarrow \mathbb{N}\) are valid representations. Indeed this is the same description as random processes where at each time point, the value is random \(N(t)\). For such a point process representation that satisfies the usual conditions, the martingale theory allows a decomposition and we can obtain the compensator process (Doob-Meyer decomposition). More rigorous conditions are omitted here, refer to Karr [58] for details.

The compensator of a point process can be simply represented by the conditional intensity function defined as

\[
\lambda^*(t) = \lim_{\delta \to 0} \frac{\mathbb{E}[N((t, t+\delta]|\mathcal{H}_t)]}{\delta}
\]

(3–2)

where \(\mathcal{H}_t\) is the \(\sigma\)-algebra of events occurring at times up to but not including \(t\) [25]. The conditional intensity function completely describes any point process [25]. The conditional intensity function can be parameterized with a generalized linear model (GLM) which often provides a unique maximum likelihood fitting and is popular in point process modeling [12, 79, 87, 99].

\(^1\) For detailed description of Poisson process see B
The first order statistic of the conditional intensity representation is the intensity function, also known as the (mean) rate function $\lambda : \mathbb{R} \rightarrow \mathbb{R}^+$,

$$\lambda(t) = \mathbf{E}[\lambda^*(t)].$$  \hfill (3–3)

For Poisson process, $\lambda^*(t) = \lambda(t)$ meaning the process is independent of the history, and the intensity function is simply related to the intensity measure by $\int_a^b \lambda(t)dt = \Lambda([a,b])$. The peri-stimulus time histogram (PSTH) is the conventionally used estimator for $\lambda(t)$ [97, 98].

### 3.1.3 Probability Measure on Spike Train Space

Here we take a rather unconventional route to avoid unnecessary complication that may hinder the presentation of the proposed statistics. Instead of assuming an arbitrary probability space as the previously introduced approaches have taken, we directly define the point processes to be a probability measure on the spike train space. This allows simple descriptions for our methods and is appropriate for non-stationary fixed time interval observations. In order to define measures, we first need a measurable space for spike trains.

Let $\Omega$ be the set of all finite spike trains on a bounded interval $\mathcal{X} \subset \mathbb{R}$, that is, each $\omega \in \Omega$ can be represented as a finite set of events (action potentials) $\omega = \{t_1 < t_2 < \ldots < t_n\}$ for some $n \in \mathbb{N}$. We stratify $\Omega$ into disjoint partitions $\Omega_n = \mathcal{X}^n$; $\Omega_n$ consists of all realizations with $n$ events (see figure 3-2). We can use the topology induced from $\mathbb{R}$ and define measurable spaces $(\Omega_n, \mathcal{F}_n)$ where $\mathcal{F}_n = \mathcal{B}(\Omega_n)$. Let $\mathcal{F} = \sigma(\bigcup_n \mathcal{F}_n)$ be the $\sigma$-algebra generated from the Borel sets. Note that $\Omega = \cup_n \Omega_n$ is a disjoint union, and $(\Omega, \mathcal{F})$ is a measurable space. We define the point process as a probability measure on $(\Omega, \mathcal{F})$.

Note that any measurable set $A \in \mathcal{F}$ can be partitioned into $\{A_n = A \cap \Omega_n\}_{n=0}^\infty$, such that each $A_n$ is measurable in corresponding measurable space $(\Omega_n, \mathcal{B}(\Omega_n))$. Here $A$ denotes a collection of spike trains involving varying number of action potentials and
corresponding action potential timings, whereas $A_n$ denotes a subset of these spike trains involving only $n$ action potentials each.

It is convenient to have a reference measure on the space. Let $\mu(A) = \delta_{\Omega_0}(A) + \sum_{n=1}^{\infty} \mu_n(A \cap \Omega_n)$ denote the extension of Lebesgue measures $\mu_n$ for $\Omega_n (n > 0)$ to $\Omega$, and $\delta_{\Omega_0}$ is a Dirac measure for the single element in $\Omega_0$ which represents the empty spike train. We will use $\mu$ as the reference measure when necessary. However, it will turn out that the choice of $\mu$ is not important since $\phi$-divergences are independent of the reference measure.

![Figure 3-2](image)

Figure 3-2. Illustration of how the spike train space is stratified. It can be written as a union of Euclidean spaces. See section 3.1.3 for details.

A point process is defined as a probability measure $P$ on the measurable space $(\Omega, \mathcal{F})$ [25]. We call this the \textit{stratified approach} of representing point processes. Similar approach can be used to induce probability measures on the smoothed spike train space (see section 5.4.1).

\section{3.2 Precisely Timed Spike Trains}

When the same stimulation is presented to a neuronal system, the train of action potentials that are observed as a result sometimes show a highly repeatable spatio-temporal
pattern at the millisecond time scale. Recently these precisely timed spike trains are abundantly reported both in vivo and in vitro preparations [26, 28, 54, 66, 74, 102, 103]. It has been speculated that this spatio-temporal pattern can support temporal coding [122]. Despite being highly reproducible, different forms of trial-to-trial variability have also been observed [118]. It is crucial to understand this variability since to utilize a precisely timed spike pattern as a temporal code, the system should presumably be robust to its variability structure, and possibly learn to reduce it [11]. In this section, we develop a parametric point process model that intuitively captures such variability structure.

Precisely timed spike trains (PTST) consists of precisely timed action potentials (PTAP) which are highly repeatable in time (Fig. 3-3). The temporal jitter with respect to stimulation onset is known as the precision of an PTAP, and the portion the PTAP identified across trials is known as the reliability. Precision and reliability are the main characterization of the variability of PTAP [118]. In the PTST, normally there are also additive action potentials that do not correspond to a PTAP which is additive to the trial-to-trial variability.

These variabilities of PTSTs are not appropriately described by the widely used point process models such as Poisson process or renewal models which describe the rate models

Figure 3-3. Illustration of different variabilities in a precisely timed spike train: precision, reliability, and additive action potentials. The jitter distribution on the bottom represents precision, the dotted box indicates a missing action potential (reliability), and the red action potential represents an additive noise which is not precisely time locked.
well [1]. Although the PTST’s have been previously quantified in practice [14, 18, 126], there has not been a formal point process model to describe them. In this section, we provide a mathematically rigorous point process model, and describe its properties.

In auditory cortex, it has been observed that some neurons only fire a single action potential with precise delay after the stimulation [27]. The concept of “binary spiking” was introduced in [28] quantifying the probability of firing and the Fano factor. We extend the binary action potential model as a unit that comprises a general precisely timed spike train where there are more than one of these PTAPs, therefore we represent PTST as a collection of PTAPs. Our model is autonomous, i.e., it does not have an input unlike the generalized linear models [59, 87] which also accurately models the spike train response given the full input trace to the system.

To obtain a model with manageable complexity, we assume that the neuronal system is stationary between trials. Changes in the system such as learning and memory due to either the stimulation or spontaneous process would be considered cross-trial non-stationary and are not considered in the present framework. This simplifying assumption is similar to independence of action potentials in a Poisson process or independence of intervals in a renewal process.

Statistical model to describe the variability of PTST is presented in two steps; first the model for a single precisely timed action potential is described, then point as a superposition of process precisely timed action potentials and additive action potentials as well as rate modulated action potentials.

3.2.1 Precisely Timed Action Potential Model

We use the random counting measure to describe the precisely timed action potential. 

**Definition 2.** The counting process $N(t)$ of the precisely timed action potential (PTAP) can be described by two independent random variables $R$ and $T$. $R$ is a Bernoulli random variable, i.e. $P(R = 1) = p$ and $P(R = 0) = 1 - p$. $T$ is an arbitrary non-negative real-valued random variable with finite mean and corresponding cumulative distribution
\( F(t) \). The counting process is defined as,

\[
N(t) = R \cdot 1_{T \leq t}
\]  

(3-4)

where \( 1_{(\cdot)} \) is the indicator function.

\( R \) represents the reliability, if it takes the value 0, the corresponding event does not occur.\(^2\) \( T \) represents the actual occurrence pattern of the event which is the jitter distribution of the action potential (Fig. 3-3). Note that \( T \) is completely ignored when \( R \) takes 0.

Note that definition 2 describes the full joint distribution for a finite collection of random variables \( \{N(A_i)\} \) for any disjoint intervals \( \{A_i\} \), where \( N(A) \) denotes the counting random variable for number of events in \( A \). Thus it defines a random process. \( N(t) \) only takes value of either 0 or 1, with the following probability,

\[
\Pr[N(t) = 1] = pF(t)
\]

\[
\Pr[N(t) = 0] = 1 - \Pr[N(t) = 1]
\]  

(3-5)

Clearly, this point process is highly temporally correlated. Once a point occurs at some time, there cannot be another point anywhere else. In other words, the conditional probability \( \Pr[N(A) = 1|N(B) = 1] = 0 \) whenever \( A \cap B = \emptyset \).

Temporal correlation and reliability is widely measured with either Fano factor, coefficient of variation, or less often with peri-stimulus time variance in the literature [75, 99]. Since these quantities depend on the window size when temporal correlation is present, we use a windowed version of Fano factor.

---

\(^2\) In the context of survival analysis, this is similar to the censoring random variable.
Definition 3 (Windowed Fano Factor). The windowed Fano factor of a counting process \( N(t) \) for an interval \( A \) is defined as,

\[
\text{FF}(N, A) = \frac{\text{var}(N(A))}{\mathbb{E}[N(A)]}
\]

(3–6)

Note that \( \text{FF}(N, [0, \infty)) \) is the Fano factor of the entire process.

It is well known that the Fano factor of any interval is unity for Poisson process [23]. If the Fano factor is smaller than 1, we use the term sub-Poisson for the corresponding counting random variable.

For a PTAP and an interval \( A = [a, b) \), let \( q = p(F(b) - F(a)) \), then

\[
\text{FF}(N, A) = \frac{\text{var}(N(A))}{\mathbb{E}[N(A)]} = \frac{q(1 - q)}{q} = 1 - q \leq 1
\]

Thus, \( N(A) \) behaves like a Bernoulli process with probability \( q \). Note that as the window size \((b - a)\) becomes smaller, the Fano factor approaches unity as in the case of renewal process [75].

The conditional intensity function is given by,

\[
\lambda^*(t) = \lambda(t|\mathcal{H}_t) = \lim_{\Delta \to 0} \frac{\Pr[N(t, t + \Delta) = 1|\mathcal{H}_t]}{\Delta} = pf(t) \frac{1}{1 - F(t)} \mathbf{1}_{T > t}
\]

(3–7)

where \( \mathcal{H}_t \) is the filtration (or history) at time \( t \), whenever the derivative of \( f(t) = \frac{dF(t)}{dt} \) exists. Note that the conditional intensity function is a random variable.

The marginal intensity function is given by,

\[
\lambda(t) = \mathbb{E}[\lambda(t|\mathcal{H}_t)] = \lim_{\Delta \to 0} \frac{\Pr[N(t, t + \Delta) = 1]}{\Delta} = pf(t)
\]

(3–8)

3.2.2 Precisely Timed Spike Train Model

We introduce the notion of precisely timed spike train as a superposition of independent point processes representing PTAPs and other rate modulations in addition.
Definition 4. The counting process for precisely timed spike train (PTST) is defined as,

\[ N(t) = \sum_{i=1}^{M} N_i(t) + N_\nu(t) \]  (3–9)

where \( \{N_i(t)\}_{i=1}^{M} \) is the collection of \( M \) indistinguishable and independent PTAPs with corresponding \( \{p_i, F_i(t)\}_{i=1}^{M} \), \( N_i(t) = R_i \cdot 1_{T_i \leq t} \), and \( N_\nu(t) \) is a counting process for inhomogeneous Poisson process with intensity function \( \nu(t) \).

The rate process \( N_\nu \) accounts for the additive action potentials that are assumed to be a rate response from the stimuli and/or the background activity. In figure 3-4, this is illustrated. Note that the summation in (3–9) removes the origin of individual action potentials.

Superposition of independent sub-Poisson processes (or equivalently the summation of the corresponding counting processes) results in a sub-Poisson process, hence the PTST is restricted to be sub-Poisson. In fact, the Fano factor of the PTST is given as

\[ \text{FF}(N, A) = \sum_{i=1}^{M} q_i q_i + q_\nu \leq 1 \] where \( q_i = p_i \int_A f(s)ds \) and \( q_\nu = \int_A \nu(s)ds \). It is easy to show that the marginal intensity function for the PTST is

\[ \lambda(t) = \sum_{i=1}^{M} p_i \frac{dF_i(t)}{dt} + \nu(t) = \sum_{i=1}^{M} \lambda_i(t) + \nu(t). \]  (3–10)

A likelihood measure \( \theta \) for a spike train model is a probability measure on the set of all possible spike trains including the empty spike train (\( \emptyset \)).

\[ \theta(\emptyset) = \prod_{i=1}^{M} (1 - p_i) \exp \left[ - \int \nu(t)dt \right] \]

\[ \frac{d\theta}{d\mu}(t_1, t_2, \ldots, t_N) = \sum_{\text{perm}} \prod_{k=L+1}^{M} (1 - p_{i_k}) \prod_{k=1}^{L} \lambda_{i_k}(t_{i_k}) \]

\[ \prod_{k=L+1}^{N} \nu(t_{i_k}) \exp \left[ - \int \nu(t)dt \right] \]  (3–11)

where \( N \leq M \) is the number of action potentials in a realization at corresponding times \( t_1 < t_2 < \cdots < t_N \), and the summation is over all possible assignments of \( N \)}
action potentials to $L$ PTAPs and $N - L$ rate process. Due to ambiguity created by the summation of counting processes, the likelihood measure for PTST is combinatorial in nature. Notice that for the empty spike train the measure has finite probability, and for a non-empty spike train and only the density (Radon-Nikodym derivative with respect to a Lebesgue measure) is well defined otherwise [116].

In figure 3-5, we fit the real data from neuronal culture with the proposed PTST model. Both the mean and variance fits well, and the generated spike trains are almost indistinguishable from the real data.
The proposed PTST model is designed from observations to capture the trial-to-trial variability of precisely timed spike trains. It provides an intuitive and natural point process model for understanding the noise that the neural system has to overcome if the precise timing information would be used optimally. Knowing the noise structure is critical for analyzing the temporal code [123], building an optimal neural decoder (e.g. classifier) [30, 118] and deriving synaptic learning rules to enhance neural coding [11]. It also restricts the amount of information that can be encoded in such a manner by setting the resolution for distinguishable spike patterns. The space of all possible spike trains that encode information has finite volume by the biological constraints such as firing rate and intrinsic noise sources. The modeled variability gives clue about how much of that space is available for coding. Hence, the current work is a complementary approach to understanding the neural code based on precisely timed spike trains.

Figure 3-5. Precisely timed spike train modeling of neuronal culture data. Data and model generated spike trains, and peri-stimulus mean and variance histogram (3 ms sliding window).
In this section, we only presented the model for a single spike train, however the extension to multiple spike trains is straight-forward. Using a population of precisely timed spike trains would be advantageous to decrease the uncertainty of representation, and make the overall performance better. In fact, the multiple interaction process (MIP) proposed by Kuhn and coworkers has a similar structure [64]. MIP has been invented to theoretically analyze the effect of different kinds of synchrony on the firing rate of the post-synaptic neuron. The proposed model is a biologically relevant alternative for studying statistics of neuron models.

The more widely used approach on point process modeling of spike trains is to use the renewal model [9, 120]. However, modeling the inter-spike interval (ISI) is not appropriate for precisely timed sequence of events. Since the precisely timed events often originate from external world (e.g. patterns in sensory input) the modeling the interval distribution makes less sense than modeling the timings measured by external reference time. As demonstrated by Mainen and Sejnowski [69] and other in vivo visual stimulation studies [e.g., 102], stimulation with features distributed over time (as in frozen noise or natural stimulation) is the origin of PTAPs. If there is no temporal structure, as in the case of constant current injection, the timing error accumulates and reduces the temporal correlation of APs among realizations and increases variability across trials despite the highly repeatable input [69].

The simplifying assumption in the proposed model is the independence among PTAPs. Several data sets showed in their count statistics that some action potentials tend to fire together or fire exclusively (data not shown). This correlated behavior is not captured with our model, but straight-forward to include in the model.

### 3.3 Point Process Divergence and Dissimilarity

Divergence is a bivariate function of probability measures $d(P, Q) : \mathcal{M} \times \mathcal{M} \to \mathbb{R}^+$ that evaluates how close two probability measures are, and takes the value of zero if and only if $P$ and $Q$ are identical. In statistics, a family of divergences called φ-divergence (or
f-divergence) is widely used [67].

\[ d_\phi(P, Q) = \int \phi\left(\frac{dQ}{dP}\right) dP, \]  

(3–12)

where \( \phi \) is a convex function and \( \phi(1) = 0 \). \( \phi \)-divergences include other well known divergences such as Kullback-Leibler, and Hellinger divergences. Divergences are closely related to convex functions and information theory, e.g., Kullback-Leibler divergence between the product of marginal probabilities and the joint probability is the mutual information [22]. Hypothesis testing using divergences result in a distribution-free test, meaning that the test does not assume underlying distribution of the samples, but any distribution can be plugged-in.

In neuroscience, the conventionally Wilcoxon test, t-test, and ANOVA are used in discriminating the change in the number of action potentials, time to first spike, and inter-spike interval (e.g. [63]). The statistics used in these tests are powerful, yet are not divergences – for example given the distribution of number of action potentials, there could be many point processes that generate value 0 for such statistics; such statistics are known as dissimilarities [96]. Therefore, the test derived from them does not necessarily guarantee discrimination of the underlying probability laws.

Recently proposed distance measure for point process by Paiva and coworkers [84], is also only a dissimilarity, but not a divergence. The dissimilarity statistic for general point processes is defined from their (marginal) intensity functions \( \lambda_P(t), \lambda_Q(t) \) which can readily be estimated from data.

\[ d_{L2}(P, Q)^2 = \int (\lambda_P(t) - \lambda_Q(t))^2 dt \]  

(3–13)

Note that this is a comparison in the first order statistic only. We will compare these dissimilarities with the proposed methods in the experiments.

A generalization of (3–13) suggests the use of square-integrable conditional intensity functions in place of marginal intensity function \( \int (\lambda_P^*(t) - \lambda_Q^*(t))^2 dt \). However, this
quantity is a random variable, hence an expectation is necessary to get a real valued function.

\[
d^*_L(P, Q)^2 = \mathbb{E} \left[ \int (\lambda^*_P(t) - \lambda^*_Q(t))^2 \, dt \right] \tag{3-14}
\]

Although (3-14) is a divergence, it is difficult to estimate this quantity in practice unless a parametric model is used.
For distribution-free hypothesis testing, one of the most widely used tests are standard tests on the cumulative distribution functions (CDF) such as Kolmogorov–Smirnov (K-S) test and Cramér–von-Mises (C-M) test. The main advantage is that the CDF can be easily estimated consistently using empirical CDF without any free parameter. The empirical distribution function for samples \( \{x_1, \ldots, x_n\} \) is defined as,

\[
F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{x_i \geq x}.
\]  

The convergence is at least almost sure by the law of large numbers, however it can be shown to be stronger:

**Theorem 1** (Glivenko–Cantelli). \( F_n \) converges to \( F \) uniformly almost surely.

Two sided Kolmogorov–Smirnov test statistic uses the \( L_\infty \) metric of the difference between empirical distributions: \( KS = \sup_x |F_n(x) - G_n(x)| \), while the Cramér–von-Mises test statistic uses the \( L_2(\mathbb{R}, F + G) \) squared distance: \( CM = \int (F_n(x) - G_n(x))^2 d(F + G) \).

These statistics will significantly deviate from zero when \( F \neq G \). K-S test only measures the maximum deviation, and the C-M test measures square sum of deviations, however one does not always performs better than the other.

The concept of CDF is only well defined for real random variables where a full ordering is provided. In the space of spike trains, there is no naturally provided ordering. However within each stratum it is possible to have a multi-dimensional CDF. The stratified space approach of representing point processes allows these traditional statistical tests to be implemented. Therefore the extension of K-S and C-M test to the point processes requires careful application of stratification.

---

\(^1\) The level of the stratified model is developed in chapter 3
4.2 Extended Kolmogorov–Smirnov Divergence

A Kolmogorov-Smirnov (K-S) type divergence between $P$ and $Q$ can be derived from the $L_1$ distance between the probability measures, following the equivalent representation,

$$d_1(P, Q) = \int \Omega |P - Q| \geq \sup_{A \in \mathcal{F}} |P(A) - Q(A)|. \quad (4-2)$$

Since (4–2) is difficult and perhaps impossible to estimate directly without a model, our strategy is to use the stratified spaces $(\Omega_0, \Omega_1, \ldots)$ defined in section 3.1.3, and take the supremum only in the corresponding conditioned probability measures to approximate the lower bound. Let $\mathcal{F}_i = \mathcal{F} \cap \Omega_i := \{ F \cap \Omega_i | F \in \mathcal{F} \}$. Since $\cup_i \mathcal{F}_i \subset \mathcal{F}$,

$$d_1(P, Q) \geq \sum_{n \in \mathbb{N}} \sup_{A \in \mathcal{F}_n} |P(A) - Q(A)| = \sum_{n \in \mathbb{N}} \sup_{A \in \mathcal{F}_n} |P(\Omega_n)P(A|\Omega_n) - Q(\Omega_n)Q(A|\Omega_n)|.$$

Since each $\Omega_n$ is a Euclidean space, we can induce the traditional K-S test statistic by further reducing the search space to $\tilde{\mathcal{F}}_n = \{ \times_{i \in \Omega}(-\infty, t_i] | t = (t_1, \ldots, t_n) \in \mathbb{R}^n \}$. This results in the following inequality,

$$\sup_{A \in \tilde{\mathcal{F}}_n} |P(A) - Q(A)| \geq \sup_{A \in \tilde{\mathcal{F}}_n} |P(A) - Q(A)| = \sup_{t \in \mathbb{R}^n} \left| F_P^{(n)}(t) - F_Q^{(n)}(t) \right|, \quad (4-3)$$

where $F_P^{(n)}(t) = P[T_1 \leq t_1 \land \ldots \land T_n \leq t_n]$ is the cumulative distribution function (CDF) corresponding to the probability measure $P$ in $\Omega_n$. Hence, we define the K-S divergence as

$$d_{KS}(P, Q) = \sum_{n \in \mathbb{N}} \sup_{t \in \mathbb{R}^n} \left| P(\Omega_n)F_P^{(n)}(t) - Q(\Omega_n)F_Q^{(n)}(t) \right|. \quad (4-4)$$

Given a finite number of samples $X = \{x_i\}_{i=1}^{N_P}$ and $Y = \{y_j\}_{j=1}^{N_Q}$ from $P$ and $Q$ respectively, we have the following estimator for equation (4–4).

$$\hat{d}_{KS}(P, Q) = \sum_{n \in \mathbb{N}} \sup_{t \in \mathbb{R}^n} \left| \hat{P}(\Omega_n)\hat{F}_P^{(n)}(t) - \hat{Q}(\Omega_n)\hat{F}_Q^{(n)}(t) \right| = \sum_{n \in \mathbb{N}} \sup_{t \in X_n \cup Y_n} \left| \hat{P}(\Omega_n)\hat{F}_P^{(n)}(t) - \hat{Q}(\Omega_n)\hat{F}_Q^{(n)}(t) \right|. \quad (4-5)$$
where \( X_n = X \cap \Omega_n \), and \( \hat{P} \) and \( \hat{F}_P \) are the empirical probability and empirical CDF, respectively. Notice that we only search the supremum over the locations of the realizations \( X_n \cup Y_n \) and not the whole \( \mathbb{R}^n \), since the empirical CDF difference \( \left| \hat{P}(\Omega_n)\hat{F}^{(n)}_P(t) - \hat{Q}(\Omega_n)\hat{F}^{(n)}_Q(t) \right| \) only changes values at those locations.

**Theorem 2** \((d_{KS} \text{ is a divergence})\).

\[
d_1(P, Q) \leq d_{KS}(P, Q) \leq 0 \tag{4-6}
\]

\[
d_{KS}(P, Q) = 0 \iff P = Q \tag{4-7}
\]

**Proof.** The first property and the \( \iff \) proof for the second property are trivial. From the definition of \( d_{KS} \) and properties of CDF, \( d_{KS}(P, Q) = 0 \) implies that \( P(\Omega_n) = Q(\Omega_n) \) and \( F_P^{(n)} = F_Q^{(n)} \) for all \( n \in \mathbb{N} \). Given probability measures for each \( (\Omega_n, F_n) \) denoted as \( P_n \) and \( Q_n \), there exist corresponding unique extended measures \( P \) and \( Q \) for \( (\Omega, F) \) such that their restrictions to \( (\Omega_n, F_n) \) coincide with \( P_n \) and \( Q_n \), hence \( P = Q \).

**Theorem 3** \((\text{Consistency of K-S divergence estimator}) \). *As the sample size approaches infinity,*

\[
\left| d_{KS} - \hat{d}_{KS} \right| \xrightarrow{a.u.} 0 \tag{4-8}
\]

**Proof.** Note that \( \sum \sup \cdot - \sum \sup \cdot \leq \sum |\sup \cdot - \sup \cdot| \). Due to the triangle inequality of the supremum norm,

\[
\left| \sup_{t \in \mathbb{R}^n} \left| P(\Omega_n)F_P^{(n)}(t) - Q(\Omega_n)F_Q^{(n)}(t) \right| - \sup_{t \in \mathbb{R}^n} \left| \hat{P}(\Omega_n)\hat{F}^{(n)}_P(t) - \hat{Q}(\Omega_n)\hat{F}^{(n)}_Q(t) \right| \right| \leq \sup_{t \in \mathbb{R}^n} \left| P(\Omega_n)F_P^{(n)}(t) - Q(\Omega_n)F_Q^{(n)}(t) \right| - \left| \hat{P}(\Omega_n)\hat{F}^{(n)}_P(t) - \hat{Q}(\Omega_n)\hat{F}^{(n)}_Q(t) \right|.
\]
Again, using the triangle inequality we can show the following:

\[
\left| P(\Omega_n) F_P^{(n)}(t) - Q(\Omega_n) F_Q^{(n)}(t) \right| - \left| P(\Omega_n) \hat{F}_P^{(n)}(t) - Q(\Omega_n) \hat{F}_Q^{(n)}(t) \right|
\leq \left| P(\Omega_n) F_P^{(n)}(t) - Q(\Omega_n) F_Q^{(n)}(t) - \hat{P}(\Omega_n) \hat{F}_P^{(n)}(t) + \hat{Q}(\Omega_n) \hat{F}_Q^{(n)}(t) \right|

= \left| P(\Omega_n) F_P^{(n)}(t) - Q(\Omega_n) F_Q^{(n)}(t) - Q(\Omega_n) \hat{F}_Q^{(n)}(t) \right|
+ P(\Omega_n) \left| \hat{F}_P^{(n)}(t) - \hat{F}_P^{(n)}(t) \right| + Q(\Omega_n) \left| F_Q^{(n)}(t) - \hat{F}_Q^{(n)}(t) \right|
+ \hat{F}_P^{(n)}(t) \left| P(\Omega_n) - \hat{P}(\Omega_n) \right| + \hat{F}_Q^{(n)}(t) \left| Q(\Omega_n) - \hat{Q}(\Omega_n) \right|.
\]

Then the theorem follows from the Glivenko-Cantelli theorem, and \( \hat{P}, \hat{Q} \xrightarrow{a.s.} P, Q \). \( \square \)

### 4.3 Extended Cramér–von-Mises Divergence

We can extend equation (4.4) to derive a Cramér–von-Mises (C-M) type divergence for point processes. Let \( \mu = \frac{P + Q}{2} \), then \( P, Q \) are absolutely continuous with respect to \( \mu \). Note that, \( F_P^{(n)}, F_Q^{(n)} \in L_2(\Omega_n, \mu|_n) \) where \( \mu|_n \) denotes the restriction on \( \Omega_n \), i.e. the CDFs are \( L_2 \) integrable, since they are bounded. Analogous to the relation between K-S test and C-M test, we would like to use the integrated squared deviation statistics in place of the maximal deviation statistic. By integrating over the probability measure \( \mu \) instead of the supremum operation, and using \( L_2 \) instead of \( L_\infty \) distance in (4.4), we define

\[
d_{CM}(P, Q) = \sum_{n \in \mathbb{N}} \int_{\mathbb{R}^n} \left( P(\Omega_n) F_P^{(n)}(t) - Q(\Omega_n) F_Q^{(n)}(t) \right)^2 d\mu|_n(t). \tag{4.9}
\]

This can be seen as a direct extension of the C-M criterion. The corresponding estimator can be derived using the strong law of large numbers,

\[
\hat{d}_{CM}(P, Q) = \sum_{n \in \mathbb{N}} \left[ \frac{1}{2} \sum_i \left( \hat{P}(\Omega_n) \hat{F}_P^{(n)}(x_i^{(n)}) - \hat{Q}(\Omega_n) \hat{F}_Q^{(n)}(x_i^{(n)}) \right)^2 \right]
+ \frac{1}{2} \sum_i \left( \hat{P}(\Omega_n) \hat{F}_P^{(n)}(y_i^{(n)}) - \hat{Q}(\Omega_n) \hat{F}_Q^{(n)}(y_i^{(n)}) \right)^2. \tag{4.10}
\]
Theorem 4 (\(d_{CM}\) is a divergence). For \(P\) and \(Q\) with square integrable CDFs,
\[
d_{CM}(P, Q) \geq 0 \tag{4–11}
\]
\[
d_{CM}(P, Q) = 0 \iff P = Q. \tag{4–12}
\]

**Proof.** Similar to theorem 2. \(\square\)

Theorem 5 (Consistency of C-M divergence estimator). As the sample size approaches infinity,
\[
\left| d_{CM} - \hat{d}_{CM} \right| \xrightarrow{a.u.} 0 \tag{4–13}
\]

**Proof.** Similar to (4–8), we find an upper bound and show that the bound uniformly converges to zero. To simplify the notation, we define \(g_n(x) = P(\Omega_n)F_P^{(n)}(x) - Q(\Omega_n)F_Q^{(n)}(x)\), and \(\hat{g}_n(x) = \hat{P}(\Omega_n)\hat{F}_P^{(n)}(x) - \hat{Q}(\Omega_n)\hat{F}_Q^{(n)}(x)\). Note that \(\hat{g}_n \xrightarrow{a.u.} g\) by the Glivenko-Cantelli theorem and \(\hat{P} \xrightarrow{a.s.} P\) by the strong law of large numbers.

\[
\left| d_{CM} - \hat{d}_{CM} \right| = \frac{1}{2} \sum_{n \in \mathbb{N}} \int g_n^2 dP|_n + \sum_{n \in \mathbb{N}} \int g_n^2 dQ|_n - \sum_{n \in \mathbb{N}} \sum_{i} \hat{g}_n(x_i) - \sum_{n \in \mathbb{N}} \sum_{i} \hat{g}_n(y_i)
\]
\[
= \sum_{n \in \mathbb{N}} \left[ \int g_n^2 dP|_n - \int \hat{g}_n^2 d\hat{P}|_n + \int g_n^2 dQ|_n - \int \hat{g}_n^2 d\hat{Q}|_n \right]
\]
\[
\leq \sum_{n \in \mathbb{N}} \left[ \left| \int g_n^2 dP|_n - \int \hat{g}_n^2 d\hat{P}|_n \right| + \left| \int g_n^2 dQ|_n - \int \hat{g}_n^2 d\hat{Q}|_n \right| \right]
\]

where \(\hat{P} = \sum_i \delta(x_i)\) and \(\hat{Q} = \sum_i \delta(y_i)\) are the corresponding empirical measures. Without loss of generality, we only find the bound on \(\left| \int g_n^2 dP|_n - \int \hat{g}_n^2 d\hat{P}|_n \right|\), then the rest is bounded similarly for \(Q\).

\[
\left| \int g_n^2 dP|_n - \int \hat{g}_n^2 d\hat{P}|_n \right| = \left| \int g_n^2 dP|_n - \int \hat{g}_n^2 d\hat{P}|_n + \int \hat{g}_n^2 dP|_n - \int \hat{g}_n^2 d\hat{P}|_n \right|
\]
\[
\leq \left| \int (g_n^2 - \hat{g}_n^2) dP|_n \right| + \left| \int \hat{g}_n^2 d(P|_n - \hat{P}|_n) \right|
\]
Applying Glivenko-Cantelli theorem and strong law of large numbers, these two terms converges since $\hat{g}_n^2$ is bounded. Hence, we show that the C-M test estimator is consistent.

4.4 Simulation Results

We present a set of two-sample problems and apply various statistics to perform hypothesis testing. As a baseline measure, the widely used Wilcoxon rank-sum test (or equivalently, the Mann-Whitney U test) is performed on the count distribution (e.g. [60]); this is a non-parametric median test for the total number of action potentials. In addition, difference in the rate function, i.e., 1-dimensional first-order summarizing statistic of the point process, is measured with an integrated squared deviation statistics $\lambda_{L2} = \int (\lambda_1(t) - \lambda_2(t))^2 \, dt$, where $\lambda(t)$ is estimated by smoothing spike timing with a Gaussian kernel, evaluated at a uniform grid at least an order of magnitude smaller than the standard deviation of the kernel. We report the performance of the test with varying kernel sizes.

All tests are quantified by the power of the test given a significance threshold (type-I error) at 0.05. The null hypothesis distribution is empirically computed by either generating independent samples or by permuting the data to create at least 1000 values.

4.4.1 Poisson Process

Poisson process is the simplest point process model widely used as a basic stochastic neuron model. We test if the methods can detect difference in the rate profile while maintaining the average rate constant. In Poisson process, the count distribution has variance equal to the mean rate, hence for higher rates the data becomes sparsely spread among the dimensions $\Omega_n$. For this example of rate 3, 90% of the count distribution is concentrated in the range of one to seven spikes.

Since the rate function fully describes this process, the performance of the rate based statistic $\lambda_{L2}$ works the best in this case (see Figure 4-1). Also because the count distribution for $H_0$ and $H_1$ are identically Poisson distributed, Wilcoxon test (denoted $N$)
Figure 4-1. Statistical power of cumulative based divergences on Poisson processes. (Left) Spike trains from the null and alternate hypothesis. The rate function is constant during each 100 ms interval. The rate of $H_0$ changes from 20 to 10 spk/s, and for $H_1$ it changes from 10 to 20 spk/s. (Right) Comparison of the power of each method. The Wilcoxon test on mean count (labeled with $N$) stays around the threshold level (0.05). All other methods are empirically consistent for this example. The error bars are standard deviation over 10 Monte Carlo runs.

fails to detect the difference. K-S and C-M based divergences performs similarly, yet it is interesting to note that C-M is consistently better than the K-S divergence.

4.4.2 Stationary Renewal Processes

Renewal process is a widely used point process model that compensates the deviation from Poisson process [75]. Stationary renewal process with gamma interval distribution is simulated. Mean rate are the same, therefore rate function statistic and Wilcoxon test does not yield consistent result, while the proposed measures obtained high power with a small number of samples. The C-M test is more powerful than K-S in this case; this can be interpreted by the fact that the difference in the cumulative is not concentrated but spread out over time because of the stationarity.
Figure 4-2. Gamma distributed renewal process with shape parameter $\theta = 3$ ($H_0$) and $\theta = 0.5$ ($H_1$). The mean number of action potential is fixed to 10. (Left) Spike trains from the null and alternate hypothesis. (Right) Comparison of the power of each method. The error bars are standard deviation over 20 Monte Carlo runs.

Figure 4-3. Precisely timed spike train model ($H_0$) versus equi-intensity Poisson process ($H_1$). Spike trains from the null and alternate hypothesis for $L = 4$. 
Figure 4-4. Comparison of the power of each method for \( L = 1, 2, 3, 4 \) on precisely timed spike train model \((H_0)\) versus equi-intensity Poisson process \((H_1)\). See Figure 4-3 for example spike trains. (Left) Power comparison for methods except for \( N \). The rate statistic \( \lambda_{L,2} \) are not labeled, since they are not able to detect the difference. (Right) Wilcoxon test on the number of action potentials. The error bars are standard deviation over 10 Monte Carlo runs.

4.4.3 Precisely Timed Spike Trains

As described in section 3.2, a precisely timed spike train in an interval is modeled by \( L \) number of probability density and probability pairs \( \{(f_i(t), p_i)\}_{i=1}^{L} \). Each \( f_i(t) \) corresponds to the temporal jitter, and \( p_i \) corresponds to the probability of generating the spike. Each realization of the PTST model produces at most \( L \) spikes. The equi-intensity Poisson process has the rate function \( \lambda(t) = \sum_i p_i f_i(t) \). We test if the methods can differentiate between the PTST \((H_0)\) and equi-intensity Poisson process \((H_1)\) for \( L = 1, 2, 3, 4 \) (see Figure 4-3). Note that \( L \) determines the maximum dimension for the PTST. \( f_i(t) \) were equal variance Gaussian distribution on a grid sampled from a uniform random variable, and \( p_i = 0.9 \).

As shown in Figure 4-4, only the proposed methods perform well. Since the rate function profile is identical for both models, the rate function statistic \( \lambda_{L,2} \) fails to differentiate. The Wilcoxon test does work for intermediate dimensions, however its performance is highly variable and unpredictable. This is because the assumption that the
count distributions come from a location family is violated in this example. In contrast to the example in section 4.4.1, the K-S test is consistently better than the C-M statistic in this problem.

4.4.4 Neuron Model

Figure 4-5. Izhikevich neuron model. (Left) example spike trains from the model for the baseline current (top) and with increased gain (middle). The bottom trace is the waveform of the injected current. (Right) Comparison of the power of each method for different input scaling. TTFS is the K-S statistic for the jitter distribution of the time to first spike.

We investigate the sensitivity of the proposed methods in a neurophysiological scenario. We simulated a repeated injection of current to a neuron model, and observed the output spike train pattern. By varying the gain factor of input, we investigate the sensitivity of various statistics to a realistic statistical change. Izhikevich’s simplified neuron model for Class 2 neuron is stimulated with noisy current injection [48, Ch 8]. The dynamics of the model is fully described by the following equations,

\[ \dot{v} = 0.04(v + 42)(v + 82) \quad \text{if } v \geq 30, \text{ then} \]

\[ \dot{u} = 0.2(0.26v - u) \quad v \leftarrow -65 \]
where spike times are recorded whenever \( v \) is reset. Frozen noise is generated from white Gaussian wave form with duration 40 ms. The noise is injected to the neuron model with additional noise current that is generated for each trial. The neuron model is sensitive to certain feature of the noise [69], hence precisely timed action potentials are observed (see Figure 4-5). When the frozen noise was scaled up, the action potentials generally reduced their variance and fired in earlier time on average (see Figure 4-5).

The input current is adjusted such that 2 action potentials are generated on average. The sample size is 30 spike trains for each condition. The C-M and K-S measures both perform at least as good as the best rate based measure. Additionally, a K-S test on the jitter distribution to the first event (time to first spike), ignoring the second spike timing, is compared and shown to be slightly worse than the proposed method.

### 4.4.5 Serial Correlation Model

![Figure 4-6. Renewal process with \((H_0)\) or without serial correlation \((H_1)\). (Left) Spike trains from the null and alternate hypothesis. (Right) Comparison of the power of each method. The error bars are standard deviation over 10 Monte Carlo runs.](image)

Serial correlation in point process is defined by the autocovariance of the intervals (time between spikes). Due to internal dynamics of neurons, the spike trains can have
serial correlation [33]. We simulated a renewal process without serial correlation and a point process with same marginal interval distribution but with a non-zero serial correlation [15]. We restrict our problem to two intervals only, where each interval distribution is a convolution of two uniform distributions.

It turns out this problem is quite difficult and hundreds of samples are needed to discriminate them (see Figure 4-6). Both K-S and C-M type tests perform equally well, while the Wilcoxon test fails to differentiate between the two point processes. An optimized kernel size for the rate function statistic outperforms both K-S and C-M type test.

4.5 Optimal Stimulation Parameter Selection

4.5.1 Problem

In the context of brain machine interfaces, one of the challenges is to create a sensory neural prosthetics; a device that will artificially generate perception of appropriate sensation. One way of achieving this is through electric stimulation of the sensory pathway such as thalamus. If one agrees to the notion that similar sensory cortex activation leads to similar sensory perception, the problem is to optimize stimulation for maximum similarity. Thus, a principled method for measuring similarity is essential. Since we may not be able to find the exact parameter that generates the target response pattern, the similarity measure needs to well behave for a wide range providing distinct similarity values.

An experimental data that consists of multi-trial measurement for sensory cortex response resulting from natural touch stimuli of the rat paw is used. The natural stimulation serves as the target point process, while a fixed number of parameterized microstimuli responses create the search space.

4.5.2 Method

The data used in this analysis is from a single rat implanted with multi-site electrodes. The anesthetized rat’s neuronal activity was recorded from 16 cortical
electrodes in S1 and 16 thalamic electrodes in ventral posterolateral nucleus (VPL) area. The rat was anesthetized by isofluorane followed by a nembutal injection and maintained with isofluorane. Neurons with significant natural response were identified by natural touch stimulation via thwacking of the first digit on the rat’s right paw at 2 Hz for 120 seconds, yielding 240 thwack responses. The thalamic microstimulation was performed on two electrodes, chosen for their significant response to natural touch (based on the audible quality of the spiking). The stimulations consisted of single biphasic pulses with specified pulse duration and amplitude, but always at 2 Hz frequency. During the trial 19 distinct pairs of pulse duration and amplitude were applied, with 140 responses from each pair randomly permuted throughout the recording. Thus, the goal was to compare the 240 natural touch responses to the 140 responses for each pair.

In order to compare the temporal response, the natural touch and microstimulation needed to be aligned. A difficulty with microstimulation is that it produces artifact that

---

**Figure 4-7.** Spike trains from sensory cortex in response to natural and electrical stimulus. Each box shows the raster of 140 spike trains (except for the natural which has 240) of duration 40 ms. The spike trains are sorted by the number of spikes and time to first spike.
prevents recording for nearly 6 ms on the entire amplifier array. The corresponding initial 6 ms response to natural touch was removed. A window of 40 ms was used for comparison. The firing rate returned to the spontaneous level by the end of the window as in figure 4-9.

![Figure 4-8. Dissimilarity/divergences across parameter sets. Each measure is shifted and scaled to be in the range of 0 to 1. $\lambda_{L2}$ used 2.5 ms bins with no smoothing.](image)

![Figure 4-9. The natural response (left), the microstimulation set selected by $\lambda_{L2}$ (center), and the set selected by K-S and C-M (right). Top row shows the spike trains stratified into number of spikes and then spike time. Bottom row shows the average response PSTH binned at 2.5 ms; the windowed variance is shown as a thin green line.](image)

As an application of the proposed measure, we demonstrate how they can be used as a cost function to be minimized. The problem is to find the optimal electrical stimulation...
parameters that would produce spiking pattern in the sensory cortex that is closest to the target pattern generated from natural sensory input. Given 240 responses from natural input, and around 140 responses from each stimulation parameter (ranging over duration and amplitude of a biphasic current injection), we use the proposed divergences to measure the most similar response pattern, taking the variability structure over trials into account.

The results from applying the C-M, K-S, and $\lambda_{L2}$ measures between the set of natural stimuli responses and each parameter set are shown Figure 4-8. The overall trend among dissimilarity/divergences is consistent, but the location of the minima does not coincide for $\lambda_{L2}$. The natural stimuli and the optimal responses are shown in Figure 4-9. The set selected by the proposed methods match the early timing pattern better. In fact, the deviation of windowed variance from the mean supports the non-Poisson nature of the response, probably due to the time locked first action potential.

4.6 Conclusion

We have proposed two novel measures of divergence between point processes based on cumulative distribution functions. The proposed measures have been derived from the basic probability law of a point process and we have shown that these measures can be efficiently estimated consistently from data. Using divergences for statistical inference let us break free from using only first and second order statistics, and enables distribution-free spike train analysis.

The time complexity of both methods is $O(\sum_n n [N_P(n)N_Q(n) + N_P^2(n) + N_Q^2(n)])$ where $N_P(n)$ is the number of spike trains from $P$ that has $n$ spikes. In practice this is often faster than the binned rate function estimation which has time complexity $O(BN)$ where $B$ is the number of bins and $N = \sum_n n(N_P(n) + N_Q(n))$ is the total number of spikes in all the samples. In several examples demonstrated here, the $L_2$ distance of rate function statistic outperformed our proposed method. However, it involves the search for the smoothing kernel size and bin size which can make the process slow and prohibitive.
In addition, it brings the danger of multiple testing, since some smoothing kernel sizes may pickup spurious patterns that are only fluctuations due to finite samples size.
CHAPTER 5
ϕ-DIVERGENCE AND HILBERTIAN METRIC

Recall the definition of ϕ-divergence (3–12)

\[ D_\phi(P, Q) = \int \phi \left( \frac{dQ}{dP} \right) dP, \] (5–1)

The ϕ-divergence family encompasses several well known divergences such as Kullback-Leibler divergence, Hellinger distance, total variation, and \( \chi^2 \)-divergence, that are widely applied in estimation theory, decision theory, formulation of classification bounds, and information theory [67]. While ϕ-divergences are not metric in general, a subclass of ϕ-divergences are squared Hilbertian metrics on the space of probability laws. We investigate two special cases in the intersection: Hellinger divergence and symmetric \( \chi^2 \)-divergence. We extend these to point processes and present estimators via Radon-Nikodym derivatives. We present two distinct methods for estimating the Radon-Nikodym derivative; via a direct density estimation, or via approximating the derivative with kernel methods. In addition, two datasets are analyzed for change detection, and clustering applications.

5.1 Hilbertian Metric

Statistical divergence is not a metric in general; the symmetry nor the triangle inequality is satisfied. Symmetrization can be achieved by defining \( D(Q, P) + D(P, Q) \), however satisfying the triangle inequality is not trivial. Since a lot of algorithms can operate in a metric space, it is desirable to study the divergences that have the metric properties; hence working in the metric space of probability measures.

Moreover, as we will see a family of divergences known as Hilbertian metric provides a distance on the probability space that can be isometrically embedded in a Hilbert space. These divergences are closely related to kernel methods; in fact there exists a family of corresponding symmetric positive definite kernels on probability measures. Therefore, we can provide an inner product structure in addition to the distance through these divergences.
A metric $d$ is a Hilbertian metric if and only if $-d^2(x,y)$ is a conditionally positive definite kernel\(^1\) (cpd) [7, 108]. Since probability and their density take positive real values, one can induce a metric between probability measures from a distance on positive reals. In order to ensure the distance is independent of the reference measure, which is crucial for its evaluation, the distance must be a $\frac{1}{2}$-homogeneous Hilbertian metric.

**Definition 5** (α-homogeneous function [43]). A symmetric function $f$ is α-homogeneous if for all $c \in \mathbb{R}^+$,

$$f(cx, cy) = c^\alpha f(x, y). \quad (5-2)$$

Hein and Bousquet defined a family of such metrics that leads to Hilbertian metrics on probability measures.

**Theorem 6** (Fuglede [36], Hein and Bousquet [43]). The function $d_{\alpha|\beta} : \mathbb{R}^+ \times \mathbb{R}^+ \rightarrow \mathbb{R}$ defined as,

$$d_{\alpha|\beta}^2(x, y) = \frac{2^{\frac{1}{\beta}}(x^\alpha + y^\alpha) - 2^{\frac{1}{\alpha}}(x^\beta + y^\beta)}{2^{\frac{1}{\alpha}} - 2^{\frac{1}{\beta}}} \quad (5-3)$$

is a $\frac{1}{2}$-homogeneous Hilbertian metric on $\mathbb{R}^+$, whenever $\alpha \in [1, \infty], \beta \in [\frac{1}{2}, \alpha]$ or $\beta \in [-\infty, -1]$.

**Proposition 7** (Hein and Bousquet [43]). Let $P$ and $Q$ be two probability measures on $X$, and $\mu$ be a measure that dominates both probability measures. Let $d_{\mathbb{R}^+}$ be a $\frac{1}{2}$-homogeneous Hilbertian metric on $\mathbb{R}^+$. Then $d_{\mathcal{M}}$ defined as,

$$d_{\mathcal{M}}^2(P, Q) = \int_X d_{\mathbb{R}^+}^2 \left( \frac{dP}{d\mu}, \frac{dQ}{d\mu} \right) d\mu \quad (5-4)$$

is a Hilbertian metric on $\mathcal{M}$. $d_{\mathcal{M}}^2$ is independent of the dominating measure $\mu$.

---

\(^1\) If $\sum_i \sum_j c_i c_j K(x_i, x_j) \geq 0$ for all $\sum_i c_i = 0$, then it is cpd.
Proof. Since \( d^2_{\mathbb{R}^+} \) is a 1-homogeneous function, the definition is invariant to the choice of dominating measure;

\[
d^2_{\mathcal{M}}(P, Q) = \int_X d^2_{\mathbb{R}^+} \left( \frac{dP}{d\mu}, \frac{dQ}{d\mu} \right) d\mu = \int_X \frac{d\mu}{d\nu} d^2_{\mathbb{R}^+} \left( \frac{dP}{d\mu}, \frac{dQ}{d\mu} \right) d\nu
\]

by the chain rule [42] where \( \mu \ll \nu \). By the linearity of the integration, the cpd property of \(-d^2_{\mathbb{R}^+}\) transmits to \(-d^2_{\mathcal{M}}\), hence it is a Hilbertian metric. \(\square\)

Schölkopf showed the connection between the cpd kernels and their behavior as a distance based feature space [108].

**Proposition 8** (cpd and pd kernel Schölkopf [108]). Let \( \kappa : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R} \) be a symmetric kernel. Then, \( \tilde{\kappa} \) defined as follows,

\[
\tilde{\kappa}(x, y) = \frac{1}{2} (\kappa(x, x) - \kappa(z, x) - \kappa(x, z) + \kappa(z, z))
\]

(5–5)

is pd if and only if \( \kappa \) is cpd.

Now it is possible to build pd kernels from cpd kernels. Note that from (5–5) we can induce an RKHS based on \( \tilde{\kappa} \). Let \( \phi(x) \) denote the projected vector corresponding to \( x \in \mathcal{X} \). Then, the squared distance in the RKHS becomes,

\[
\|\phi(x) - \phi(y)\|^2 = \tilde{\kappa}(x, x) + \tilde{\kappa}(y, y) - 2\tilde{\kappa}(x, y)
\]

(5–6)

\[
= \frac{1}{2} (\kappa(x, x) - \kappa(z, x) - \kappa(x, z) + \kappa(z, z))
\]

\[
+ \kappa(y, y) - \kappa(z, y) - \kappa(y, z) + \kappa(z, z))
\]

\[
- (\kappa(x, y) - \kappa(z, y) - \kappa(x, z) + \kappa(z, z))
\]

\[
= -\kappa(x, y) + \frac{1}{2} (\kappa(x, x) + \kappa(y, y))
\]

(5–7)

Therefore, using the pd kernel \( \tilde{\kappa} \) and using an algorithm that only requires distances, essentially the RKHS can be linked to a cpd kernel.
In terms of Hilbertian metrics for probability measures, using (5–4) and (5–3), and taking the \( z \) in (5–5) to be the zero measure results in a class of pd kernels \( \kappa_{\alpha|\beta}(P, Q) \) [43].

\[
\kappa_{1|1}(P, Q) = \int \frac{p(x)q(x)}{p(x) + q(x)} d\mu(x), \quad \kappa_{\frac{1}{2}|1}(P, Q) = \int \sqrt{p(x)q(x)} d\mu(x) \\
\kappa_{1|1}(P, Q) = -\frac{1}{\log 2} \int p(x) \log \left( \frac{p(x)}{p(x) + q(x)} \right) + q(x) \log \left( \frac{q(x)}{p(x) + q(x)} \right) d\mu(x)
\]

where \( p = \frac{dP}{d\mu} \) and \( q = \frac{dQ}{d\mu} \). These kernels induce RKHS where the distance corresponds to using symmetric \( \chi^2 \)-measure, Hellinger metric, and Jensen-Shannon divergence.

### 5.2 Radon-Nikodym Derivative

Both \( \phi \)-divergences and Hilbertian metrics can be written in the integral form:

\[
\int f \left( \frac{dP}{dQ} \right) dR
\]

where \( P, Q \) and \( R \) are \( \sigma \)-finite measures. Here, \( \frac{dP}{dQ} \) is the Radon-Nikodym derivative of \( P \) with respect to \( Q \).

**Definition 6** (Absolute continuity). Let \( (\Omega, \mathcal{F}) \) be a measurable space, and \( P, Q \) two measures. If for all \( A \in \mathcal{F} \), \( Q(A) = 0 \) implies \( P(A) = 0 \), then \( P \) is absolutely continuous with respect to \( Q \). This is written as \( P \ll Q \).

**Theorem 9** (Radon-Nikodym derivative [42]). If \( \mu \) is a \( \sigma \)-finite measure, and \( \nu \ll \mu \), then there exists a finite valued measurable function \( f \) on \( \Omega \) such that

\[
\nu(E) = \int_E f d\mu
\]

for every measurable set \( E \). \( f \) is the Radon-Nikodym derivative, and denoted as \( \frac{d\nu}{d\mu} \). It is unique up to \( \mu \).

Radon-Nikodym derivative of probability measures \( P \) and \( Q \) is equivalent to the likelihood ratio. The question is how to estimate \( \frac{dP}{dQ} \) from finite samples from \( P \) and \( Q \).

A measure \( P \) is a diffuse measure if \( P(\{x\}) = 0 \) for all \( x \in \Omega \). If the space is equipped with a reference measure \( \mu \) such that for any diffuse probability measures \( P \) and \( Q \), \( (P, Q \ll \mu) \), then we can use the chain rule of Radon-Nikodym derivatives [42] and write...
(5–9) as,

\[
\int f \left( \frac{dP}{d\mu} / \frac{dQ}{d\mu} \right) dR
\]

Hence, given an estimator for \( \frac{dP}{d\mu} \) the divergence can be estimated.

In the case of stratified approach, we defined a reference measure derived from the Lebesgue measure of Euclidean spaces (section 3.1.3). Hence, for a point process \( P \), \( \frac{dP}{d\mu} \) can be measured via Parzen windowing using density estimation kernels on Euclidean spaces (assuming the point processes are diffusive).

The probability measure \( P \) for point process can also be decomposed according to the partition of \( \Omega \). Define \( P_n(A) = P(A \cap \Omega_n) \), then we can write \( P = \sum_{n=0}^{\infty} P_n \). Also denote \( P_n(\Omega) \) as \( p_n \), the probability of having \( n \) action potentials. \( P \) is assumed to be absolutely continuous with respect to \( \mu \) (written as \( P \ll \mu \)), thus we can take the Radon-Nikodym derivative,

\[
\frac{dP}{d\mu}(\omega) = P_0(\Omega)\delta_{\Omega_0}(\omega) + \sum_{n=1}^{\infty} \frac{dP_n}{d\mu_n}(\omega_n) = p_0\delta_{\Omega_0}(\omega) + \sum_{n=1}^{\infty} p_n f_n(\omega_n)
\]

where \( f_n \) is the unordered joint location density and is symmetric on the permutation of its arguments, and \( \omega_n = \omega \cup \Omega_n \). \( f_n \) has a close relationship with the density of Janossy measure \( j_n(\omega) = n!p_n f_n(\omega) \). For a spike train \( \omega = (t_1, \cdots, t_n) \), the Janossy measure has a simple interpretation; \( j_n(t_1, \cdots, t_n)dt_1 \cdots dt_n \) is the probability that there are exactly \( n \) events in the point process, one in each of the \( n \) distinct infinitesimal intervals \((t_i, t_i + dt_i)\) [25]. Note that if \( n \) action potentials are ordered, it should be divided by \( n! \) to represent the joint density of ordered times which has a smaller state space.

Given a sequence of observations \( X = \{\omega_i\}_{i=1}^{m} \), we can use the decomposition (5–11) for estimation of the finite point process. Let the subsequences \( X^{(n)} = \{\omega_i; \omega_i \in \Omega_n, i = 1, \ldots, m\} \) be the set of all spike trains with length \( n \). Frequency based estimate of the
total count distribution and the kernel density estimation of $f_n$ can then be written as,

$$
\hat{p}_n = \frac{\text{card}(X^{(n)})}{\text{card}(X)} \quad n = 0, 1, \quad (5-12)
$$

$$
\hat{f}_n(x) = \frac{1}{\text{card}(X^{(n)})} \sum_{\omega_i \in X^{(n)}} \kappa_n(x - \omega_i; \sigma_n) \quad n = 1, \ldots \quad (5-13)
$$

where $\kappa_n(\cdot; \sigma)$ is a symmetric $n$-dimensional density estimation kernel with bandwidth parameter $\sigma$ which is also symmetric on the permutation of its arguments, and $\text{card}(\cdot)$ denotes the cardinality of a set. $\{\hat{p}_n, \hat{f}_n\}$ completely describes a general finite point process as we discussed before. In the experimental section, we used a spherical normal distribution for $\kappa(\cdot, \sigma)$. The optimal bandwidth of the kernel is usually set by a cross validation method. However, to speed up computation we used a modified Silverman’s rule that guarantees consistent density estimation instead [114]: $\sigma_n = \text{card}(X^{(n)})^{-1/(n+4)} \cdot \sigma_1$ where $n > 1$ is the dimension and $\sigma_1$ is the bandwidth for a single sample in one dimension.

### 5.3 Hellinger Divergence

First, we investigate Hellinger divergence for point processes. There are two forms:

a Hilbertian metric form (5–14) with $\alpha = \frac{1}{2}, \beta = 1$, and the $\phi$-divergence form with $\phi(x) = (1 - \sqrt{x})^2 \quad (5–15)$.

$$
\begin{align*}
\tilde{d}_H^2(P, Q) &= \int_{\Omega} \left( \frac{\sqrt{dP}}{d\mu} - \frac{\sqrt{dQ}}{d\mu} \right)^2 d\mu \\
&= \int_{\Omega} \left( 1 - \sqrt{\frac{dP}{d\mu}} / \sqrt{\frac{dQ}{d\mu}} \right)^2 Q \quad (5–15)
\end{align*}
$$
5.3.1 Estimator

Using the nonparametric estimator of finite point process (5–12) and (5–13) to (5–14) with respect to the reference measure \( \mu \), we obtain the Hellinger distance estimator:

\[
d_{H1}^2(\hat{P}, \hat{Q}) = \int_\Omega \left( \sqrt{\frac{d\hat{P}}{d\mu}} - \sqrt{\frac{d\hat{Q}}{d\mu}} \right)^2 d\mu \tag{5–16}
\]

Since the integration of (5–16) is not easily computable, we resort to the Monte Carlo method for estimation.

We would like to use the average probability measure \( \nu = \frac{\hat{P} + \hat{Q}}{2} \) for importance sampling, however, \( \nu \) is usually not absolutely continuous with respect to \( \mu \). Using the Lebesgue decomposition theorem \([42]\), \( \mu \) can be decomposed to \( \mu = \mu_\perp + \mu_\ll \), where \( \mu_\perp \perp \nu \) and \( \mu_\ll \ll \nu \). Since \( d^2(0,0) = 0 \), the integration with \( \mu_\perp \) leads to zero, and hence, the importance sampling can be achieved as,

\[
d_{H1}^2(\hat{P}, \hat{Q}) = \int_x d^2 \left( \frac{d\hat{P}}{d\mu}, \frac{d\hat{Q}}{d\mu} \right) d\mu
\]

\[
= \int_x d^2 \left( \frac{d\hat{P}}{d\mu}, \frac{d\hat{Q}}{d\mu} \right) d\mu_\perp + \int_x d^2 \left( \frac{d\hat{P}}{d\mu}, \frac{d\hat{Q}}{d\mu} \right) d\mu_\ll
\]

\[
= \int_x d^2 \left( \frac{d\hat{P}}{d\mu}, \frac{d\hat{Q}}{d\mu} \right) \frac{d\mu_\ll}{d\nu} d\nu = \int_x d^2 \left( \frac{d\hat{P}}{d\mu}, \frac{d\hat{Q}}{d\mu} \right) \frac{2}{\frac{d\hat{P}}{d\mu_\ll} + \frac{d\hat{Q}}{d\mu_\ll}} d\nu
\]

\[
= \int_x d^2 \left( \frac{d\hat{P}}{d\mu}, \frac{d\hat{Q}}{d\mu} \right) \frac{2}{\frac{d\hat{P}}{d\mu} + \frac{d\hat{Q}}{d\mu}} d\nu. \tag{5–17}
\]

Using (5–17), we can either use all the original spike trains that were used to estimate \( \hat{P} \) and \( \hat{Q} \), or generate arbitrary number of spike trains from \( \hat{P} \) and \( \hat{Q} \). The latter approach provides more controlled variance but is slower.

We can use an alternative estimator derived from (5–15) as,

\[
d_{H2}^2(\hat{P}, \hat{Q}) = \int_\Omega \left( 1 - \sqrt{\frac{d\hat{P}}{d\mu}/\frac{d\hat{Q}}{d\mu}} \right)^2 d\hat{Q} = E_{\hat{Q}} \left[ \left( 1 - \sqrt{\frac{d\hat{P}}{d\mu}/\frac{d\hat{Q}}{d\mu}} \right)^2 \right] \tag{5–18}
\]

Note that expectation can be estimated by the strong law of large numbers.
5.3.2 Illustrative Example

Two point processes with two events or less with same marginal intensity function but with different event correlation structure are chosen to illustrate the method (Fig. 5-1). In point process A, the two timings are correlated; the interspike interval (ISI) has a narrow distribution. In point process B, the two timings are independent; each action potential has a precise timing. Both point processes have a lossy noise; each action potential has a probability of disappearing with probability \( p = 0.1 \). Note that process A represents renewal type of neural code, while process B represents precisely timed action potential. We took 40 realizations of each process, and computed the Hilbertian metric corresponding to Hellinger metric.

We choose point process A and B such that \( P_n \) and \( f_1 \) are identical for both; the only difference is in \( f_2 \). Non-parametric estimation from the samples show the correlation can be detected visually (Fig. 5-1). Due to the small number of realizations for \( f_1 \), the estimate is crude, but recall that it will be weighted by \( \hat{p}_1 \) which is also small.

The Monte Carlo integration of (5–17) with Hellinger metric converged within hundreds of iterations. This is illustrated in figure 5-2 where the squared distance between two sets of realizations from process A and one from process B is compared. The 90\% confidence interval can distinguish process A and process B within 50 iterations (see also figure 5-3).

The statistical power of the test increases as the number of sample increases. Table 5-1 summarizes the power of discriminating process A and B with threshold of 0.05. The top row is computed using only the samples themselves (twice the number used to estimate the probability measure since there are two of them) and the bottom row uses 500 generated samples from the estimated measures (see section 3.2).

5.3.3 Detecting Learning

Let us consider a hypothetical experiment where one investigates the effectiveness of a drug or a stimulation protocol that modulates or induces synaptic plasticity. We collect
Figure 5-1. (Top rasters) 40 realizations each from point process A (top) and B (bottom), x-axis is time in seconds. The spike trains are sorted in terms of time of first action potential (right). In both A and B, $t_1 \sim U(0.2, 0.3)$. $t_2 \sim t_1 + 0.3 + \mathcal{N}(0, 0.01)$ in A, and $t_2 \sim U(0.5, 0.6) + \mathcal{N}(0, 0.01)$ in B. Both $t_1$ and $t_2$ have 0.1 probability of loss. (Below) Estimated parameters from the spike trains shown above. $p_n$, $f_1$ and $f_2$ are depicted from left to right. An nonsymmetric $f_2$ is plotted for simplicity.
Figure 5-2. 1000 parallel Monte Carlo integrations to show the convergence. Max and min, 90\% empirical confidence interval, estimated standard deviation, and mean of estimation of (5–17) for two fixed sets of realizations of point process A (blue), and between two fixed sets of realizations of point process A and B (red) are plotted.

Figure 5-3. Empirical distribution of Hellinger divergence. Histogram of statistics estimated from 40 realizations (left) and 100 realizations (right).

data from piecewise stationary segments where little plasticity is presumed. The goal is to detect the plasticity by computing the divergence between datasets.

A randomly connected network of 1000 neurons (800 excitatory regular spiking, 200 inhibitory fast spiking [48]) with delay and spike timing dependent plasticity (STDP) on excitatory to excitatory synapses was simulated similar to [47]. The network had current injection to randomly preselected 50 neurons for 2 ms once a second. The response of each neuron in the 100 ms window after current injection was collected as trials.

As an initialization, STDP was turned on to adapt the weights for the first 20 s, for the second 20 s STDP was turned off to reach the steady state. The following 120 s
consists of three 40 s block; the first 40 s correspond to the baseline measurements, during the second 40 s STDP was turned on (or off for the control experiment) to induce changes, and the last 40 s corresponds to the trials from the modified system (or null hypothesis for the control). Each neuron was separately analyzed. Note that the stimulation protocol is a typical stimulation given to cultured neurons on microelectrode arrays (MEAs) [126].

Although the data originated from one long experiment cut into 40 trials, and divergence from individual neurons were collected together, the data from STDP and no STDP showed significant difference (Fig. 5-4). Assuming the neurons to be an ensemble, 91.9% were rejected based on the empirical surrogate test. Detailed observation revealed that STDP in general brings the timing of firing closer to the stimulation time, while some neurons did not change at all.

5.3.4 Artificial Spike Trains

In table 5-2, the above proposed method is tested on 4 distinct point processes that have emphasis on features related to neuroscience; correlated spike timings [34], firing rate change [46], precisely timed spike trains [26, 28, 103], and serial correlation [15].

To answer the question of whether a set of spike trains originate from the null hypothesis, we performed hypothesis testing by generating the empirical distribution of the divergence given finite number of samples. The generation of surrogate distribution with carefully chosen assumptions can compensate for the bias and gives estimated p-value to the test. In case of table 5-2, we also know the alternative hypothesis, hence we computed the statistical power.

Non-parametric method is a generalist, while each of the parametric method is a specialist. As expected, one can see from the table that the proposed method is a

<table>
<thead>
<tr>
<th># of samples</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>samples only</td>
<td>0.201</td>
<td>0.420</td>
<td>0.763</td>
<td>0.887</td>
<td>0.965</td>
<td>0.993</td>
<td>0.999</td>
<td>0.999</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>500 MC</td>
<td>0.202</td>
<td>0.463</td>
<td>0.749</td>
<td>0.904</td>
<td>0.983</td>
<td>0.995</td>
<td>0.999</td>
<td>1.000</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 5-1. Statistical power of integration methods on number of samples.
Figure 5-4. Significance of divergence based on Hellinger distance values before and after learning protocol. (Left) Solid curve represents the null hypothesis; sorted divergence values for each neuron in stationary condition (without learning). Horizontal dotted line is the 0.05 significance threshold (0.161) for rejection. Crosses represent the corresponding divergence values for each neuron when STDP was enabled for 40 s. 91.9% of the neurons had significantly different spiking statistic after STDP learning ($p < 0.05$). (Right) Histogram of the null and alternative hypothesis. The stimulated neurons and inhibitory neurons are excluded ($n = 754$).

generalist. One needs a generalist when the change is subtle or unknown. When large dataset with various possibility is given, a generalist would be a good choice to use in detecting changes. Given infinite number of samples, the proposed method is guaranteed to detect the change if there is a change.

5.3.5 Non-stationarity Detection

Independence and stationarity across trials are essential requirements for the analysis of experimental observations. Testing whether the measured dataset has the desired property is a challenging statistical problem, even more so when the measurements are in the form of spike trains. The best we can do is to use sensitive methods to detect non-stationarity and support the claim of stationarity from failure of detection. The point process divergence can be used to perform hypothesis testing for the inter-trial
Table 5-2. Statistical power (higher the better) of various methods for hypothesis testing with threshold 0.05. The power was computed with respect to empirical surrogate distribution. Data set I: a two action potential model with or without correlation, II: homogeneous Poisson with different rate, III: precisely timed spike train and Poisson equivalent, IV: renewal process with or without serial correlation. 40 spike trains are used for I, II, and III, 320 spike trains are used for IV. FR: absolute difference between mean firing rate, FF: Fano factor difference, ISI: KS-statistic of inter-spike interval distribution, N: KS-statistic of total spike count distribution, $L_210$, $L_2100$: mean square difference of smoothed PSTH with corresponding kernel size, TTFS: KS-statistic of time to first spike distribution HL10, HL100: proposed non-parametric method with 100 ms and 10 ms kernel size $\sigma_1$.

<table>
<thead>
<tr>
<th>Exp</th>
<th>FR</th>
<th>FF</th>
<th>ISI</th>
<th>N</th>
<th>$L_2100$</th>
<th>$L_210$</th>
<th>TTFS</th>
<th>HL10</th>
<th>HL100</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>0.0685</td>
<td>0.0589</td>
<td>0.9516</td>
<td>0.0502</td>
<td>0.0657</td>
<td>0.0164</td>
<td>0.0415</td>
<td>0.0579</td>
<td>1.0000</td>
</tr>
<tr>
<td>II</td>
<td>0.9990</td>
<td>0.2869</td>
<td>0.2067</td>
<td>0.9275</td>
<td>0.9990</td>
<td>0.9971</td>
<td>0.4579</td>
<td>0.9217</td>
<td>0.8927</td>
</tr>
<tr>
<td>III</td>
<td>0.0859</td>
<td>0.9690</td>
<td>1.0000</td>
<td>0.1632</td>
<td>0.0995</td>
<td>0.1391</td>
<td>0.2531</td>
<td>0.8657</td>
<td>1.0000</td>
</tr>
<tr>
<td>IV</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0483</td>
<td>0.0000</td>
<td>0.0541</td>
<td>0.1140</td>
<td>0.0309</td>
<td>0.7942</td>
<td>0.6386</td>
</tr>
</tbody>
</table>

Figure 5-5. Performance of test with Hellinger divergence depends on the number of samples. The blue (thin) and black (thick) curves correspond to two different scales of the A4 vs A5 discrimination problem. The dotted line is the $\phi$-divergence form of the estimator (5–18). The error bars are standard deviation estimated from 10 Monte Carlo runs.
non-stationarity. As an application of the proposed method, we apply it to investigate the short-term non-stationarity and/or stationarity of datasets collected from neural cultures.

Periodically repeated stimulation to neural system is a standard technique to probe and analyze the state of the system. Certainly this is not a natural input for in vivo neural systems where the stimulus is more or less random, but the resulting spike train signals are easier to analyze; in many cases the response patterns are strictly time locked to the stimulation. From the dataset observed from neural cultures, we have previously observed that trial-to-trial variability plays a major role for classification or decoding of the repeated input stimulus conditions [30].

For a single probing trial, the measured spike train is not stationary in general. However, multiple trials can be considered independent and stationary through proper initialization, for example by controlling the inter-trial interval. By inter-trial stationarity, we mean that the probability law behind the set of trials do not change over time. In other words, if a set of trials obey inter-trial stationarity, each trial is independent and identically distributed. When the observations are spike trains, this means that the trials are independent realizations of the same point process (non-stationary within the trial in general). Given a measure of divergence between two sets of spike trains, we can perform statistical test to show if the underlying point processes are different, hence, showing inter-trial non-stationarity.

When electrically stimulated, the response of neural culture recorded through a micro-electrode array (MEA) changes as the stimulation is periodically repeated. Various resources of neurons are used and short term plasticity is first observed from the action potential timings. However, after a while the responses stabilize due to homeostatic mechanisms. The time period that this happens is related to the time constants of the neurophysiological processes. Hence, when stimulated for a long time, the response becomes more or less stationary. In neural cultures, most action potentials occur time locked to the stimulation time, therefore this is visually verifiable. Assuming a certain
portion of the trials to be inter-trial stationary, we would like to determine by how much we can extend the assumption.

To utilize the multi-channel recording nature of the experiments, we assumed that the channels have similar non-stationarity pattern over time. By averaging over the channels, a smoother estimate of the divergence is obtained. The distribution for null hypothesis must also be changed accordingly by taking the mean of across channels.

We have selected 5 to 15 active channels with phasic responses which seems to be changing in the experimental setup. We removed the channels that does not have less than or equal to 5 action potentials for 90% of the trials. The inactive channels that cannot be reliably estimated (mean number of action potentials less than 0.5) are removed from analysis as well. This process also reduces the bias and increases sensitivity in some cases. Including all the channels for analysis does not change the result significantly.

We assume that the culture reaches a stationary state at the end (last 80 trials out of 600 (or 1200) trials over 5 minutes), and we want to determine how much more one can extend the window that is within the stationarity condition. The null hypothesis is that the set of trials of interest came from the same point process as the last 40 trials.

We used a resampling technique to create the surrogate distribution of divergences for the null hypothesis. First recall that the last 80 trials are i.i.d, by the previous assumption. We randomly choose 40 of the last 80 trials to divide into two. The divergence between these two halves is repeatedly computed for 2000 times to create the empirical distribution under the null hypothesis. The actual divergence value is compared against this to test if it significantly deviates from the null hypothesis.

Figure 5-6 shows that the cultures are non-stationary for the first 350 samples corresponding to 3 minutes. The dashed line, and the dotted line correspond to empirical and Gaussian assumed 95 per cent confidence level for the null hypothesis. The divergence values (blue and green) have decreasing trend during this period and cross the threshold. The non-stationarity (deviation from the null-hypothesis) is steep for the first 200 samples,
Figure 5-6. Nonstationarity detection. (top) The Hellinger divergence of a moving window of 40 trials. Values are averaged over 10 selected channels. (bottom two rows) example raster plots. The three thick black bars indicate the window of 40 trials the moving window was compared against.

followed by a plateau. These observations can also be made with other methods except ISI and FF.

The right panel of figure 5-7 shows a more discontinuous state change at around 400 samples. Despite a 5 minute rest at 600 samples time point, the culture response is still within the pseudo-stationary region. Hence, the sudden change is likely a long-term plasticity effect. Again similar observations can be made with different divergence measures. Depending on cultures, we could distinguish short-term changes and long-term changes ($N = 6$).
5.3.6 Kernel Size

The proposed estimator has a free parameter – kernel size (also known as bandwidth). If the kernel size is too small, every small difference will be considered different, therefore resulting in no discriminability. On the other hand, if the kernel size is too big, no matter how different the spike timings are, they will be considered very similar, and again the discrimination power would be lost. In theory, if the kernel size goes to zero as the number of samples goes to infinite, it is possible to have a consistent density estimator [114].
However, in practice we only have finite number of samples, and choosing the best kernel size becomes an issue.

In figure 5-8, it can be seen that the kernel size plays a significant role. The type of feature, number of samples, problem difficulty (jitter size) are all related to the problem, and hence it is non-trivial to estimate the best kernel size. However, for the current kernel size scaling rules, it can be seen that the physiological jitter size is in the range of the optimal kernel size, and the performance is not very sensitive to small changes in the kernel size. Hence, we recommend using the physiologically relevant kernel sizes (1 millisecond to 1 second range) and optimize for performance if possible as usually done with the spike train distance measures [123, 124].

Figure 5-8. Jitter 20 ms is more difficult that 10 ms. The peak performance for jitter 10 ms case is maintained for about 0.5 order of magnitude. Number of samples is fixed to 30. $\phi$-divergence based estimator (solid lines) are compared against the importance sampling (dotted lines). They have essentially same performance.
5.4 Symmetric $\chi^2$-divergence

In this section, we particularly focus on the $\chi^2$-divergence, where $\phi(t) = (t - 1)^2$, that is,

$$D_{\chi^2}(P, Q) = \int \left( \frac{dP}{dQ} - 1 \right)^2 dQ. \quad (5\text{-}19)$$

Notice that, the existence of $D_{\chi^2}$ requires the Radon-Nikodym derivative $dP/dQ$, also known as the likelihood ratio, to be $L_2(Q)$ integrable.

The $\chi^2$-divergence (or any $\phi$-divergence as such) can be consistently estimated as

$$\hat{D}_{\chi^2} = \int \left( \frac{d\hat{P}}{d\hat{Q}} - 1 \right)^2 d\hat{Q} \quad (5\text{-}20)$$

where $d\hat{P}/d\hat{Q}$ is a consistent estimator of the Radon-Nikodym derivative and $\hat{Q}$ is the empirical probability measure. However, the estimation of the Radon-Nikodym derivative only makes sense when $P \ll Q$ since the term $dP/dQ$ becomes undefined otherwise. For example, in the extreme case when $P$ and $Q$ have disjoint supports. To tackle this problem we propose to work with $D_{\chi^2}(P, (P + Q)/2)$ instead, since $P \ll (P + Q)/2$. This divergence is the same as the symmetric $\chi^2$-divergence

$$D_S(P, Q) = \int \left( \frac{d(P - Q)}{d(P + Q)} \right)^2 d(P + Q), \quad (5\text{-}21)$$

addressed in [43], except for a multiplicative factor: $D_S(P, Q) = 2D_{\chi^2}(P, (P + Q)/2) = 2D_{\chi^2}(Q, (P + Q)/2)$. Moreover, the symmetric $\chi^2$-divergence is a metric on the space of all probability measures [43].

In this section, we explore two conceptually different approaches to estimate the Radon-Nikodym derivative and the symmetric $\chi^2$-divergence. The first method is identical to the previous section, whereas the second method maps the spike trains in a more structured space, and utilizes strictly positive definite kernels on that space to estimate the Radon-Nikodym derivative. The use of strictly positive definite kernels to estimate Radon-Nikodym derivative and divergence functionals has recently gained considerable
interest in $\Omega = \mathbb{R}^d$ [56, 77]. We extend this idea and show that this approach can also be applied on a more abstract space.

5.4.1 Point Process Representation

In addition to the stratified space approach (section 5.2), we present an alternative method to approximate the Radon-Nikodym derivative using the smoothed spike train space.

The stratification approach inherently distributes the samples with different number of action potentials in different groups, and therefore, two spike trains with different number of action potentials never interact. This poses a problem in estimation if the count distribution is flat i.e. the spike trains tends to have different number of action potentials.

An alternate approach to represent a spike train is to project the spike train in a different, and perhaps more structured, space. Let $\mathcal{S}$ be the space of all $L_2$ integrable functions over $\mathcal{X}$ i.e., $\mathcal{S} = L_2(\mathcal{X})$. Given $\omega = \{t_1, \ldots, t_n\} \in \Omega$, define a mapping $G : \Omega \rightarrow \mathcal{S}$ as $G(\omega)(t) = \sum_{i=1}^{n} g(t, t_i)|_{\mathcal{X}}$ such that $G$ is injective. When $g$ is a translation invariant function that decays at $\infty$, $G$ can be considered a smoothed spike train [84].

There are many different $g$’s that make the mapping $G$ injective. For example, when $g(x, y)$ is a bounded strictly positive definite function. To see this, consider two spike trains $\omega_1 = \{t_1^1, \ldots, t_n^1\}$ and $\omega_2 = \{t_1^2, \ldots, t_m^2\}$ such that all spike timings are distinct i.e., $\omega_1 \cap \omega_2 = \emptyset$, and assume that $G(\omega_1) = G(\omega_2)$. Since the kernel matrix $[K]_{ij} = g(t_i, t_j)$ where $t_i, t_j \in \omega_1 \cup \omega_2$ is full rank (by the virtue of strict positive definiteness), $G(\omega_1)(t) = G(\omega_2)(t)$ for all $t \in \omega_1 \cup \omega_2$ implies that $\omega_1 = \omega_2$.

The $\sigma$-algebra of $\mathcal{S} = L_2(\mathcal{X})$ is induced by $\mathcal{B}(\Omega)$ and $G$ to be $\sigma(\mathcal{S} \cup \{G(A) | A \in \mathcal{B}(\Omega)\})$. Then $G$ is measurable, and we can define an induced probability measure $U$ such that $U(\mathcal{S} \setminus G(\Omega)) = 0$ and $U(G(A)) = P(A)$ for $A \in \sigma(\Omega)$. Let $U$ and $V$ be two probability laws on $\mathcal{S}$ induced by $P$ and $Q$, respectively, then, the following two propositions show that the Radon-Nikodym derivative in $\Omega$ can be transferred to $\mathcal{S}$.

**Proposition 10.** If $P \ll Q$ then $U \ll V$
Proof. Let \( A \subset G(\Omega) \), then

\[
V(A) = 0 \Rightarrow Q(G^{-1}(A)) = 0 \Rightarrow P(G^{-1}(A)) = 0 \Rightarrow P(A) = 0.
\]

Similar proof follows when \( A \not\subset G(\Omega) \).

Proposition 11. \( \frac{dP}{dQ}(\omega) = \frac{dU}{dV}(G(\omega)) \)

Proof. For any integrable function \( \phi : \mathcal{S} \to \mathbb{R} \),

\[
\int_{\Omega} \phi(G(\omega))dP(\omega) = \int_{\mathcal{S}} \phi(f)dU(f) = \int_{\mathcal{S}} \phi(f) \frac{dU}{dV}(f)dV(f) = \int_{\Omega} \phi(G(\omega)) \frac{dU}{dV}(G(\omega))dQ(\omega)
\]

Therefore, \( dP(\omega) = dU/dV(G(\omega))dQ(\omega) \).

Corollary 12. The \( \chi^2 \)-divergence between \( P \) and \( Q \) is the same as the \( \chi^2 \)-divergence between \( U \) and \( V \), i.e.

\[
\int_{\Omega} \left( \frac{dP}{dQ}(\omega) - 1 \right)^2 dP(\omega) = \int_{\mathcal{S}} \left( \frac{dU}{dV}(f) - 1 \right)^2 dU(f).
\]

This corollary shows that the problem of estimating the Radon-Nikodym derivative and \( \chi^2 \)-divergence in \( \Omega \) can be framed in \( \mathcal{S} \), which is more structured.

5.4.2 Estimation of Symmetric \( \chi^2 \)-divergence

Following the two representations described in the previous section, we propose two distinct estimators of the symmetric \( \chi^2 \)-divergence. Notice that, in this section we assume that \( P \ll Q \) to simplify the derivations. However, the proposed methods can be trivially extended to estimate symmetric \( \chi^2 \)-divergence by replacing \( Q \) by \( (P + Q)/2 \).

5.4.2.1 Stratification approach

The Radon-Nikodym derivative \( dP/dQ \) in \( \mathbb{R}^n \) can be consistently estimated, under appropriate conditions, by the ratio of the Parzen estimates \( \frac{\hat{dP}/d\mu}{\hat{dQ}/d\mu} \) where \( \mu \) denotes the Lebesgue measure [17]. In our case, since \( dP/dQ = \sum_{n=0}^{\infty} dP_n/dQ_n \) we
estimate the Radon-Nikodym derivative as
\[
\frac{dP}{dQ} = \sum_{n=0}^{\infty} \frac{dP_n/d\mu_n}{dQ_n/d\mu_n}
\]
Using this estimate, the estimator of the $\chi^2$-divergence becomes
\[
\hat{D}_{\chi^2} = \int \left( \frac{dP}{dQ} - 1 \right)^2 d\hat{Q} = \sum_{n=0}^{\infty} \int \left( \frac{dP_n/d\mu_n}{dQ_n/d\mu_n} - 1 \right)^2 d\hat{Q}_n
\]
\[
= \frac{1}{\text{card}(Y^{(0)})} \left( \frac{m \text{card}(X^{(0)})}{n \text{card}(Y^{(0)})} - 1 \right)^2 + \sum_{n=1}^{\infty} \frac{1}{\text{card}(Y^{(n)})} \sum_{y_i \in Y^{(n)}} \left( \frac{l \sum_{x_j \in X^{(n)}} \kappa_n (\omega_j^1 - \omega_i^1; \sigma_n)}{m \sum_{y_j \in Y^{(n)}} \kappa_n (\omega_j^2 - \omega_i^2; \sigma_n)} - 1 \right)^2.
\]
where $\{\omega_i^1\}_{i=1}^m$ and $\{\omega_i^2\}_{i=1}^l$ are samples from P and Q respectively. Notice that, although the individual estimators for each $n$ is consistent by the virtue of the consistency of the Radon-Nikodym derivative estimate, and the empirical measure, the entire estimator might not be consistent since it requires adding a countable number of such estimators. Moreover, the convergence rate of the estimated Radon-Nikodym derivative depends on the dimensionality of the Euclidean space. Therefore, the convergence rate of the entire estimator is governed by the slowest convergence rate. However, these issues can be tackled by assuming that number of spikes in a spike train is upper bounded i.e. $n < n_{\text{max}}$.

5.4.2.2 Smoothed spike train approach

In the smoothed spike train representation, we focus on estimating the Radon-Nikodym derivative in $S$ rather than in $\Omega$ i.e. to estimate $\frac{dU}{dV}$ from samples $\{f_i = G(\omega_i^1)\}_{i=1}^n$ and $\{g_i = G(\omega_i^2)\}_{i=1}^l$. Although this problem can be approached in several ways, we follow the approach suggested in [56].

Using the triangle inequality, we get,
\[
|D_{\chi^2} - \hat{D}_{\chi^2}| = \left| \int (dU/dV - 1)^2 dV - \int (d\hat{U}/d\hat{V} - 1)^2 d\hat{V} \right|
\leq \left| \int (d\hat{U}/d\hat{V} - 1)^2 d(V - \hat{V}) \right| + \left| \int (dU/dV - d\hat{U}/d\hat{V})^2 dV \right|.
\]  (5–22)
This inequality shows that the error between the actual and estimated Radon-Nikodym derivative is bounded by two $L^2$ distances. The first term goes to zero as $n \to \infty$ since $\hat{V} \to V$ almost surely. Therefore, in order to get a consistent estimate of $\mathbb{D}_{\chi^2}$ it is important to get an appropriate estimate of Radon-Nikodym derivative that makes the second term arbitrarily close to zero.

Following [56], we assume that $\frac{dU}{dV}(f) = \sum_{i=1}^{l} \alpha_i \tilde{k}(g - g_i)$ where $\alpha_i$’s are a real coefficients and $\kappa(f - g)$ is a strictly positive definite kernel [10]. This expansion is justified due to the following proposition which states that the functions of the form $\sum_{i=1}^{\infty} \alpha_i \tilde{k}(g - g_i)$ are dense in $L^2(S, V)$ i.e. it can approximate any function $p \in L^2(S, V)$ with arbitrary accuracy in the $L^2$ sense. Notice that since $\frac{dU}{dV} \in L^2(S, V)$, this implies that the proposed expansion can approximate the Radon-Nikodym derivative arbitrarily.

**Theorem 13.** Let $\tilde{k}(x, y)$ be a symmetric strictly positive definite continuous kernel on $S \times S$ and $V$ is a probability measure on $S$ such that $\int_S \tilde{k}^2(x, y) dV(x) < \infty$ for all $y \in S$, then $\text{span} (\tilde{k}(x, y) : y \in G \subset S)$ is dense in $L^2(S, V)$, where $G$ denotes the subset where the measure $V$ lies.

**Proof.** Let us assume that the span is not dense in $L^2(S, V)$, then there exists a function $g \in L^2(S, V)$ such that $\int g(y) \kappa(x, y) dV(y) = 0$. Therefore, $\int \int g(x) g(y) \kappa(x, y) dV(x) dV(y) = 0$. Since $\tilde{k}$ is strictly positive definite, this implies that $g$ is zero a.e. $V$. \[\Box\]

Theorem 13 is a closely related to the notion of Universal kernel; instead of continuous functions $\mathcal{C}(S)$ for a compact $S$, we showed that it is dense in $L^2(S)$ functions.

Using the kernel expansion, the second term in (5–22) can be expressed as,

$$
\int \left( \frac{dU}{dV}(f) - \sum_{i=1}^{l} \alpha_i \tilde{k}(f, g_i) \right)^2 dV(f)
$$

$$
= \int (\frac{dU}{dV})^2(f) dV(f) - 2 \sum_{i=1}^{l} \int \alpha_i \tilde{k}(f, g_i) dU(f) + \int \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j \tilde{k}(f, g_i) \tilde{k}(f, g_j) dV(f)
$$

$$
\approx C - \frac{2}{m} \sum_{i=1}^{l} \sum_{j=1}^{m} \int \alpha_i \tilde{k}(f_j, g_i) + \frac{1}{l} \sum_{i=1}^{l} \sum_{j=1}^{l} \sum_{k=1}^{l} \alpha_i \alpha_j \tilde{k}(g_k, g_i) \tilde{k}(g_k, g_j)
$$

83
where $C$ is a constant. Therefore, the second term is $\left(5-22\right)$ is minimized for $
abla = (l/m)(K_{QQ}K_{QQ} + \lambda/I)^{-1}K_{QP}1$ where $\lambda$ is a regularization parameter required to avoid overfitting and $[K_{PQ}]_{ij} = \tilde{\kappa}(G(\omega_i), G(\omega_j))$ is the gram matrix. The estimated $\alpha$ can then be used to estimate $\hat{\mathcal{D}}_{\alpha^2} = (K_{QQ}\alpha - 1)^2/1$.

Examples of kernels that satisfy the required properties are the nCI kernels (section 2.3).

$$\tilde{\kappa}(f, g) = \exp\left(-\int_X (f(t) - g(t))^2 dt/\sigma^2\right)$$ and $$\tilde{\kappa}(f, g) = \int_X \exp\left(-\frac{(f(t) - g(t))^2}{\sigma^2}\right) dt.$$ For the simulations, we use the first kernel, set the $\sigma$ to be the median distance between samples $f_i$'s and set the regularization parameter to be $1/n$.

5.4.3 Results

We present 3 hypothesis testing experiments to demonstrate the pros and cons of each method. For each example, there are two classes of spike trains, and the test is to find whether the spike trains originate from the same probability law. The statistical power, which measures the type 2 error given a fixed type 1 error ($\alpha = 0.05$), is estimated via a surrogate test. We compare the results with the baseline dissimilarity $\lambda_{L2}$ which computes the $L_2$ distance between PSTH (peri-stimulus time histogram) as a statistic. We observe that the kernel based estimator consistently approaches high power quickly for all examples.

5.4.3.1 Two action potentials

Two point processes with two events or less with same marginal intensity function but with different event correlation structure are chosen to illustrate the methods (Fig. 5-9 left). In point process from $H_0$, the two timings are correlated; the interspike interval (ISI) has a narrow distribution. In point process from $H_1$, the two timings are independent; each action potential has a precise timing. Both point processes have a lossy noise; each action potential has a probability of disappearing with probability $p = 0.1$. Note that $H_0$
Figure 5-9. Two action potentials that are correlated ($H_0$) and independent ($H_1$). $\sigma_1 = 10 \text{ ms}$ is used for the stratified kernel. (Left) Spike trains from the null and alternate hypothesis. (Right) Comparison of the power of each method. The error bars are standard deviation over 5 Monte Carlo simulations.

represents renewal type of neural code, while process $H_1$ represents precisely timed action potentials.

Since the dimension of the problem is at most 2, this problem is easier for the stratified estimator (Fig. 6-2 right). Nevertheless, the kernel based estimator quickly catches up as the number of sample increases. The firing rate function based dissimilarity fails to discriminate the two processes because the intensity function is designed to be identical.

5.4.3.2 Inhomogeneous Poisson process

In Poisson process, the mean rate function fully describes the process. We simulate two inhomogeneous Poisson process where the rate changes at 100 ms. In figure 5-10, the $\lambda_{L2}$ with the right bin size outperforms others. The stratified approaches suffers from the spread of samples in different strata and the curse of dimensionality, while the kernel based estimator quickly approaches high power.
Figure 5-10. Poisson process with rate changing in step function from 3 to 2, and 2 to 3 at 100 ms. \( \sigma_1 = 100 \text{ ms} \) is used for the stratified kernel. The mean number of action potential is fixed to 5. (Left) Spike trains from the null and alternate hypothesis. (Right) Comparison of the power of each method. The error bars are standard deviation over 5 Monte Carlo runs.

5.4.3.3 Stationary renewal processes

Renewal process is a widely used point process model to compensate the deviation from Poisson process [75]. Stationary renewal process with gamma interval distribution is simulated.

Despite the fact that the count distribution is widely distributed and the high dimensionality, the stratified estimator performs the best. This is because the difference in the count distribution is easily captured when the bandwidth is relatively large. The kernel based method does not suffer from this fact and performs well.

The mean firing rate based estimator is not consistent in this example, in the sense that the performance of the method does not improve with the sample size. As the number of sample approaches infinite it fails to detect. However, interestingly in the finite sample range, the variance in the rate function depends on the regularity of the firing pattern, therefore it is able to detect the difference.
Figure 5-11. Gamma distributed renewal process with shape parameter $\theta = 3$ ($H_0$) and $\theta = 0.5$ ($H_1$). $\sigma_1 = 100$ ms is used for the stratified kernel. The mean number of action potential is fixed to 10. (Left) Spike trains from the null and alternate hypothesis. (Right) Comparison of the power of each method. The error bars are standard deviation over 20 Monte Carlo runs.

5.4.4 Auditory Neuron Clustering

We explore the auditory response of rat thalamic neurons. As seen in figure 3-1, the trial-to-trial variability of these neurons have some component of precise timing. A ICRA frozen noise was presented 56 times to the system and corresponding 56 single neuron response is recorded. We applied smoothed spike train based $\chi^2$ divergence estimator (theorem 13) to all pairs of sets of spike trains. The aim is to cluster similar response point processes together. This could possibly help identify neural assemblies.

67 single units from medial geniculate body (49 ventral, 18 dorsal) are analyzed. Each trial resulted in 2.5 seconds of spike train recording. We applied distance based hierarchical clustering (figure 5-12). Note that compared to mean rate function clustering, the symmetric $\chi^2$ measure gives different results. Also, there is a bias (around 0.2) in the distance estimation.
Figure 5-12. Point process clustering of auditory neurons with $\chi^2$ divergence compared against $\lambda_{L2}$. The neurons were divided into three groups according to their mean rate: high ($> 3.5$, $n = 17$), low ($< 1.5$, $n = 25$), medium (rest, $n = 25$). Agglomerative hierarchical cluster tree minimizing the average pairwise distance is used to create the hierarchy. X-axis corresponds to neuron index, and Y-axis is time for the raster and distance for the dendrogram.
5.4.5 Discussion

In this section, we have introduced two distinct methods of estimating $\chi^2$-divergence and Radon-Nikodym derivative between two point processes, namely the stratification approach and the smoothed spike train approach. The stratification approach is simpler in nature, and can be evaluated more efficiently than the later. To be specific, the stratification approach is $O(ml)$ in computation whereas the smoothed spike train approach is $O(l^3 + ml^2)$ in computation. However, the computation can be improved by exploiting the structure of the Gram matrices, and has been elaborately studied in the literature \[35\]. The stratification approach, on the other hand, suffers when the count distribution of the spike trains is flat i.e. the spike trains tend to contain different number of spikes (as demonstrated in the subsection 5.4.3.1). This poses a problem in estimation since each stratum contains lesser spikes in a small sample size situation. Therefore, the smoothed spike train approach is applicable to a larger variety of point processes, but, at the cost of higher computation.

The problem of estimating Radon-Nikodym derivative, however, can also be approached in several other ways that remain to be explored. For example, notice that, although the space of spike trains is not Euclidean, it is still a (semi-)normed space where the norm is induced by either the stratification or the smoothed spike train representation, or other available distance metrices such as Victor-Purpura metric \[123\]. Recently, there have been considerable work on estimating regression functionals on semi-normed space by introducing bounded kernels on the norm of the space \[24\]. Similar approach can be considered for estimating Radon-Nikodym derivative by employing similar kernels, albeit properly normalized, with the help of Parzen type estimation.

Moreover, in this section we have only addressed the problem of estimating symmetric $\chi^2$-divergence. However, in practice, other divergences, such as KL-divergence, are also widely used. The proposed methods can be extended to estimate any arbitrary divergence following the approach presented in \[77\]. This requires the strictly positive definite kernel
to be rich enough to present a wide variety of continuous functions or in a sense to be universal [73]. However, we leave these interesting extensions as future work.
CHAPTER 6
KERNEL BASED DIVERGENCE

6.1 Introduction

In this chapter, we follow a kernel based approach, since a kernel – bivariate positive definite function – can be defined on any arbitrary space, and it provides an implicit similarity measure on the space that can be used to design divergence and dissimilarity on the probability laws defined on that space [43]. This approach has also been explored by Diks and Panchenko [29], and has recently been popularized by Gretton and coworkers in the context of two sample problem [40]. We present a family of (strictly) positive definite kernels on the space of spike trains, and dissimilarities (divergences) measures induced by them. We show that this framework encompasses many well known measures of dissimilarity e.g. measures based on intensity functions and count statistics. We also show that a class of positive definite kernels on spike trains introduced before are actually strictly positive definite, thus, explaining their superior performance over other kernels [84].

The rest of the chapter is organized as follows. In section 6.2 we discuss the notion of kernel based divergence and dissimilarity measures. In section 6.3 we discuss some special form of strictly positive definite kernels on $\mathbb{R}^n$. We later use these kernels to design kernels on spike train space. In section 6.5 and 6.6 we discuss a family of (strictly) positive definite kernel and their corresponding (divergence) dissimilarity measures using the stratified representation and the smoothed representation, respectively. In section 6.7 we provide some simulation results showing the performance of the proposed kernels.

6.2 Kernel Based Divergence and Dissimilarity

Let $(X, \mathcal{F})$ be a measurable space and $\mathcal{F} = \mathcal{B}(X)$ is the Borel algebra induced from the topology on $X$.

**Definition 7** (Positive definite kernel). Let $(X, \mathcal{F})$ be a measurable space, and $K : X \times X \to \mathbb{C}$ be a $(\mathcal{F} \otimes \mathcal{F} / \mathcal{B}(\mathbb{C}))$-measurable function. The kernel $K$ is called positive
definite if and only if for any finite non-zero complex Borel measure \( \mu : \mathcal{F} \to \mathbb{C} \),
\[
\iint K(x,y)d\mu(x)d\mu^*(y) \geq 0.
\]
Furthermore, if the inequality is strict, then the kernel \( K \) is called strictly positive definite.

Note that the definition does not ensure symmetry. Let \( \mathcal{M}_+ \) denote the set of probability measures on \((\mathcal{X}, \mathcal{F})\). Given a positive definite kernel \( K : \mathcal{X} \times \mathcal{X} \to \mathbb{C} \), we define a measure of dissimilarity \( \mathcal{D}_K : \mathcal{M}_+ \times \mathcal{M}_+ \to \mathbb{R}^+_0 \) as,
\[
\mathcal{D}_K(P, Q) = \iint K(x,y)d\mu(x)d\mu(y),
\tag{6–1}
\]
where \( \mu = P - Q \). Due to the positive definiteness of \( K \), \( \mathcal{D}_K(P, Q) \) is non-negative, and \( P = Q \) implies \( \mathcal{D}_K(P, Q) = 0 \).

A dissimilarity measure defined in such a way compares some particular features of \( P \) and \( Q \). For example, if \( \mathcal{X} = \mathbb{R} \) and \( K(x, y) = x \cdot y \), then \( \mathcal{D}_K(P, Q) = (E_P[X] - E_Q[Y])^2 \), i.e. the kernel \( K(x, y) = x \cdot y \) only compares the mean of the random variables \( X \sim P \) and \( Y \sim Q \).

A divergence measure \( \mathbb{D}_K(P, Q) : \mathcal{M}_+ \times \mathcal{M}_+ \to \mathbb{R}^+_0 \) can be constructed using \( (6–1) \) by incorporating a strictly positive definite kernel. Due to the strict positiveness of \( K \), \( \mathbb{D}_K(P, Q) \) is non-negative, and zero if and only if \( P = Q \).

As a matter of fact, one can show that the resulting divergence is a metric induced from the distance in the reproducing kernel Hilbert space (RKHS) if the positive definite kernel is Hermitian [29].

One of the major advantages of \( (6–1) \) is the simplicity of its estimator.

**Definition 8 (Estimator).** Given samples \( \{x_i\}_{i=1}^{N_P} \) and \( \{y_j\}_{j=1}^{N_Q} \) from \( P \) and \( Q \) respectively, the estimator for \( \mathcal{D}_K(P, Q) \) can be defined as,
\[
\hat{\mathcal{D}}_K(P, Q) = \sum_{i=1}^{N_P} \sum_{j=1}^{N_P} K(x_i, x_j) - 2 \sum_{i=1}^{N_P} \sum_{j=1}^{N_Q} K(x_i, y_j) + \sum_{i=1}^{N_Q} \sum_{j=1}^{N_Q} K(y_i, y_j) \tag{6–2}
\]
This estimator is consistent i.e. \( \hat{\mathcal{D}}_K(P, Q) \xrightarrow{a.s.} \mathcal{D}_K(P, Q) \).
While (6–1) provides a large class of divergences, instead of using a fixed kernel $K$, which, casually speaking, compares two probability laws on the entire sample space, it is possible to use a kernel that compares two probability laws only on their support. This motivates us to design a measure dependent divergence as follows.

**Definition 9** (Measure dependent divergence). A measure dependent divergence is defined as,

$$
\mathbb{D}_K(P, Q) = \int \int K_\xi(x, y)d\mu(x)d\mu(y),
$$

where $\mu = P - Q$, $\mu \ll \xi$ (absolute continuity), $\xi$ depends on $P$ and $Q$, and $K_\xi$ is a strictly positive definite kernel for all non-zero measures absolutely continuous with respect to $\xi$ (c.f. definition 7).

It directly follows from the definition that (6–3) is a divergence. For example, if $\xi = \frac{P + Q}{2}$, then the induced divergence compares the difference between $P$ and $Q$, only on the support on $P + Q$ instead of the entire space.

6.3 Strictly Positive Definite Kernels on $\mathbb{R}^n$ and $L_2$

In $\mathbb{R}^n$, strictly positive definite kernels are well studied [37]. We briefly discuss a few families of strictly positive definite kernels of special interest.

6.3.1 Composite Kernels on $\mathbb{R}^n$

**Definition 10** ($\mathcal{S}$-admissible functions). Let $\mathcal{S}$ be the Schwarz space of rapidly decreasing functions. $g : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ is $\mathcal{S}$-admissible if for every $h \in \mathcal{S}$, the following integral equation has a solution $f$, \[\int g(x, u)f(u)du = h(x).\]

**Theorem 14** (Composition kernels). If $g : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{C}$ is a $\mathcal{S}$-admissible or a strictly positive definite function, $\xi$ is a measure, and $\text{supp}(\xi) = \mathbb{R}^n$, then the following kernel is strictly positive definite on $\mathbb{R}^n$, \[K(x, y) = \int g(x, u)g^*(y, u)d\xi(u).\]

[Proof at the end of chapter]

Examples of basis functions $g(x, u)$ used for composition kernels are $e^{ixu}$, $\mathbb{I}(x \leq u)$, and $e^{-(x-u)^2}$ for $\mathbb{R}$. In $\mathbb{R}^n$, we can use the tensor product kernels: $\prod_i g(x_i, u_i)$.
6.3.2 Schoenberg Kernels (or radial basis function) on $L_2$

**Definition 11** (Completely monotone functions [106]). A function $\phi : (0, \infty) \to \mathbb{R}$ is said to be completely monotone on $[0, \infty)$ if $\phi \in C^\infty(0, \infty)$, continuous at zero, and $(-1)^l \phi^{(l)}(r) \geq 0$, $l \in \mathbb{N}, r > 0$ where $\phi^{(l)}$ denotes the $l$-th derivative.

Examples of completely monotone functions on $[0, \infty)$ are $\alpha, e^{-\alpha x}, \frac{1}{(x+\alpha^2)^\beta}$ where $\alpha$ and $\beta$ are constants.

**Theorem 15** (Strictly positive definite Schoenberg kernels on $L_2$). If a function $\phi : [0, \infty) \to \mathbb{R}$ is completely monotone on $[0, \infty)$ but not a constant function then $K(x, y) = \phi(\|x - y\|^2)$ is strictly positive definite on $L_2$ where $\|\cdot\|$ denotes the $L_2$ norm. [Proof at the end of chapter]

6.4 Representation of Spike Trains and Point Process Spaces

6.4.1 Smoothed Spike Train Space

An alternative representation for the spike train space is to use an injective transform $S : \Omega \to L_2(\mathfrak{X})$ from the space of spike trains to the space of $L_2$ integrable continuous functions with finite number of discontinuities [44, 84]. Given a spike train $\omega = \{t_i\}_i \in \Omega_n$, this is usually achieved by the linear smoothing $\omega \mapsto \lambda(t) = \sum_{i=1}^n g(t - t_i)|x$ where $g$ is the impulse response of the smoothing filter [84, 121]. The function $g$ is often taken to be strictly positive definite in the sense of Bochner [10], since it makes $S$ injective.

6.5 Stratified Spike Train Kernels

The stratification introduced in section 6.3 provides an opportunity to build kernels $\Omega$ by combining kernels on $\mathbb{R}^n$.

**Theorem 16** (Stratified strictly positive definite kernel). Let $\{K^{(n)}\}_{n=0}^\infty$ be a family of (strictly) positive definite kernels $K^{(n)} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{C}$ for every $n \in \mathbb{N}$. Define,

$$K_s(\omega_i, \omega_j) = \begin{cases} K^{(n)}(\omega_i, \omega_j) & \text{if both } \omega_i, \omega_j \in \Omega_n, \\ 0 & \text{otherwise} \end{cases}$$

(6-4)

Then $K_s$ is a (strictly) positive definite kernel on spike trains,
Proof. By direct sum.

The corresponding dissimilarity (divergence) can be simplified as

\[
\mathbb{D}_s(P,Q) = \sum_{n=0}^{\infty} \int \int K^{(n)}(\omega_i, \omega_j) d\mu|_n(\omega_i) d\mu|_n(\omega_j)
\]

(6–5)

where \( \mu = P - Q \), and \( \cdot|_n \) denotes the restriction of the measure to \( \mathcal{F}_n \). Below, we present a few special cases of this type of kernels.

If \( K^{(n)}_1 = 1 \) for all \( n \in \mathbb{N} \), then the dissimilarity becomes,

\[
\mathbb{D}_1(P,Q) = \sum_{n=0}^{\infty} \int (P(\Omega_n) - Q(\Omega_n))^2 d\xi|_n(\omega_1^n) d\xi|_n(\omega_2^n)
\]

(6–6)

where \( F \) denotes the cumulative distribution function in \( \mathbb{R}^n \). Given \( \xi = \frac{P+Q}{2} \), it can be shown that the resulting divergence is the cumulative distribution function (CDF) based Cramér–von-Mises type divergence proposed in chapter 4.

Another interesting kernel is found by using the following composite kernels on \( \mathbb{R}^n \),

\[
K^{(n)}_s(\omega_i, \omega_j) = \int \mathbb{I}(\omega_i \leq \omega) \mathbb{I}(\omega_j \leq \omega) d\xi(\omega)
\]

where \( \xi \) is a positive measure, \( \text{supp}(\xi) = \Omega_n \), \( \mathbb{I}(\omega_i \leq \omega_j) = \prod_d \mathbb{I}(\omega_i^d \leq \omega_j^d) \), and \( \omega = [\omega^1, \ldots, \omega^d] \). The corresponding divergence is given by,

\[
\mathbb{D}_s = \sum_n \int \int \left[ \mathbb{I}(\omega_i^{(n)} \leq \omega^{(n)}_1) \mathbb{I}(\omega_j^{(n)} \leq \omega^{(n)}_2) \right] d(P - Q)|_n(\omega_1^{(n)}) d(P - Q)|_n(\omega_2^{(n)})
\]

(6–6)

where \( F \) denotes the cumulative distribution function in \( \mathbb{R}^n \). Given \( \xi = \frac{P+Q}{2} \), it can be shown that the resulting divergence is the cumulative distribution function (CDF) based Cramér–von-Mises type divergence proposed in chapter 4.

If \( g(x, u) \) is the delta function \( \delta(x, u) \) such that \( \int f(x) \delta(x, u) dx = f(u) \) for every continuous function \( f(x) \), then the induced divergence is given by,

\[
\mathbb{D}_{K_p}(P,Q) = \sum_n \int \left( P(\Omega_n)f^{(n)}_P(u) - Q(\Omega_n)f^{(n)}_Q(u) \right)^2 d\xi|_n(u)
\]

(6–7)
where \( f_P^{(n)} \) and \( f_Q^{(n)} \) are the (continuous) density functions in \( \Omega \), i.e. the induced divergence is just the \( L_2 \) distance between two density functions. In practice the \( \delta \) function can be replaced by a smoothing function \( g \) that satisfies the condition in theorem 14.

### 6.6 Kernels on Smoothed Spike Trains

A (strictly) positive definite kernel \( K(\omega_1, \omega_2) \) on \( \Omega \) can be defined using a (strictly) positive definite kernel \( \tilde{K}(\tilde{\lambda}_1, \tilde{\lambda}_2) \) on the \( L_2 \) space since,

\[
\int K(\omega_1, \omega_2) d\mu(\omega_1) d\mu(\omega_2) = \int \tilde{K}(\tilde{\lambda}_1, \tilde{\lambda}_2) \tilde{\mu}(\tilde{\lambda}_1) \tilde{\mu}(\tilde{\lambda}_2) \geq 0,
\]

where \( \tilde{\mu} \) denotes the measure in \( L_2 \) induced by the measure \( \mu \) in \( \Omega \) i.e. \( \tilde{\mu}(A) = \mu(\mathcal{S}^{-1}(A)) \) if \( A \subset \mathcal{S}(\Omega) \) or zero otherwise.

Paiva and coworkers proposed the following kernel on spike trains \[84]\.

\[
K_{L_2}(\omega_1, \omega_2) = \int \tilde{\lambda}_1(t) \tilde{\lambda}_2(t) dt,
\] (6-8)

where \( \tilde{\lambda} \) is obtained by linear smoothing (see section 5.4.1). It is easy to see that this is a symmetric positive definite kernel. Therefore, the dissimilarity statistics \( D_{K_{L_2}} \) is not a divergence. \( K_{L_2} \) induces a spike train RKHS where the van Rossum like distance \[121]\ is induced by the inner product.

Direct sum of (strictly) positive definite kernels defines a (strictly) positive definite kernel. This motivates the following construction.

**Theorem 17.** Given a bounded strictly positive definite function \( \kappa : \mathbb{R} \to \mathbb{R} \), and a positive function \( h : \mathbb{R} \to \mathbb{R}^+ \), the kernel \( K_\kappa : \Omega \times \Omega \to \mathbb{R} \) is strictly positive definite.

\[
K_\kappa(\omega_1, \omega_2) = \int h(t) \kappa(\tilde{\lambda}_1(t) - \tilde{\lambda}_2(t)) dt
\] (6-9)

**Proof.** It is easy to see from the following form.

\[
\int \int K_\kappa(\omega_1, \omega_j) d\mu(\omega_i) d\mu(\omega_j) = \int h(t) \left[ \int \int \kappa(\tilde{\lambda}_{\omega_i}(t) - \tilde{\lambda}_{\omega_j}(t)) d\mu(\omega_i) d\mu(\omega_j) \right] dt.
\]
As a special case, the nonlinear kernel of the following form proposed in [84] is strictly positive definite.

\[ K_a(\omega_1, \omega_2) = \int_x \exp \left\{ -\left( \tilde{\lambda}_1(t) - \tilde{\lambda}_2(t) \right)^2 \right\} \, dt. \tag{6-10} \]

Finally, following theorem 15, any kernel of the form,

\[ K(\omega_1, \omega_2) = \phi(-\|\tilde{\lambda}_1(t) - \tilde{\lambda}_2(t)\|^2), \tag{6-11} \]

where \( \phi(\cdot) \) is completely monotone function and \( \|\cdot\| \) is the \( L_2 \) norm, is strictly positive definite. In [84], the authors have also proposed the following special case,

\[ K_b(\omega_1, \omega_2) = \exp \left\{ -\int \left( \tilde{\lambda}_1(t) - \tilde{\lambda}_2(t) \right)^2 \, dt \right\}. \tag{6-12} \]

### 6.7 Simulation Results

#### 6.7.1 Kernel PCA

<table>
<thead>
<tr>
<th>( \theta )</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>( K_b )</td>
<td>0.0289</td>
<td>0.4367</td>
<td>0.9999</td>
<td>0.9999</td>
</tr>
<tr>
<td>( K_a )</td>
<td>0.1072</td>
<td>0.9893</td>
<td>0.9999</td>
<td>0.9999</td>
</tr>
<tr>
<td>( K_s )</td>
<td>0.0367</td>
<td>0.0782</td>
<td>0.7623</td>
<td>0.9999</td>
</tr>
</tbody>
</table>

Table 6-1. Statistical power estimated from 1000 Monte Carlo runs for each kernel given 40 spike trains.

When a hermitian strictly positive definite kernel is used, the kernel based divergences between two probability laws can be seen as the distance between the corresponding mean elements in the RKHS [117]. Therefore, the performance of a specific kernel can be visualized by the separation of the mean element in the RKHS. However, since it is not possible to visualize the mean element in the infinite dimensional RKHS, we perform kernel principal component analysis (KPCA) and use the first two eigenvectors for visualization [107].
Figure 6-1. KPCA results of different strictly positive definite kernels are visualized on the first two principal components. $K_b$ is global, $K_a$ is local, and $K_s$ is stratified (see text for details). 100 spike trains each from two renewal processes with fixed rate Gamma distributed intervals are generated (top). $\theta$ is the shape parameter of the Gamma distribution; $\theta = 1$ corresponds to the Poisson process. One of the two classes was always fixed to be Poisson; the first column corresponds to the indistinguishable case. The diamond and circle are the mean of each class. The mean number of spikes is set to 10.

We compare the performance of the following three kernels; $K_b$ (6–12), $K_a$ (6–10) and $K_s$ (6–4). For $K_b$, we have obtained the smoothing using heavyside function i.e. $\tilde{\lambda}(t) = \sum_i \mathbb{I}(t \geq t_i)$, whereas for $K_a$ the smoothing is obtained using rectangular window i.e. $\tilde{\lambda}(t) = \sum_i \mathbb{I}(t_i - \tau \leq t \leq t_i + \tau)$. Therefore, while $K_b$ remembers all the history over time, $K_a$ is only sensitive to local events within the window. For $K_s$ we have taken $K^{(n)}$ to be an $n$-dimensional spherical Gaussian.

Since a strictly positive definite kernel is characteristic [117], the mean in the RKHS for each class is unique whenever the generating point processes are distinct (the right 3 columns in the figure, the first column is for identical point processes) [117].
Figure 6-2. Hypothesis test on two action potentials that are correlated ($H_0$) and independent ($H_1$). (Left) Spike trains from the null and alternate hypothesis. (Right) Comparison of the power of each method. Y-axis indicates the statistical power that measures the type 2 error given a fixed type 1 error ($\alpha = 0.05$) as estimated via surrogate test. Various kernels are compared; see test of figure 6-1 for details. For $\lambda_{L2}$ the number indicates the bin size for rate estimation.

However, the KPCA projections show that the representation significantly differs between different kernels. We emphasize that it is not necessary for the kernel to cluster the sample points to derive a good divergence measure; only the distance between the mean matters. We show the performance of the KPCA in figure 6-1 and report the power of the corresponding hypothesis test in table 6-1. We observe that each kernel has high discriminating power on this dataset.

### 6.7.2 Statistical Power

In this section, we empirically compare the small sample power of the proposed divergences against a baseline dissimilarity ($\lambda_{L2}$) which computes the $l^2$ distance between PSTH (peri-stimulus time histogram). We generate two classes of spike trains and test whether they follow the same probability law. Both point processes share same marginal intensity function but they have different correlation structure (Fig. 6-2 left). In the first process ($H_0$ in figure), the two event timings are correlated; the interspike interval (ISI)
has a narrow distribution. In the second process ($H_1$ in figure), the two event timings are independent; each action potential has a precise timing. Both point processes have a lossy noise; each action potential has a probability of disappearing with probability $p = 0.1$. Note that $H_0$ represents renewal type of neural code, while process $H_1$ represents precisely timed action potentials.

Since the highest dimension of the problem is 2, this problem is easier for the stratified kernels (Fig. 6-2 right). Nevertheless, the smoothed kernel based divergences quickly catch up as the number of sample increases. The firing rate function based dissimilarity ($\lambda_{L2}$) performs poorly, since the intensity function is designed to be identical.

6.8 Discussion

In this chapter, we have introduced a family of (strictly) positive definite kernels on the space of spike trains to design dissimilarity (divergence) between point processes. We have shown that the proposed kernels often lead to popular dissimilarity (divergence) measures that are widely used in practice, and that, many existing kernels are actually strictly positive definite in nature. However, given this collection of kernels, it is a natural question to ask, which kernel is better. In the following paragraphs, we briefly discuss the pros and cons of the proposed kernels.

The kernels defined on the stratified space are more efficient to compute, and easy to design. However, this type of kernels suffer from small sample sizes, since if the count distribution is flat (not concentrated) i.e the samples are scattered in different partitions $\Omega_n$, the divergence become difficult to evaluate. In addition, the stratified approach suffers from the curse of dimensionality, when the number of spikes in a spike train is large. Hence, it may be necessary to reduce the time window of observation to reduce the dimension. On the other hand, the kernels defined on the smoothed space do not suffer from these problems, since these kernels can compare two spike trains involving different spike number of spikes. However, on the contrary, these kernels are rather difficult to compute due to the integration required for evaluating the kernel.
In the smoothed representation, we have introduced three different kernels. The first kernel (6–8) is simply the inner product between the smoothed spike trains. This kernel is easier to evaluate than the other kernels on this space. However, this kernel is a positive definite kernel (not strictly) and the corresponding dissimilarity measure can be expressed as the $L_2$ distance between two intensity functions [84]. The second kernel (6–10) is a strictly positive definite kernel. However, this kernel is difficult to evaluate since it requires evaluating an integral which might be expensive. The final kernel (6–12) provides a compromise between performance and computational cost since the kernel is strictly positive definite but the computational cost is similar to that of the inner product kernel. Notice that the computational cost of these kernels can be reduced by choosing appropriate smoothing functions $g$ such as heavyside function or rectangular function.

In Table 6-2, we provide a brief description of the proposed method along with their complexity and relation to the literature.

6.9 Proof of Theorem 14

First we show that it is positive definite.

$$
\mathbb{D}_K(P, Q) = \iint K(x, y) d\mu(x) d\mu^*(y) = \iint \left( \int g(x, u) g^*(y, u) d\xi(u) \right) d\mu(x) d\mu^*(y)
$$

$$
= \int \left( \int g(x, u) d\mu(x) \right)^2 d\xi(u) \geq 0
$$

To show the strictly positive definiteness, we need to show when $d_K(P, Q) = 0$, $\mu = 0$. Suppose $d_K(P, Q) = 0$, then,

$$
\int \left( \int g(x, u) d\mu(x) \right)^2 d\xi(u) = 0 \iff \int g(x, u) d\mu(x) = 0 \quad \xi\text{-a.e. } u
$$

and, since supp($\xi$) = $\mathbb{R}^n$, this holds everywhere. If $g$ is strictly positive definite, $\iint g(x, u) d\mu(x) d\mu(y) = 0$ implies $\mu = 0$. In the case where $g$ is $\mathcal{G}$-admissible, we multiply both sides by an arbitrary function $f(u)$ and integrate.

$$
\int f(u) \int g(x, u) d\mu(x) du = 0 \Rightarrow \int \left( \int g(x, u) f(u) du \right) d\mu(x) = 0
$$
<table>
<thead>
<tr>
<th>eq#</th>
<th>Kernel</th>
<th>divergence?</th>
<th>time complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>(6–8)</td>
<td>$K = \int_x \tilde{\lambda}_1(t) \tilde{\lambda}_2(t) dt$</td>
<td>no</td>
<td>$O(N^2M^2)^a$</td>
</tr>
<tr>
<td></td>
<td>smoothed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(6–10)</td>
<td>$K = \int_x \exp \left{ - \left( \tilde{\lambda}_1(t) - \tilde{\lambda}_2(t) \right)^2 \right} dt$</td>
<td>yes</td>
<td>$O(N^2T)$</td>
</tr>
<tr>
<td>(6–12)</td>
<td>$K = \exp \left{ - \int \left( \tilde{\lambda}_1(t) - \tilde{\lambda}_2(t) \right)^2 dt \right}$</td>
<td>yes</td>
<td>$O(N^2M^2)^a$</td>
</tr>
<tr>
<td>(6–8)</td>
<td>$K = \int_x \tilde{\lambda}^S_1(t) \tilde{\lambda}^S_2(t) dt$</td>
<td>–</td>
<td>$O(N^2T)$</td>
</tr>
</tbody>
</table>

| stratified | $K^{(n)} = 1$               | no          | $O(NM)$                   |
| stratified | $K^{(n)}_\chi = \int \chi(x) \chi(x) dx$                        | yes         | $O(N^2M^2)$               |
| stratified | $K^{(n)}_p = \int \kappa(x-u) \kappa(y-u) du$                 | yes         | $O(N^2M^2)$               |

$^a$ using an exponential smoothing kernel

$^b$ nonlinear smoothing

$^c$ presented in the context of independence

Table 6-2. List of kernels for spike trains of special interest and their corresponding time complexity. The worst case time complexity is presented where $N = \max(N_P, N_Q)$, $M$ is the maximum number of spikes in all the samples, and $T$ is the number of numerical integration steps. Note that the stratified kernels can be more efficiently evaluated because the spike trains with different number of events are not compared. Although we refer to the literature (marked with †) for the kernels, they were not explicitly proven to be strictly positive definite before to the best of our knowledge.

For any $h \in \mathcal{S}$, $\int g(x,u)f(u)du = h(x)$ has a solution $f$, therefore, $\int h(x)d\mu(x) = 0$. Since, $h(x) \in \mathcal{S}$, the above condition implies that $\mu = 0$. This follows from the property of the Schwartz space.

### 6.10 Proof of Theorem 15

Since $\phi$ is a completely monotone function, it has a Laplace transform representation,

$$\phi(x) = \int_0^\infty e^{-xt}dm(t)$$

where $m$ is a finite positive Borel measure. Let $\{\psi_k\}$ be an orthonormal basis of $L_2$. We can represent any function in $f \in L_2$ as $f = \sum_j a_j \psi_k$. Then, for any set of $\{a_l\} \subset \mathbb{C}$ and
\{f_i\} \subset L_2,$

$$
\sum_l \sum_k \alpha_l \alpha_k^* \phi(\|f_l - f_k\|^2) = \int_0^\infty \sum_l \sum_k \alpha_l \alpha_k^* \exp(-t\|f_l - f_k\|^2)dm(t)
= \int_0^\infty \sum_l \sum_k \alpha_l \alpha_k^* \exp(-t\|a_l - a_k\|^2)dm(t) \geq 0
$$

where $a$ is a vector of coefficients $\{a_j\}$ and the inequality is from the fact that Gaussian is a positive definite function. It remains to show that the strict equality implies that all $\alpha_l$ are zero. Suppose the equality holds, that is,

$$
\sum_l \sum_k \alpha_l \alpha_k^* \exp(-t\|f_l - f_k\|^2) = 0 \quad m\text{-a.e.} t
$$

Take a sequence of independent zero mean unit variance Gaussian random variables $\{x_j\}$. Due to the Fourier transform property of Gaussian,

$$
E\left[ \exp \left( i \sum_j a_j x_j \right) \right] = \exp \left( -\frac{1}{2} \sum_j a_j^2 \right).
$$

$$
\Rightarrow \sum_l \sum_k \alpha_l \alpha_k^* \exp \left( -\frac{1}{2} \sum_j (a_l^j - a_k^j)^2 \right) = E \left[ \sum_k \alpha_k \exp \left( i \sum_j a_k^j x_j \right) \right]^2 = 0.
$$

Hence,

$$
\sum_k \alpha_k \exp \left( i \sum_j a_k^j x_j \right) = 0 \quad a.e.
$$

By taking the conditional expectations,

$$
\sum_k \alpha_k \exp \left( i a_k^j x_1 \right) E \left[ \exp \left( i \sum_{j=2} a_k^j x_j \right) \right] = \sum_k \alpha_k \beta_k \exp \left( i a_k^j x_1 \right) = 0 \quad a.e..
$$

Multiplying both sides with the conjugate and taking expectation we get,

$$
\sum_l \sum_k \alpha_l \beta_l \alpha_k^* \beta_k \exp \left( \frac{1}{2} (a_l^j - a_k^j)^2 \right) = 0
$$
From the strictly positive definiteness of Gaussian kernel, and from the fact that $\beta_k > 0$, we conclude that $\alpha_k = 0$ for all $k$. 
CHAPTER 7
CONCLUSION

The non-stationary, non-Poisson, and heterogeneous nature of trial-to-trial variability in the spike train observation calls for flexible analysis tools for measuring similarities. Instead of assuming a parametric structure on spike trains, our goals is to develop statistical divergences to directly measure the similarity between point processes, connecting the fundamental concepts of noise and similarity.

Divergences are direct extensions of ITL concepts where the full description of probability laws are utilized instead of partial statistics – they measure distances in the space of point processes. Divergence serves as a corner stone to build statistical learning and inference. We posed several neuroscience problems such as change detection, clustering and parameter estimation as statistical inferences.

We proposed three divergence families and corresponding estimators. They are motivated from different principles and estimation strategies. The cumulative based divergence extends the K-S test through the use of CDF representation on the stratified space. The main advantage is the lack of free parameter, however using the empirical CDF without smoothing leads to low statistical power of the estimator. The cumulative based point process divergence has a low time complexity (figure 7-1). Due to stratification, it works best with small number of action potentials per trial. Therefore, it is suitable for searching massive datasets with relatively high number of trials and small number of action potentials per trial.

The Hilbertian metric and $\phi$-divergence brings an inner product structure to the space of point processes. We presented the Hellinger and symmetric $\chi^2$-divergences both of which are quite powerful (figure 7-2). While both the divergences can be estimated via Parzen windowing on the stratified space, the symmetric $\chi^2$-divergence allows a direct estimation of the likelihood ratio while maintaining the $L_2$ consistency. Although the density estimation based divergence estimation performs better in low dimensional
problems (less number of spikes), they suffer from slow convergence in high dimensional
data. In contrast, the direct estimation of likelihood ratio performs equally well regardless of the dimension; however, the computational complexity is much higher (the slowest, see figure 7-1). Being Hilbertian, the divergence is directly connected to a symmetric positive definite kernel on point processes, thus enabling a plethora of kernel algorithms.

The kernel based divergence (based on strictly positive definite kernels) approaches the problem directly in the sense that no intermediate statistical object such as cumulative distribution, probability density, nor likelihood ratio, is estimated. However, it turns out that quadratic divergence measures such as the extended C-M test, that is developed as cumulative based divergence, as well as the $L_2$ distance (in the case of mCI kernel) are special cases of kernel based divergence. The estimation is straight-forward, and has intermediate time complexity. When a symmetric kernel is applied, it can be interpreted as the distance between the mean in the RKHS; this is equivalent to the $D_{ED}$ criterion (Euclidean divergence) in the ITL framework. Moreover, the symmetric kernel induces a
Hilbertian metric on the point process space as well, but this is fundamentally different from the $\phi$-divergence subclass.

![Figure 7-2](image.png)

Figure 7-2. Comparison of statistical power among proposed methods. (left) PTST against equi-intensity Poisson process. (right) Renewal against equi-intensity Poisson process.

All the proposed measures are binless and asymptotically consistent. In theory, consistency means that given arbitrarily large number of trials, the estimator converges to the theoretical value. However, when the number of trials is finite, as is always the case in practice, we may have poor estimation. The binless property implies that it utilizes all the given information that would have been lost if binning were performed. For example, the direct estimation of entropy with binning require orders of magnitude larger number of trials for estimation compared to binless methods [123, 125].

Each proposed estimator has different behavior in the small sample size regime. Generally speaking, the stratified kernels are more sensitive to the change in count distributions, and the non-stratified kernel-based methods are more sensitive to firing rate profile changes. In addition, the choice of parameters such as kernel type and kernel size creates complexity in using the divergence. Applying multiple divergences or tuning the parameter can lead to incorrectly high rejection rate in hypothesis testing, and careful design of multiple-testing is necessary [6]. We leave this as an open problem. In practice, the proposed methods should be applied when the simple test on rate function fails.
One of the immediate extension of the proposed method is to generalize it to multiple spike trains. Pairwise measure is only useful when observing and analyzing a single spike train, and the need for multi-variate analysis is evident; hypothesis testing can be used to determine whether two point processes are statistically dependent or not. However, estimation of joint point process is a much more difficult problem. The stratified space approach is very likely to fail due to the combinatorial explosion of the number of stratum. Therefore, we should focus on developing kernels of the form $K : \Omega^d \times \Omega^d \rightarrow \mathbb{R}$. The simplest form is the product kernel, however it is possible to design other efficient kernels directly having the multi-dimensional form. We also leave this for future directions.

Another interesting future application of divergence measures is to use it as a cost function to obtain adaptive filtering and learning algorithms. In ITL, this approach has brought fruitful results such as robust adaptive filtering and information theoretic clustering \[101\]. These divergences can potentially improve the MSE or MLE based estimation process as well. Designing synaptic plasticity rules by optimizing the divergences is also an interesting avenue \[11, 16, 61, 119\].

In addition individual divergences provide different structure to the point process space. Being divergences, they behave similarly when the two point processes are very close to each other. However, when far apart, they emphasize different aspects of distance. In an optimization problem, when the optimum (zero divergence) is achieved, the choice of divergence would only alter the convergence rate and estimation quality. In contrast, if the optimum is not achievable, each divergence may give distinct answers even theoretically. Therefore, further investigation of each divergence is necessary as a future work.

The simulated examples clearly show that the divergences are superior to dissimilarities. However, the analysis of the real data sets shows that despite being a dissimilarity the $L_2$ distance of mean rate function (or equivalently the kernel dissimilarity with the mCI kernel) performs comparable to the proposed divergences. There are two explanations for this phenomena. First, the experimentalist’s bias plays a key role. We have not designed
the experiments to demonstrate our superiority, and the experimentalist is likely to discard the conditions or protocols that do not show significant change in the mean rate profile, thus creating datasets that can be mostly described by the mean rate function. Second, the change in the neural system often accompanies change in the mean firing rate. This is perhaps the reason why scientists still relies on firing rate and count statistic. It is undeniable that the mean firing rate contains significant amount of information. However, the point of distribution-free divergences is to analyze “additional” information that the scientists may be blind to when only using the first order statistic.

We propose a two stage approach to find evidence that the full statistic is needed for real data as a future work. For a target sensory system, collect as many trials as possible with slightest changes in the input, so that the output spike trains cannot be fully classified via mean rate function only. Then using a divergence, verify if it is possible to sub-classify the ambiguously classified results from the mean rate function. This may reveal novel tuning properties of neurons as well.

One last disadvantage of using divergence that we are aware of for hypothesis testing is that when the hypothesis is rejected, the divergence does not explicitly indicate which feature caused the rejection. This is because all aspects are merged into a single number representation. Therefore, it is necessary to further analyze the data to understand it better. One possible line of future work is to design a divergence that can be broken into small explicit features, hence naturally allowing hierarchical tests.

I would like to clarify which parts of this thesis are my original work (many of which are results of collaboration). The theoretical contribution of the dissertation is the development of mathematical frameworks for (1) Hilbert space representation of spike trains, (2) alternative point process formulation, (3) precisely timed point process model, (4) statistical inference through point process divergences, and (5) novel divergences and estimators for point process. In addition, theoretical connections among ITL, RKHS, and (strictly) positive definite functions are better understood.
The practical contributions are development of a set of data analysis and modeling tools for neuroscience applications. A few problems with real data are demonstrated through the thesis: (1) non-stationarity detection, (2) plasticity detection, (3) optimal parameter selection, (4) clustering and visualization of spike trains, and (5) clustering point processes. The estimators and algorithms described in the thesis are implemented in MATLAB\textsuperscript{TM} as an open source project called IOCANE and freely available on the web.\textsuperscript{1}

The general point process view provided by the current body of work describes what is the noise, and measures similarity based on the noise structure, but it tells little about what the signal is. However, we strongly believe that analyzing the noise is the first step towards better understanding the signal and its processing in the neural system. Our analysis framework is focused on dealing with non-stationary and arbitrary noise structure that is often observed in experiments, and provides principled statistical tools for analysis. We hope the proposed methods contribute to the neuroscience community and transform how we think about spike train similarities.

\textsuperscript{1} \url{http://www.cnel.ufl.edu/~memming/iocane/}
\url{http://code.google.com/p/iocane/}
APPENDIX A
REPRODUCING KERNEL HILBERT SPACES FOR INFORMATION THEORETIC LEARNING

The initial goal of ITL was to propose alternative cost functions for adaptive filtering [31]. Entropy characterizes the uncertainty in the error distribution, however, the difficulty of Shannon’s entropy resides on non-parametric estimation within the constraints of optimal filtering (e.g. smooth costs). For this reason we embraced [32] a generalization of Shannon’s entropy proposed by Alfred Rényi. It is a parametric family of entropies given by
\[ H_\alpha(X) = \frac{1}{1-\alpha} \log \int f^\alpha(x)dx, \quad \alpha > 0, \]
where \( f(x) \) is the PDF of the continuous random variable \( X \). It is easy to show that the limit of \( \alpha \to 1 \) yields Shannon’s entropy and [105] shows that the singularity is not essential.

There is a practical nonparametric estimator for the quadratic Rényi’s entropy (\( \alpha = 2 \)),
\[ H_2(X) = -\log \int f^2(x)dx = -\log E[f(x)], \quad (A-1) \]
i.e. is the \( -\log \) of the first moment of the PDF itself. Since the logarithm function is monotonic, the quantity of interest is just the argument of it.
\[ \mathcal{V}(X) = \int f^2(x)dx, \quad (A-2) \]

---

which is called the information potential\(^2\) (IP), so named due to a similarity with potential fields in physics [100].

A non-parametric asymptotically unbiased and consistent estimator for a given PDF \(f(x)\) is defined as [95]

\[
\hat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} \kappa(x, x_i),
\]

where \(\kappa(\cdot, \cdot)\) is called the Parzen window, or kernel. Here the Parzen kernel will be chosen as a symmetric non-negative definite function with non-negative value just like in kernel-based learning theory, such as the Gaussian, polynomial etc [37].\(^3\) Then by evaluating the expectation of the Parzen’s PDF approximation in (A–2), the integral can be directly estimated from the data as

\[
\hat{V}(X) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \kappa(x_i, x_j),
\]

where \(\{x_i\}_{i=1}^{N}\) is the data sample and \(N\) is the total number, which is the estimator for IP. The concept and properties of information potential (and its estimator) have been mathematically studied and a new criterion based on information potential has been proposed, called the MEE (Minimization Error Entropy), to adapt linear and nonlinear systems [31]. MEE serves as an alternative cost function to the conventional MSE (Mean Square Error) in linear/nonlinear filtering with several advantages in performance when the error distribution is not Gaussian. If we think for a moment we see the big difference between MSE and MEE: MSE is the second order moment of the data and MEE is the first moment of the PDF of the data. Since all the information contained in the random

\(^2\) Note that in previously published papers, we called the estimator for this quantity (A–4) as the information potential. In this paper, we generalize the concept and called the statistical descriptor behind it (A–2) as the IP and refer to (A–4) as the estimator of IP. The physical interpretation still holds.

\(^3\) Not all Parzen window kernels are nonnegative definite (e.g. rectangular kernel), and not all non-negative definite kernel is nonnegative valued
variable is represented in its PDF, we can expect better performance from the later than from MSE.

In information theory, mutual information is used to quantify the divergence between the joint PDF and the product of marginal PDFs of two random variables. Another well-known divergence measure is the Kullback-Leibler divergence \[.\] However, both are difficult to estimate in practice without imposing simplifying assumptions about the data. Numerical methods are required to evaluate the integrals. This IP and two divergence measures among PDFs, one based on their Euclidean distance and the other on Cauchy-Schwarz inequality, have been proposed to surpass these limitations \[.\]

Given two probability density functions \(f(x)\) and \(g(x)\), their Euclidean divergence is defined as

\[
D_{ED}(f, g) = \int (f(x) - g(x))^2 \, dx = \int f(x)^2 \, dx - 2 \int f(x)g(x) \, dx + \int g(x)^2 \, dx. \tag{A-5}
\]

The divergence measure based on Cauchy-Schwarz inequality is given by

\[
D_{CS}(f, g) = -\log \frac{\int f(x)g(x) \, dx}{\sqrt{\int f^2(x) \, dx \cdot \int g^2(x) \, dx}}. \tag{A-6}
\]

Notice that both \(D_{ED}(f, g)\) and \(D_{CS}(f, g)\) are greater or equal to zero, and the equality holds if and only if \(f(x) = g(x)\). Notice the form of the integrals. We have in both the first moment of each PDF, and a new term \(\int f(x)g(x) \, dx\) that is the first moment of the PDF \(g(x)\) over the other PDF \(f(x)\) (or vice versa) which is called the cross information potential (CIP) \[.\]. CIP measures the similarity between two PDFs as can be expected due to its resemblance to Bhattacharyya distance and other distances as explained in Gokcay and Principe \[.\]. CIP appears both in Euclidean and Cauchy-Schwarz divergence measures and if one substitutes \(g(x)\) by \(f(x)\) in CIP it becomes the argument of Rényi’s quadratic entropy. As expected all these terms can be estimated directly from data as in (A–4).
The IP estimator (A–4) can be interpreted in an RKHS. Indeed according to the Mercer’s theorem [72], any symmetric non-negative definite kernel function that is square integrable has an eigen-decomposition as \( \kappa(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}_\kappa} \), where \( \Phi(x) \) is the nonlinearly transformed data in the RKHS \( \mathcal{H}_\kappa \) induced by the kernel function and the inner product is performed in \( \mathcal{H}_\kappa \). Therefore, we can re-write (A–4) as
\[
\hat{V}(x) = \left\langle \frac{1}{N} \sum_{i=1}^{N} \Phi(x_i), \frac{1}{N} \sum_{j=1}^{N} \Phi(x_j) \right\rangle_{\mathcal{H}_\kappa} = \left\| \frac{1}{N} \sum_{i=1}^{N} \Phi(x_i) \right\|^2.
\]
Similar interpretations of the Cauchy-Schwarz divergence in \( \mathcal{H}_\kappa \) were developed in [53].

The RKHS \( \mathcal{H}_\kappa \) is data independent since the kernel is pre-designed and acts on individual data samples, which means that extra computation involving functional evaluations in \( \mathcal{H}_\kappa \) are required when statistical quantities are estimated. The example of the IP estimator is still simple and can exploit the kernel trick, but in general this may not be the case.

The difficulty is that the inner product structure of \( \mathcal{H}_\kappa \) is not translating the statistics of the data.

### A.1 RKHS Framework for ITL

From the various definitions in information-theoretic learning summarized above, we see that the most general quantity of interest is the integral of the product of two PDFs \( \int f(x)g(x)dx \) which we called the CIP. Therefore this will be our starting point for the definition of the ITL RKHS that will include the statistics of the input data in the kernel.

#### A.1.1 The \(L_2\) space of PDFs

Let \( \mathcal{E} \) be the set that consists of all square integrable one-dimensional probability density functions, i.e., \( f_i(x) \in \mathcal{E}, \forall i \in I \), where \( \int f_i(x)^2dx < \infty \) and \( I \) is an index set. We then form a linear manifold
\[
\left\{ \sum_{i \in I} \alpha_i f_i(x) \right\} \quad \text{(A–7)}
\]
for any countable $I \subset \mathbb{I}$ and $\alpha_i \in \mathbb{R}$. Complete the set in (A–7) using the metric
\[
\|f_i(x) - f_j(x)\| = \sqrt{\int (f_i(x) - f_j(x))^2 \, dx}, \quad \forall i, j \in \mathbb{I}
\] (A–8)
and denote the set of all linear combinations of PDFs and its limit points by $L_2(\mathcal{E})$. $L_2(\mathcal{E})$ is an $L_2$ space on PDFs. Moreover, by the theory of quadratically integrable functions, we know that the linear space $L_2(\mathcal{E})$ forms a Hilbert space if an inner product is imposed accordingly. Given any two PDFs $f_i(x)$ and $f_j(x)$ in $\mathcal{E}$, we can define an inner product as
\[
\langle f_i(x), f_j(x) \rangle_{L_2} = \int f_i(x)f_j(x) \, dx, \quad \forall i, j \in \mathbb{I}
\] (A–9)
Notice that this inner product is exactly the CIP. This definition of inner product has a corresponding norm of (A–8). Hence, $L_2(\mathcal{E})$ equipped with the inner product (A–9) is a Hilbert space. However, it is not an RKHS because the inner product is not reproducing in $L_2(\mathcal{E})$, i.e., the point evaluation of any element in $L_2(\mathcal{E})$ cannot be represented via the inner product between two functionals in $L_2(\mathcal{E})$. Next we show that the inner product (A–9) is symmetric non-negative definite, and by the Moore-Aronszajn theorem it uniquely defines an RKHS.

**A.1.2 RKHS $\mathcal{H}_V$ Based on $L_2(\mathcal{E})$**

First, we define a bivariate function on the set $\mathcal{E}$ as
\[
\mathcal{V}(f_i, f_j) = \int f_i(x)f_j(x) \, dx, \quad \forall i, j \in \mathbb{I}
\] (A–10)
In RKHS theory, the kernel function is a measure of similarity between functionals. Notice that (A–10) corresponds to the definition of the inner product (A–9) and the cross information potential between two PDFs, hence it is natural and meaningful to define the kernel function as $\mathcal{V}(f_i, f_j)$. Next, we show that (A–10) is symmetric non-negative definite in $\mathcal{E}$.

**Property 1 (Non-Negative Definiteness):** The function (A–10) is symmetric non-negative definite in $\mathcal{E} \times \mathcal{E} \rightarrow \mathcal{R}$. 

115
Proof. The symmetry is obvious. For any \( N \), any set \( \{f_1(x), f_2(x), \ldots, f_N(x)\} \subseteq \mathcal{E} \) and any not all zero real numbers \( \{\alpha_1, \alpha_2, \ldots, \alpha_N\} \), by definition we have

\[
\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \mathcal{V}(f_i, f_j) = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \int f_i(x) f_j(x) \, dx \tag{A-11}
\]

\[
= \int \left( \sum_{i=1}^{N} \alpha_i f_i(x) \right) \left( \sum_{j=1}^{N} \alpha_j f_j(x) \right) \, dx = \int \left( \sum_{i=1}^{N} \alpha_i f_i(x) \right)^2 \, dx \geq 0.
\]

Hence, \( \mathcal{V}(f_i, f_j) \) is symmetric non-negative definite, and it is also a kernel function. \( \square \)

According to the Moore-Aronszajn theorem, there is a unique RKHS, denoted by \( \mathcal{H}_V \), associated with the symmetric non-negative definite function (A–10). We construct the RKHS \( \mathcal{H}_V \) bottom-up. Since the bivariate function (A–10) is symmetric and non-negative definite, it also has an eigen-decomposition by Mercer’s theorem [72] as

\[
\mathcal{V}(f_i, f_j) = \sum_{k=1}^{\infty} \lambda_k \psi_k(f_i) \psi_k(f_j), \tag{A-12}
\]

where \( \{\psi_k(f_i), k = 1, 2, \ldots\} \) and \( \{\lambda_k, k = 1, 2, \ldots\} \) are sequences of eigenfunctions and corresponding eigenvalues of the kernel function \( \mathcal{V}(f_i, f_j) \) respectively. The series above converges absolutely and uniformly on \( \mathcal{E} \times \mathcal{E} \) [72].

Then we define a space \( \mathcal{H}_V \) consisting of all functionals \( \mathcal{G}(\cdot) \) whose evaluation for any given PDF \( f_i(x) \in \mathcal{E} \) is defined as

\[
\mathcal{G}(f_i) = \sum_{k=1}^{\infty} \lambda_k a_k \psi_k(f_i), \tag{A-13}
\]

where the sequence \( \{a_k, k = 1, 2, \ldots\} \) satisfies the following condition

\[
\sum_{k=1}^{\infty} \lambda_k a_k^2 < \infty. \tag{A-14}
\]

Furthermore we define an inner product of two functionals in \( \mathcal{H}_V \) as

\[
\langle \mathcal{G}, \mathcal{F} \rangle_{\mathcal{H}_V} = \sum_{k=1}^{\infty} \lambda_k a_k b_k, \tag{A-15}
\]

where \( \mathcal{G} \) and \( \mathcal{F} \) are of form (A–13), and \( a_k \) and \( b_k \) satisfy property (A–14).
It can be verified that the space $\mathcal{H}_V$ equipped with the kernel function (A–10) is indeed a reproducing kernel Hilbert space and the kernel function $\mathcal{V}(f_i, \cdot)$ is a reproducing kernel because of the following two properties:

1. $\mathcal{V}(f_i, f_j)$ as a function of $f_i(x)$ belongs to $\mathcal{H}_V$ for any given $f_j(x) \in \mathcal{E}$ because we can rewrite $\mathcal{V}(f_i, f_j)$ as

$$\mathcal{V}(f_i, \cdot)(f_j) = \sum_{k=1}^{\infty} \lambda_k b_k \psi_k(f_j), \quad b_k = \psi_k(f_i).$$

That is, the constants $\{b_k, k = 1, 2, \ldots\}$ become the eigenfunctions $\{\psi_k(f_i), k = 1, 2, \ldots\}$ in the definition of $\mathcal{G}$. Therefore, $\mathcal{V}(f_i, \cdot) \in \mathcal{H}_V, \forall f_i(x) \in \mathcal{E}.$

2. Given any $\mathcal{G} \in \mathcal{H}_V$, the inner product between the reproducing kernel and $\mathcal{G}$ yields the function itself by the definition (A–15)

$$\langle \mathcal{G}, \mathcal{V}(f_i, \cdot) \rangle_{\mathcal{H}_V} = \sum_{k=1}^{\infty} \lambda_k a_k b_k = \sum_{k=1}^{\infty} \lambda_k a_k \psi_k(f_i) = \mathcal{G}(f_i).$$

This is so called the reproducing property.

Therefore, $\mathcal{H}_V$ is an RKHS with the kernel function and inner product defined above.

By the reproducing property, we can re-write the kernel function (A–12) as

$$\mathcal{V}(f_i, f_j) = \langle \mathcal{V}(f_i, \cdot), \mathcal{V}(f_j, \cdot) \rangle_{\mathcal{H}_V}$$

(A–16)

$$\mathcal{V}(f_i, \cdot) : f_i \mapsto [\sqrt{\lambda_k} \psi_k(f_i)], \quad k = 1, 2, \ldots$$

The reproducing kernel linearly maps the original PDF $f_i(x)$ into the RKHS $\mathcal{H}_V$.

We emphasize here that the reproducing kernel $\mathcal{V}(f_i, f_j)$ is deterministic and data-dependent by which we mean the norm of transformed vector in the RKHS $\mathcal{H}_V$ is dependent on the PDF of the original random variable because

$$\|\mathcal{V}(f_i, \cdot)\|^2 = \langle \mathcal{V}(f_i, \cdot), \mathcal{V}(f_i, \cdot) \rangle_{\mathcal{H}_V} = \int f_i(x)^2 dx.$$  

This is very different from the reproducing kernel $\kappa(x, y)$ used in kernel-based learning theory. The norm of nonlinearly projected vector in the RKHS $\mathcal{H}_\kappa$ does not rely on the
statistical information of the original data since

$$\|\Phi(x)\|^2 = \langle \Phi(x), \Phi(x) \rangle_{\mathcal{H}_\kappa} = \kappa(0)$$

if we use translation-invariant kernel functions [37]. Moreover, if \( X \) is a random variable, \( \Phi(X) \) is also a random variable in the RKHS \( \mathcal{H}_\kappa \). The value of \( \kappa(0) \) is a constant regardless of the original data. Consequently, the reproducing kernel Hilbert spaces \( \mathcal{H}_\nu \) and \( \mathcal{H}_\kappa \) determined by \( \mathcal{V}(f_i, f_j) \) and \( \kappa(x, y) \) respectively are very different in nature.

A.2 Connection Between ITL and Kernel Methods via RKHS \( \mathcal{H}_\nu \)

Figure A-1. The relationship among the sample space, PDF space, the proposed ITL RKHS \( \mathcal{H}_\nu \) and the RKHS \( \mathcal{H}_\kappa \). The sample space and \( \mathcal{H}_\kappa \) are connected via the nonlinear transformation \( \Phi_X(\cdot) \). The PDF space and \( \mathcal{H}_\nu \) are connected via the feature map \( \mathcal{V}(f, \cdot) \). A realization of a PDF in PDF space corresponds to a set of points in the sample space. The ensemble average of functionals in \( \mathcal{H}_\kappa \) corresponds to one functional in \( \mathcal{H}_\nu \). The kernel methods and ITL are related via the Parzen window. (reprinted from [127] with permission © 2006 IEEE)

In this section, we connect ITL and kernel methods via the RKHS framework. As we have mentioned in the previous section, because the RKHS \( \mathcal{H}_\kappa \) is induced by the data-independent kernel function, therefore the nonlinearily projected data in \( \mathcal{H}_\kappa \) is still stochastic and statistical inference is required in order to compute quantities of interest. For instance, in order to compute the statistics of the functionals, the mean and covariance are required. The expected value of functionals in the RKHS \( \mathcal{H}_\kappa \) is defined as \( E[\Phi(x)] \). The cross-covariance is defined as a unique operator \( \Sigma_{XY} \) such that for any
functionals \( f \) and \( g \) in \( \mathcal{H}_\kappa \)

\[
\langle g, \Sigma_{XY} f \rangle_{\mathcal{H}_\kappa} = E[g(y)f(x)] - E[g(y)]E[f(x)] = \text{Cov}[f(x), g(y)].
\]

(A–17)

The mean and cross-covariance operators as statistics of functionals in \( \mathcal{H}_\kappa \) become intermediate steps to develop algorithms such as the the maximum mean discrepancy (MMD) [40], the kernel independent component analysis (Kernel ICA) [5] and others. However, the proposed ITL RKHS \( \mathcal{H}_V \) is based on the CIP (integral of product of PDFs), therefore the transformed functional in \( \mathcal{H}_V \) is deterministic and only algebra is needed to carry out statistical inference in ITL RKHS. Hence our proposed RKHS offers simplicity and elegance in dealing with data statistics.

The RKHS \( \mathcal{H}_\kappa \) and the RKHS \( \mathcal{H}_V \) are related via the expectation operator. In order to justify this statement, Parzen’s non-parametric asymptotically unbiased and consistent PDF estimator (A–3) is employed to estimate those PDFs used in the ITL descriptors [52]. The Parzen window evaluates the PDFs in the sample space. Provided one chooses a non-negative definite kernel function as the Parzen window, it connects the RKHS \( \mathcal{H}_V \) to the RKHS \( \mathcal{H}_\kappa \) used in the kernel methods. As illustrated in Fig. A-1, the feature map \( \Phi(x) \) nonlinearily projects the sample space into a stochastic RKHS \( \mathcal{H}_\kappa \). Alternatively the feature map \( \mathcal{V}(f, \cdot) \) transforms the PDF space into a deterministic RKHS \( \mathcal{H}_V \). Hence the stochasticity is implicitly embedded into the feature map, and immediate algebraic operation can be applied to compute statistics. However the \( \mathcal{H}_\kappa \) methodology has to rely on intermediate steps by defining mean and covariance operators.
APPENDIX B
POISSON PROCESS

While there are many equivalent approaches to define and describe point processes – we just discussed the complete intensity function and random measure – we will use the counting process representation which is the most intuitive for people who are already exposed to random variables.

Let $\mathcal{B}$ be a Borel space. Let $N(B)$ be a non-negative valued measure for $B \in \mathcal{B}$. Let $\mathbb{B}$ be the set of all measures $N(\cdot)$. A point process is essentially a probability distribution on the set $\mathbb{B}$.

In case of usual temporal point processes, $\mathcal{B}$ is the Borel set of the real line and the realization $N([0, t))$ of the random measure signifies the number of events that occurred during the interval $[0, t)$. As a notation, we define $N(t) = N([0, t))$, and $N(s, t) = N(t) - N(s)$. The function $N(t)$ is a non-decreasing integer valued function with initial value 0.

Poisson process is the simplest and the most random point process. Most of the point processes are some form of generalization of the Poisson process. Also, several forms of general limit theorems indicates that the superposition of independent renewal process with some conditions asymptotically convergence to a Poisson process. There are many equivalent ways to define a Poisson process, here’s one.

**Definition 12** (Homogeneous Poisson process). A (homogeneous) Poisson process with the rate parameter $\lambda \in \mathbb{R}$ is defined by the following properties. When the limit of $\delta$ goes to 0,

\[
\Pr[N(t, t + \delta) = 1 | \mathcal{H}_t] = \lambda \delta + o(\delta) \quad \text{(B–1)}
\]
\[
\Pr[N(t, t + \delta) > 1 | \mathcal{H}_t] = o(\delta) \quad \text{(conditional orderliness) (B–2)}
\]
\[
\Pr[N(0) = 0] = 1 \quad \text{(initial condition) (B–3)}
\]

where $\mathcal{H}_t$ is any event in the interval $[0, t)$, and $o(\cdot)$ represents higher order terms.
The conditional orderliness guarantees that there can only be one event at a single time instance (the class of point processes that has this property is known as simple point process). Since the occurrence of an event does not depend on any of the history, Poisson process is memoryless.

A discrete time analogy of Poisson process is the Bernoulli trial process (in the sense that limit of $p \to 0$, with constraint $\lambda = pn$).

Let us informally derive the interval distribution of a Poisson process. The memoryless property can be restated in terms of the random variable of the interval time $\tau$ as follows,

$$\Pr[\tau > s + t | \tau > s] = \Pr[\tau > t] \quad (B-4)$$

which means that if there were no event up till $s$, the probability distribution that there will be no event during the next $t$ has the same distribution as the conditioned distribution. Let the complementary cumulative probability distribution function of $\tau$ be $F_c(t) = \Pr[\tau > t]$. Using the Bayes rule on (B-4),

$$\Pr[\tau > s + t, \tau > s] = \Pr[\tau > t] \Pr[\tau > s]$$

$$\Pr[\tau > s + t] = \Pr[\tau > t] \Pr[\tau > s]$$

$$F_c(s + t) = F_c(t)F_c(s)$$

The only continuous function that satisfies this condition is the exponential function. Therefore, the probability distribution $f(t)$ of the intervals $\tau$ is the exponential distribution.

Exponential interval distribution is the hallmark of continuous-time Markov chains. We will show that Poisson process is equivalent to a pure birth process. From the conditional orderliness assumption, the following probability decomposition is valid for
\( k \geq 1. \)

\[
\Pr[N(s, t + \delta) = k] = \Pr[N(s, t) = k] \Pr[N(t, t + \delta) = 0] \\
+ \Pr[N(s, t) = k - 1] \Pr[N(t, t + \delta) = 1] + o(\delta) \\
= \Pr[N(s, t) = k](1 - \lambda \delta + o(\delta)) \\
+ \Pr[N(s, t) = k - 1](\lambda \delta + o(\delta)) + o(\delta)
\]

The second equality is from the definition of Poisson process (B–1).

\[
\frac{\Pr[N(s, t + \delta) = k] - \Pr[N(s, t) = k]}{\delta} = \\
- \Pr[N(s, t) = k] \lambda + \Pr[N(s, t) = k - 1] \lambda + \frac{o(\delta)}{\delta} \tag{B–5}
\]

Taking the limit \( \delta \to 0^+ \),

\[
\frac{d\Pr[N(s, t) = k]}{dt} = -\lambda \Pr[N(s, t) = k] + \lambda \Pr[N(s, t) = k - 1] \tag{B–7}
\]

And in case of \( k = 0 \),

\[
\frac{d\Pr[N(s, t) = 0]}{dt} = -\lambda \Pr[N(s, t) = 0] \tag{B–8}
\]

These differential equations are a special case of pure birth process in continuous-time Markov process. Eq. (B–8) can be solved with initial condition given by the third property in the definition of Poisson process.

\[
\Pr[N(s, t) = 0] = e^{-\lambda(s-t)} \tag{B–9}
\]

which is equivalent to the interval distribution, because this is the probability that there is no event in the interval. By recursively solving Eq. (B–7), we get the Poisson distribution:

\[
\Pr[N(s, t) = k] = \frac{(\lambda(s - t))^k}{k!} e^{-\lambda(s-t)} \tag{B–10}
\]

In case of discrete time, this corresponds to the binomial distribution.

The homogeneous Poisson process is stationary in the following sense.
**Definition 13** (Stationary point process). A point process is stationary if the probability distribution of \( N(A) \) is invariant over the time shift of a Borel set \( A \).

An extension of the homogeneous Poisson process that makes it non-stationary is the **inhomogeneous Poisson process** where the rate \( \lambda \) depends on time.

**Definition 14** (Inhomogeneous Poisson process).

\[
\Pr[N(t, s) = k] = \frac{(\Lambda(t) - \Lambda(s))^k}{k!} e^{-(\Lambda(t) - \Lambda(s))} 
\] (B-11)

\[
\Pr[N(0) = 0] = 1 \quad \text{(initial condition)} 
\] (B-12)

\[
\Pr[N(A)], \Pr[N(B)] \text{ are independent if } A \cap B = \emptyset 
\] (B-13)

where \( \Lambda(\cdot) \) is a finite-valued, non-decreasing, and non-negative function.

Using Eq. (B-11) in the definition is slightly more general than saying,

\[
\Pr[N(t, t+\delta) = 1|\mathcal{H}_t] = \lambda(t)\delta + o(\delta) 
\]

where \( \Lambda(t) - \Lambda(s) = \int_s^t \lambda(\rho)d\rho \), because \( \Lambda(\cdot) \) might not be differentiable or even continuous, hence \( \lambda(t) \) may not be well defined. Whenever \( \Lambda \) has a discontinuity, the probability of having an event is 1, thus, the process can be decomposed into fixed events and a continuous \( \Lambda' \) when needed.

**Theorem 18** (Sample function density for an inhomogeneous Poisson process [116]). Let \( \Omega = \{W_1, W_2, \cdots, W_n, N(t) = n\} \subset [s, t) \) be a particular realization of a Poisson process \( N(\cdot) \) with rate function \( \lambda(\cdot) \). The probability of having a particular realization \( \Omega \) is given by,

\[
p(\Omega) = \begin{cases} 
\exp \left( - \int_s^t \lambda(\rho)d\rho \right), & N(s, t) = 0, \\
\prod_{i=0}^n \lambda(W_i) \exp \left( - \int_s^t \lambda(\rho)d\rho \right), & N(s, t) = n \geq 1. 
\end{cases} 
\] (B-14)

\[
= \exp \left[ - \int_s^t \lambda(\rho)d\rho + \int_s^t \ln(\lambda(\rho))N(\rho) \right] 
\] (B-15)
A large class of simple point processes can be uniquely determined in terms of the intensity function.

**Definition 15** (Intensity function).

\[
\lambda(t; \mathcal{H}_t) = \lim_{\delta \to 0^+} \frac{\Pr[N(t, t + \delta) > 0|\mathcal{H}_t]}{\delta}
\]  

(B–16)

The \( \mathcal{H}_t \) can include the history dependence, input dependence, internal state dynamics, and hidden stochastic process.

**Theorem 19** (Decomposition of Poisson process into two random variables). If \( \lambda \) is well defined, an inhomogeneous Poisson process for a fixed interval \([s, t)\) can be represented with two independent random variables. One is the Poisson random variable \( N(s, t) \), and the other is a continuous random variable \( X \) for the distribution of the location of the events. The probability distribution function of \( X \) is given by normalizing the intensity function \( \lambda(\tau) \) as follows,

\[
f_X(\tau) = \frac{\lambda(\tau)}{\int_s^t \lambda(\rho) d\rho}
\]  

(B–17)

A realization of the Poisson process is a collection of \( n \) independent realizations of \( X \), where \( n \) is a realization of \( N(s, t) \).

In case of homogeneous Poisson process, \( X \) follows uniform distribution, hence it is possible to easily generate realizations using a pseudo-random number generator which is assumed to be uniform.
APPENDIX C
MEASURE THEORY

A measure is a convenient mathematical object (function) that can represent the positions of strawberries in a field, distribution of water in the ocean, or probabilities of winning over the lottery numbers; the measure counts the number of strawberries in a given area, reports the amount of water in a certain sea, and evaluates the probability of a lottery ticket to win. This abstract unifying framework enables one to rigorously “measure” quantities over a space, and also enable integration. It also allows elegant notation for probability theory. Here we briefly describe key ideas of measure theory without proof. This material is mostly based on Daley and Vere-Jones [25], Halmos [42].

To define a measure, we need a measurable space \((\Omega, \mathcal{F})\); a non-empty set \(\Omega\) and a \(\sigma\)-algebra \(\mathcal{F}\) on \(\Omega\). Here \(\Omega\) is the space where our stuff to be measured lies, and \(\mathcal{F}\) gives special structure of the space such that things are well defined and pathological sets can be avoided. An algebra \(\mathcal{F}\) of \(\Omega\) is a set of subsets of \(\Omega\) such that it contains the empty set, and closed under set union and complement. A \(\sigma\)-algebra is an algebra that is closed under countable union. Elements in \(\mathcal{F}\) are said to be measurable.

A measure \(\mu\) on \((\Omega, \mathcal{F})\) is a non-negative extended real valued function on \(\mathcal{F}\) that is countably additive

\[
\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i}^{\infty} \mu(A_i),
\]

where \(A_i\) are disjoint measurable sets. Additivity makes sense because we want \(\mu(A) + \mu(B) = \mu(A \cup B)\) when \(A\) and \(B\) are disjoint (number of strawberries should add up for different fields). If \(\mu\) is always finite valued, then it is called a finite measure. For a special case, a probability measure is a finite measure where \(\mu(\Omega) = 1\).

Given a set \(F\), we denote the smallest \(\sigma\)-algebra that contains \(F\) as \(\sigma(F)\) and say \(\sigma(F)\) is the \(\sigma\)-algebra generated by \(F\). A measure \(\mu\) on \(\mathcal{F}\) is determined by its values on any algebra that generates \(\mathcal{F}\) (Carathéodory extension theorem).
A predicate \( P(x) \) holds almost everywhere \( \mu \) (or \( \mu \)-a.e. for short) if it is true except for a set of measure zero, that is, \( \mu(E) = 0, \forall x \in E^c, P(x) \).

A function \( f \) from a measurable space \((X, \mathcal{F})\) to a measurable space \((Y, \mathcal{G})\) is measurable if \( \forall E \in \mathcal{G}, f^{-1}(A) \in \mathcal{F} \). For a topological space \((X, \mathcal{U})\), the \( \sigma \)-algebra generated by the open sets is called the Borel \( \sigma \)-algebra (or Borel set). Borel set links the measurable functions and continuous functions – every continuous function from \( \Omega \) to the real line is measurable with respect to the Borel algebra. When the topological space is induced by a metric, Borel set and Baire set coincide. Baire set is the smallest \( \sigma \)-algebra with respect to which the continuous functions are measurable.

Real-valued Borel-measurable functions are closed under algebraic operations and limit. Moreover, they can be approximated by limit of simple functions. A simple function is a finite linear combination of indicator function of measurable sets. By the linearity of integration, integration of a Borel-measurable function with respect to a measure \( \mu \) can be defined by letting the integration of an indicator function on the measurable set \( A \) as \( \mu(A) : \int A d\mu = \mu(A) \).
REFERENCES


BIOGRAPHICAL SKETCH

Il “Memming” Park is a computational neuroscientist who aims to understand the information processing of human brain. He attended Korea Advanced Institute of Science and Technology (KAIST) where he received a Bachelor of Science degree in computer science in 2004. He pursued his graduate study with José C. Príncipe in the Computational NeuroEngineering Laboratory (CNEL) in University of Florida from Fall 2005. He received his Master of Science degree in electrical and computer engineering department in University of Florida in 2007 on efficient algorithms for binless spike train domain signal processing. He continued his study at CNEL in biomedical engineering department. He is known by his pseudonym MEMMING on the web. He was born in Germany in 1979.