ESTIMATION OF SYMMETRIC CHI-SQUARE DIVERGENCE FOR POINT PROCESSES

Il Park1, Sohan Seth2, Murali Rao3, José C. Príncipe2∗

Departments of Biomedical Engineering1, Electrical and Computer Engineering2,4, Mathematics3
University of Florida, Gainesville, Florida
{memming,sohan,principe}@cnel.ufl.edu1,2,4, rao@math.ufl.edu3

ABSTRACT

This paper addresses the estimation of symmetric χ²-divergence between two point processes. We propose a novel approach by, first, mapping the space of spike trains in an appropriate functional space, and then, estimating the divergence in this functional space using a least square regression approach. We compare the proposed approach with other available methods on simulated data, and discuss its pros and cons.

Index Terms— Point process, spike train, symmetric chi-square divergence, kernel method, hypothesis testing

1. INTRODUCTION

One of the most fundamental questions encountered in neuroscience applications, such as change detection and neural coding, is whether two sets of spike trains follow the same probability law. From a statistical perspective, this question can be framed as a hypothesis testing problem with the help of a suitable divergence measure. By definition, a divergence measure is a nonnegative statistic of two probability laws that achieves zero value if and only if the two probability laws are the same. However, traditional approaches toward this problem avoid estimating divergence between point processes (probability law over spike trains), and use simpler tests, such as Wilcoxon test (WT) on the number of action potentials, to judge the similarity of two point processes. These tests, however, inherently put some assumptions on the underlying point process, e.g. WT assumes that the count distribution is a sufficient descriptor of a point process. Such assumptions reduce the generality of the conclusions that can be drawn from these tests since in practice, two point processes might be similar in terms of, say, the mean firing rate or the count distribution, and can still be different in terms of the other statistics [8].

Therefore, a true divergence measure is absolutely necessary to address the problem of point process similarity.

Although, estimation of divergence has been elaborately studied in $\mathbb{R}^d$ [10], these methods cannot be readily applied to spike train spaces, since this space lacks basic algebraic and topological structures. Recently, however, [7] has proposed a method for estimating the Hellinger divergence between two point processes by representing the space of spike trains as a union of Euclidean spaces, also called strata, and then computing the divergence in each strata separately. This approach, although theoretically elegant, suffers in estimation since it distributes the spike train observations in strata, thus, making each strata sparse. In this paper, we address this issue in detail, and explore an alternate method of estimating divergence. We follow a different representation of the spike train space as introduced by [6] i.e. we represent each spike train as a function in a suitable functional space, and estimate the divergence in this functional space. We, however, address the problem of estimating symmetric χ²-divergence, rather than the Hellinger divergence as addressed by [7]. Given two probability laws $P$ and $Q$ on a measurable space $(\Omega, \Sigma)$, the symmetric χ² divergence is defined as

$$D_S(P, Q) = \int \left( \frac{d(P - Q)}{d(P + Q)} \right)^2 d(P + Q),$$ (1)

We consider this divergence since, like Hellinger divergence, it is a metric on the space of all probability measures [3], and as we will discuss later in detail, it can be efficiently estimated using a least square regression approach [4].

The rest of the paper is organized as follows; in section 2 we briefly discuss the stratification approach proposed by [7], and how it can be used to estimate the symmetric χ²-divergence. In section 3 we discuss the main result of the paper, and describe a novel approach to estimate the same divergence using the functional representation. In section 4 we compare these two approaches with other methods for hypothesis testing on simulated data, and in section 5 we conclude the paper with some discussion on the pros and cons of the proposed approach.

2. STRATIFIED REPRESENTATION

Let $\Omega$ be the set of all finite length spike trains, that is, each $\omega \in \Omega$ can be represented by a finite set of action potential timings $\{t_1, t_2, \ldots, t_n\}$ within $T$, the time interval of interest, where $n$ is the number of spikes. Then the non-Euclidean spike train space $\Omega$ can be partitioned in disjoint partitions
derivative can be taken as,

$$P_{\Omega}$$

The probability measure $P$ for point process can 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. Let $\mu_n(\Omega)$ denote the extension of Lebesgue measures $\mu_n$, for $\Omega_n(n > 0)$ to $\Omega$, and $\delta_n$ is a Dirac measure for the single element in $\Omega_0$ which represents the empty spike train. Assuming $P \ll \mu$, the Radon-Nikodym derivative can be taken as,

$$\frac{dP}{d\mu}(\omega) = P_0(\Omega)\delta_\Omega(\omega) + \sum_{n=1}^{\infty} \frac{dP_n}{d\mu_n}(\omega_n) = p_0\delta_\Omega(\omega) + \sum_{n=1}^{\infty} p_nf_n(\omega_n)$$

(2)

where $f_n$ is the unordered joint location density and is symmetric on the permutation of its arguments, and $\omega_n = \omega \cup \Omega_n$.

Given a sequence of observations $X = \{\omega_i\}_{i=1}^m$, we can use the decomposition (2) for estimating the Radon-Nikodym derivative as follows. 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 Parzen density estimation of $f_n$ can then be written as,

$$\hat{p}_n = \frac{|X^{(n)}|}{|X|}$$

and

$$\hat{f}_n(x) = \sum_{\omega_i \in X^{(n)}} \kappa_n(x - \omega_i; \sigma_n) / |X^{(n)}|$$

for $n = 1, \ldots, m$. where $\kappa_n(\cdot; \sigma)$ is a symmetric $n$-dimensional density estimation kernel with bandwidth parameter $\sigma$, and $\cdot \mid \cdot$ denotes the cardinality of a set.

The Radon-Nikodym derivative $d\mu/d\nu$ in $\mathbb{R}^n$ can be consistently estimated, under appropriate conditions, by the ratio of the Parzen estimates of the Radon-Nikodym derivatives $f_0^\nu = d\mu/d\mu$ and $f_\nu^\nu = d\nu/d\mu$, respectively, where $\mu$ denotes the Lebesgue measure [2]. Then, the estimator of the $\chi^2$-divergence becomes,

$$\hat{D}_{\chi^2} = \int \left( \frac{d\mu}{d\nu} - 1 \right)^2 d\nu$$

$$= \left( \frac{\hat{p}_0}{\hat{p}_\nu} - 1 \right)^2 \hat{p}_\nu + \sum_{n=1}^{\infty} \int \left( \frac{\hat{p}_n^\nu \hat{f}_n^\nu}{\hat{p}_\nu^\nu \hat{f}_\nu^\nu} - 1 \right)^2 d\nu_n$$

where $\hat{Q}_\nu$ denotes the empirical probability law. The symmetric $\chi^2$-divergence can then be computed as $D_{\chi^2}(P, Q) = D_{\chi^2}(P, (P + Q)/2) + D_{\chi^2}(Q, (P + Q)/2)$.

3. A NOVEL APPROACH

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. To resolve this issue, we follow a different approach of representing a spike train, as followed by [6].

Let $F$ be the space of all $L_2$ integrable functions over $T$ i.e. $F = L_2(T)$. Given $\omega = \{t_1, \ldots, t_n\}$, define a mapping $G: \Omega \rightarrow F$ as $G(\omega)(t) = \sum_{i=1}^{n} g(t, t_i)$ such that $G$ is injective. When $g$ is a translation invariant function that decays at $\infty$, $G$ can be considered a smoothed spike train. There are many different $g$’s that make the mapping $G$ injective e.g. a bounded strictly positive definite function [9]. We use the indicator function $I(t \geq t_i)$ for the experiments.

We consider the $\sigma$-algebra of $L_2(T)$ to be the one that makes the map $G: \Omega \rightarrow L_2(T)$ measurable. Since $G$ is measurable, we can define an induced probability measure $U$ such that $U(L_2(T) \setminus G(\Omega)) = 0$ and $U(G(A)) = P(A)$ for $A \in \sigma(\Omega)$. Let $U$ and $V$ be two probability laws on $F$ induced by $P$ and $Q$, respectively, then, the following two propositions show that the Radon-Nikodym derivative in $\Omega$ can be transferred to $F$.

**Proposition 1.** 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 2.** $dP/dQ(\omega) = dU/dV(G(\omega))$

**Proof.** For any integrable function $\phi : F \rightarrow \mathbb{R}$,

$$\int \phi(G(\omega))dP(\omega) = \int \phi(f)dU(f)$$

$$= \int_F \phi(f) dU(f) = \int_F \phi(G(\omega)) dU/G(\omega)dQ(\omega)$$

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

**Corollary 1.** The $\chi^2$-divergence between $P$ and $Q$ is the same as the $\chi^2$-divergence between $U$ and $V$, i.e.

$$\int_{\Omega} \left( \frac{dP}{dQ}(\omega) - 1 \right)^2 d\omega = \int_F \left( \frac{dU}{dV}(f) - 1 \right)^2 dU(f).$$

Therefore, in the transformation representation, we focus on estimating the Radon-Nikodym derivative in $F$ rather than in $\Omega$ i.e. to estimate $dU/dV$ from samples $\{f_i = G(\omega_i)\}_{i=1}^{m}$ and $\{g_i = G(\omega_i^2)\}_{i=1}^{m}$. Although this problem can be approached in several ways, we follow the approach suggested in [4].

Using the triangle inequality on the actual and the estimated divergence, we get,

$$D_{\chi^2} - \hat{D}_{\chi^2} = \left| \int (dU/dV - 1)^2 dV - \int (dU/dV - 1)^2 d\nu \right| \leq \left| \int (dU/dV - 1)^2 d(V - \nu) \right| + \left| \int (dU/dV - d\nu/dV)^2 dV \right|.$$

This inequality shows that the error between the actual and estimated divergence is bounded by two $L_2$ distances. The
first term goes to zero as \( n \to \infty \) since \( V \to V \) almost surely. Therefore, in order to get a consistent estimate of \( \mathbb{D}_2 \) it is important to get an appropriate estimate of Radon-Nikodym derivative that makes the second term arbitrarily close to zero.

Following [4], we assume that \( \frac{dU}{dV}(f) = \sum_{i=1}^{l} \alpha_i \tilde{k}(g - g_i) \) where \( \alpha_i \)'s are real coefficients and \( \tilde{k}(f - g) \) is a strictly positive definite kernel [1]. This expansion is justified due to the following proposition which states that the functions of the form \( \sum_{i=1}^{\infty} \alpha_i \tilde{k}(g - g_i) \) is dense in \( L_2(F, V) \) i.e. it can approximate any function \( p \in L_2(F, V) \) with arbitrary accuracy in the \( L_2 \) sense. Notice that since \( dU/dV \in L_2(F, V) \), this implies that the proposed expansion can approximate the Radon-Nikodym derivative arbitrarily.

**Proposition 3.** Let \( \tilde{k}(x, y) \) be a symmetric strictly positive definite continuous kernel on \( F \times F \) and \( V \) is a probability measure on \( F \) such that \( \int_F \tilde{k}^2(x, y) dV(x) < \infty \) for all \( y \in F \), then span \( \tilde{k}(x, y) : y \in F' \subset F \) is dense in \( L_2(F, V) \), where \( F' \) denotes the subset where the measure \( V \) lies.

**Proof.** Let \( \tilde{k} \) be the span of \( \tilde{k} \) in \( L_2(F, V) \), then \( \exists g \in L_2(F, V) \) such that \( \int F \tilde{k}(x, y) dV(y) = 0 \). Therefore, \( \int \int g(x) g(y) \tilde{k}(x, y) dV(x) dV(y) = 0 \). Since \( \tilde{k} \) is strictly positive definite, by contradiction \( g \) is zero a.e. \( V \). \( \Box \)

Using the kernel expansion, the second term in the triangle inequality can be expanded as,

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

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

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

where \( C \) is a constant. Therefore, the second term in the triangle inequality is minimized for \( \alpha \approx (l/m) (K_{QQ} + \lambda I)^{-1} K_{QP} \) where \( \lambda \) is a regularization parameter required to avoid overfitting and \( [K_{PQ}]_{ij} = \tilde{k}(G(\omega_i^1), G(\omega_j^2)) \) is the gram matrix. The estimated \( \alpha \) can then be used to estimate \( \hat{D}_{\chi^2} = (K_{QQ} - 1)^2 l / l \), and \( \mathbb{D}_S \) can again be estimated in a similar fashion. For simulation, we use the kernel \( \tilde{k}(f, g) = \exp \left( - \int_t (f(t) - g(t))^2 dt / \sigma^2 \right) \) from [6] after setting \( \sigma \) to the median of the intersample distances, and set \( \lambda = 1/n \).

Notice that, this approach allows two spike trains to always interact through an spd kernel, irrespective of their spike counts. However, it requires selecting an appropriate functional representation of the spike trains, and an appropriate kernel to evaluate the similarity between two spike trains. Although, in theory, the asymptotic performance of the proposed method does not depend on the choice of these parameters, its finite sample behavior does. Given a trivial kernel, therefore, we expect the performance of this method to improve over sample size, irrespective of the statistical nature of the underlying point process. We observe that our experimental results support this fact. The selection criteria of the best functional representation, and the best kernel, however, still remain an open issue.

### 4. RESULTS

In this section, we present 3 hypothesis testing experiments, and compare the performance of the proposed method against the stratification approach, and the \( L_2 \) distance \( (\lambda_{L_2}) \) between PSTH (peri-stimulus time histogram) based dissimilarity measure. 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. We use the permutation test to compute the statistical power, and set the size of the test to 0.05.

#### 4.1. Two action potentials

We first consider two point processes with two events or less with same marginal intensity function but with different event correlation structure (Fig. 1 left). In spike trains from \( H_0 \), the two timings are correlated, and the interspike interval (ISI) has a narrow distribution, whereas in spike trains from \( H_1 \), the two timings are independent, and 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 dimension of the problem is at most 2, this problem is easier for the stratified estimator (Fig. 1 right). Nevertheless, the kernel based estimator quickly catches up with the number of sample increases. \( \lambda_{L_2}, \) on the other hand, fails to discriminate the two processes because the intensity function is designed to be identical.

#### 4.2. Inhomogeneous Poisson process

Next, we consider two Poisson processes. Since the mean rate function is a sufficient descriptor of a Poisson process, \( \lambda_{L_2} \) can easily distinguish between two Poisson processes. We simulate two inhomogeneous Poisson process where the rate changes at 100 ms. In Fig. 2, \( \lambda_{L_2} \) with the right bin size outperforms the divergence measures. The stratified approach suffers from the spread of samples in different strata and the curse of dimensionality, while the kernel based estimator quickly approaches high power as more data is observed.

#### 4.3. Stationary renewal processes

Next, we consider a renewal process and a Poisson process. Renewal process is a widely used point process model to com-
Fig. 1. Two action potentials that are correlated ($H_0$) and independent ($H_1$). $\sigma_1 = 10$ 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.

Fig. 2. Poisson process with rate changing in step function from 3 to 2, and 2 to 3 at 100 ms. $\sigma_1 = 100$ 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.

Fig. 3. 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.

Compensate the deviation from Poisson process [5]. We simulate a stationary renewal process with gamma interval distribution. Despite the fact that the count distribution is widely distributed and the dimensionality of the problem is high, 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. However, interestingly in the finite sample case, the variance in the rate function depends on the regularity of the firing pattern, therefore the method is still able to detect the difference.

5. DISCUSSION

In this paper, we have addressed the problem of computing the symmetric $\chi^2$-divergence between two point processes. The proposed work is an extension of [7]; however, we avoid the stratification approach, and consider a smoothed representation of spike train as proposed in [6], and use a least square regression approach to estimate the divergence in a functional space. We observe that the proposed method is more robust to different statistical nature of the point processes. However, this approach is computationally more extensive compared to the stratification approach. To be specific, the stratification approach is $O(ml)$ in computation whereas the transformation approach is $O(f^3 + ml^2)$ in computation.

6. REFERENCES