Bayes Risk Consistency of Nonparametric Classification Rules for Spike Trains Data

Mirosław Pawlak, Member, IEEE, Mateusz Pabian, Student Member, IEEE, and Dominik Rzepka, Member, IEEE

Abstract

Spike trains data find a growing list of applications in computational neuroscience, imaging, streaming data and finance. Machine learning strategies for spike trains are based on various neural network and probabilistic models. The probabilistic approach is relying on parametric or nonparametric specifications of the underlying spike generation model. In this paper we consider the two-class statistical classification problem for a class of spike train data characterized by nonparametrically specified intensity functions. We derive the optimal Bayes rule and next form the plug-in nonparametric kernel classifier. Asymptotical properties of the rules are established including the limit with respect to the increasing recording time interval and the size of a training set. In particular the convergence of the kernel classifier to the Bayes rule is proved. The obtained results are supported by a finite sample simulation studies.

Index Terms

Bayes risk consistency, kernel classifiers, spike trains data, stochastic integrals

I. Introduction

EVENT driven systems are often encountered in science and engineering. In such systems data are represented by point processes that define arrival times of events. In computational neuroscience and machine learning this type of data are called spike trains [?], [?], [?]. In optical communication systems one observes a train of impulses (representing a point process) emitted by photon-sensitive detectors. Signal detection and estimation methods for such the so-called Poisson regime channels have been extensively examined in communication and information theory [?], [?], [?]. On the other hand, the mathematical theory of point processes has been extensively studied in the statistical and stochastic processes literature [?], [?]. However, the research on event type processes from the statistical classification theory [?] perspective has been initiated very recently [?], [?]. Probabilistic spiking neural networks have been introduced for supervised and unsupervised learning problems [?]. Various simulation results have been reported supporting their usefulness without, however, any accuracy studies and fundamental limits.

In this paper, we develop the Bayes strategy [?] for the spiking data supervised classification problem. This strategy can be applied to research problems where event ocurrence is the primary information carrier [?], [?]. We consider a class of temporal spiking processes that are characterized by non-random intensity functions. The intensity function plays the central role in our theory as it describes the local rate of occurrence of spikes. For such processes (Section ??) we derive the optimal Bayes rule in terms of class intensity functions. In Section ?? the limit behavior the Bayes rule with respect to the increasing length of the observation interval is examined. In Section ?? the plug-in nonparametric kernel classification rule from multiple replications of spiking processes is proposed. This is followed by the asymptotical optimality result, i.e., the convergence of the kernel rule to the Bayes rule. This result can be considered as the counterpart of the result in [?] concerning the classical plug-in nonparametric classification rules defined in the finite-dimensional Euclidean space. The spike train data are characterized by the variable-length continuous-discrete vectors of event times and their number over a given observation interval. The main mathematical tool in our asymptotic analysis is the theory of the martingale decomposition for counting processes [?].

It is also worth mentioning that the asymptotic optimality does not hold if one observes the long single realization of the underlying spiking process. In fact, the intensity estimation problem for spiking processes does not fall into the classical large-sample-smaller distance between sample points framework as the point process is casual in time [?]. Hence, for a fixed observation interval one must increase the number of events. This can be achieved by either scaling the intensity function or by using the replicates of the spiking process. The former approach can be based on the multiplicative intensity model due to Aalen [?], whereas the latter one (used in this paper) is the standard machine learning strategy, where the replicates form the training set. In this case the resulting kernel estimate will be obtained by aggregating kernel estimates from single realizations. Our asymptotic results are supported by simulation studies presented in Section ??. The preliminary version of the results developed in this paper has been reported in [?].

M. Pawlak is with the Department of Electrical and Computer Engineering, University of Manitoba, R3T 5V6 Winnipeg, Canada, and with the Department of Measurement and Electronics, AGH University of Krakow, 30-059 Kraków, Poland. E-mail: Miroslaw.Pawlak@umanitoba.ca.
M. Pabian and D. Rzepka are with the Department of Measurement and Electronics, AGH University of Krakow, 30-059 Kraków, Poland.
A preliminary version of this paper was presented at IEEE ICASSP 2023.

The symbol $\mathbf{1}(A)$ denotes the indicator function of the set A. We shall use the notation (P) for the convergence in probability, whereas $(a.s.)$ denotes the convergence with probability one. Also $a_T \prec cb_T$ denotes the asymptotic bound, i.e., $a_T \leq cb_T$ for sufficiently large T. Furthermore, $\overline{\lim}$, $\underline{\lim}$ denote the limit superior and inferior, respectively. Also, by M_f we will denote the Lipschitz constant of a function $f(t)$, i.e., $f(t)$ meets the Lipschitz condition if $|f(t_1) - f(t_2)| \leq M_f |t_1 - t_2|$ for all t_1, t_2.

II. Bayes Classification Rule

A temporal spiking process $\{N(t), t \geq 0\}$ consists of a sequence of random times $\{t_i\}$ of isolated events in time such that $N(0) = 0$. The process $N(t)$ can be defined by the counting function $N(t) = \sum_i \mathbf{1}(t_i \leq t)$ which is the number of events in $[0, t]$. We assume that the process is observed on the time window $[0, T]$ and is characterized by the non-random intensity function $\lambda(t)$ that is defined for all $t \geq 0$. This is a non-negative function that describes the local arriving rate of events such that $\mathbb{E}[N(T)] = \int_0^T \lambda(u)du$ is the average number of events in $[0, T]$. Hence, the observed on $[0, T]$ process $N(t)$ can be represented by the variable-length vector $\mathbf{X} = [t_1, \ldots, t_N; N]$, where $0 < t_1 < \cdots < t_N < T$ are the event times and $N = N(T)$. Writing $[t_1, \ldots, t_N; N]$ we emphasize the fact that the data vector consists of two parts: the occurrence times $\{t_1, \ldots, t_N\}$ and N being the number of events in $[0, T]$. The former is the continuous part of the vector $\mathbf{X}$, whereas the latter is its discrete part.

The goal of this paper is to develop a rigorous classification methodology for the aforementioned class of spiking processes based on the Bayes theory of classification [?]. Without a loss of generality we consider a two-class classification problem (see Section ?? for the generalization to the multi-class case) where class labels are denoted as ω_1, ω_2 with the priori probabilities π_1, π_2 , respectively. In order to form the optimal Bayes rule we recall the following known result [?] on the joint occurrence density of $\mathbf{X}$

$$f(\mathbf{x}) = \prod_{i=1}^{N} \lambda(t_i) \exp\left(-\int_0^T \lambda(u)du\right) \tag{1}$$

for $N = N(T) \geq 1$, whereas if $N = 0$ then $f(\mathbf{x}) = \exp\left(-\int_0^T \lambda(u)du\right)$. It is worth noting that (??) is the continuous-discrete distribution and by virtue of (??) the marginal density of the occurrence times $\{t_1, \ldots, t_N\}$ for $N \in \{0, 1, \ldots\}$ is given by

$$\begin{aligned}\sum_{n=0}^{\infty} f(t_1, \ldots, t_N; N = n) &= \exp\left(-\int_0^T \lambda(u)du\right) \\ &+ \exp\left(-\int_0^T \lambda(u)du\right) \sum_{n=1}^{\infty} \prod_{j=1}^{n} \lambda(t_j)\end{aligned}, \tag{2}$$

which is defined over the simplex regions $\mathbb{C}_n = \{(t_1, \ldots, t_n) : 0 \leq t_1 \leq \ldots \leq t_n \leq T\}$, $n = 1, 2, \ldots$. The formula in (??) defines the proper density over $\{\mathbb{C}_n\}$, i.e., we have

$$\begin{aligned}&\exp\left(-\int_0^T \lambda(u)du\right) \\ &+ \exp\left(-\int_0^T \lambda(u)du\right) \sum_{n=1}^{\infty} \int_{\mathbb{C}_n} \prod_{j=1}^{n} \lambda(t_j) dt_1 \cdots dt_n = 1\end{aligned}. \tag{3}$$

In the context of the classification problem the class occurrence densities in (??) will be denoted $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ depending whether $\mathbf{X}$ comes from class ω_1 (denoted as $\mathbf{X} \in \omega_1$) or if $\mathbf{X} \in \omega_2$, respectively. The corresponding class intensities are $\lambda_1(t)$, $\lambda_2(t)$ being the non-negative functions defined on $[0, \infty)$. Then using (??), one can form the optimal Bayes rule ψ_T^*: $\mathbf{X} \in \omega_1$ if

$$\prod_{i=1}^{N} \frac{\lambda_1(t_i)}{\lambda_2(t_i)} \exp\left(\int_0^T [\lambda_2(u) - \lambda_1(u)]\, du\right) \geq \frac{\pi_2}{\pi_1}. \tag{4}$$

assuming that $N \geq 1$ and $\exp\left(\int_0^T [\lambda_2(u) - \lambda_1(u)]\, du\right) \geq \frac{\pi_2}{\pi_1}$ if $N = 0$. Clearly, if the reverse inequality in (??) holds, then we classify $\mathbf{X}$ to ω_2. The log transform of (??) gives the alternative convenient form of the rule ψ_T^*, i.e., $\mathbf{X} \in \omega_1$ if

$$\sum_{i=1}^{N} \log\left(\frac{\lambda_1(t_i)}{\lambda_2(t_i)}\right) \geq \gamma, \tag{5}$$

where $\gamma = \int_0^T [\lambda_1(u) - \lambda_2(u)]\, du + \log\left(\frac{\pi_2}{\pi_1}\right)$. The rule in (??) can be usefully written in terms of the stochastic integral of the log-ratio $\log\left(\frac{\lambda_1(t)}{\lambda_2(t)}\right)$ with respect to the increments of the counting process $N(t)$, i.e., $\mathbf{X} \in \omega_1$ if

$$\int_0^T \log\left(\frac{\lambda_1(t)}{\lambda_2(t)}\right) dN(t) \geq \gamma\,. \tag{6}$$

Here $N(t)$ is the aforemenioned counting process with the intensity function $\lambda(t)$, where

$$\lambda(t) = \begin{cases} \lambda_1(t) & \text{if } \mathbf{X} \in \omega_1 \\ \lambda_2(t) & \text{if } \mathbf{X} \in \omega_2 \end{cases}. \tag{7}$$

For our further considerations it is useful to represent the class intensity functions on $[0, T]$ in terms of the so-called intensity factor and shape function [?]. Thus, let $\lambda_1(t) = \tau_1 p_1(t)$, $\lambda_2(t) = \tau_2 p_2(t)$, where

$$\tau_i = \int_0^T \lambda_i(u) du, \;\; p_i(t) = \lambda_i(u)/\tau_i, \;\; i = 1, 2\,. \tag{8}$$

Clearly $p_1(t)$, $p_2(t)$ are well-defined probability density functions on $[0, T]$. The representation in (??) allows us to represent the classification problem in terms of the class intensity factors and shape densities, and employ information-theoretic divergence measures. Using (??), we can rewrite the rule in (??) as follows, $\mathbf{X} \in \omega_1$ if

$$\sum_{i=1}^{N} \log\left(\frac{p_1(t_i)}{p_2(t_i)}\right) \geq \eta\,, \tag{9}$$

where $\eta = \tau_1 - \tau_2 + N\log\left(\frac{\tau_2}{\tau_1}\right) + \log\left(\frac{\pi_2}{\pi_1}\right)$. The Bayes rule ψ_T^* in (??) will be written as $W_T(\mathbf{X}) \geq \eta_T$ emphasizing the fact that the vector $\mathbf{X}$ is observed within the time window $[0, T]$.
It is worth noting that if $\lambda_1(t) = \lambda_1$ and $\lambda_2(t) = \lambda_2$, i.e., if we have the homogeneous spike train data then the Bayes rule takes the following form ψ_T^*: $\mathbf{X} \in \omega_1$

$$N\log\left(\frac{\lambda_1}{\lambda_2}\right) + T\,(\lambda_2 - \lambda_1) \geq \log\left(\frac{\pi_2}{\pi_1}\right)\,, \tag{10}$$

provided that $N \geq 1$. In the case $N = 0$ this reads as $\lambda_2 - \lambda_1 \geq \frac{1}{T}\log\left(\frac{\pi_2}{\pi_1}\right)$. The risk associated with the rule $\psi_T^*(\mathbf{x})$ in (??) (or (??)) is defined as $\mathbf{R}_T^* = \mathbf{P}(\psi_{\mathbf{T}}^*(\mathbf{X}) \neq \mathbf{Y})$ and is referred as the Bayes risk. Here $Y \in \{\omega_1, \omega_2\}$ is the true class label of $\mathbf{X}$. For our future studies we express the Bayes risk in terms of the decision function $W_T(\mathbf{X})$, i.e., we write

$$\begin{aligned} \mathbf{R}_T^* = {} & \mathbf{P}\left(\mathbf{W}_T(\mathbf{X}) \geq \eta_T | \mathbf{X} \in \omega_2\right)\pi_2 \\ & + \mathbf{P}\left(\mathbf{W}_T(\mathbf{X}) < \eta_T | \mathbf{X} \in \omega_1\right)\pi_1 \end{aligned}. \tag{11}$$

It is an important question to evaluate the Bayes risk. This includes various bounds on $\mathbf{R}_T^*$ and the behavior of $\mathbf{R}_T^*$ as a function of T. In Sections ?? and ?? we present results concerning such issues.

The presented results rely on the following local decomposition (see Appendix A) of the increment $dN(t)$ of the point process $N(t)$. Hence, we have

$$dN(t) = \lambda(t)dt + dM(t)\,, \tag{12}$$

where $\lambda(t)$ is the intensity function of $N(t)$, and $dM(t)$ is a zero mean process with uncorrelated but non-stationary increments. The formula in (??) can be viewed as the local signal plus noise decomposition, where the noise process $dM(t)$ reveals the local martingale structure [?]. Appendix A gives the pertinent results concerning the martingale decomposition of the underlying spiking process.

The decomposition in (??) allows us to express the classification rule in (??) (or its version in (??)) in the convenient stochastic integral form. In fact, by virtue of (??) and (??) we write the left-hand side of (??) as

$$\begin{aligned} \int_0^T \log\left(\frac{\lambda_1(t)}{\lambda_2(t)}\right) dN(t) = {} & \int_0^T \log\left(\frac{\lambda_1(t)}{\lambda_2(t)}\right) \lambda(t)dt \\ & + \int_0^T \log\left(\frac{\lambda_1(t)}{\lambda_2(t)}\right) dM(t) \end{aligned}, \tag{13}$$

where $\lambda(t)$ is given in (??) and $M(t)$ is the corresponding noise process defined in (??). The first term in (??) is the bias term of the optimal decision function, whereas the second one is the zero mean random variable contributing to the statistical variability of the rule. In Section ?? we show that the normalized version of this term converges exponentially fast to zero as $T \to \infty$ with probability one.

III. The Bayes Rule and Risk: Bounds and Asymptotic Behavior

A. The Bayes Decision Function

In this section we examine the optimal decision function derived in (??) or its alternative form in (??). Owing to the decomposition in (??) and using (??) we can arrive to the following equivalent form of the rule ψ_T^* in (??), $\mathbf{X} \in \omega_1$ if

$$U_T(\mathbf{X}) \geq \alpha_T + \log\left(\frac{\pi_2}{\pi_1}\right), \tag{14}$$

where

$$U_T(\mathbf{X}) = \int_0^T g(t) dM(t) . \tag{15}$$

Here $g(t) = \log\left(\frac{\lambda_1(t)}{\lambda_2(t)}\right) = \log\left(\frac{p_1(t)}{p_2(t)}\right) + \log\left(\frac{\tau_1}{\tau_2}\right)$ and

$$\begin{aligned} \alpha_T = \tau_1 - \tau_2 + \log\left(\frac{\tau_2}{\tau_1}\right) \int_0^T \lambda(t) dt \\ + \int_0^T \log\left(\frac{p_2(t)}{p_1(t)}\right) \lambda(t) dt \end{aligned}, \tag{16}$$

where $\lambda(t)$ is specified in (??).

It is worth noting that $U_T(\mathbf{X})$ in (??) represents the stochastic part of the Bayes rule. This takes the form of the stochastic integral with respect to the increments of the martingale process $M(t)$. It is known [?] that the martingale property is preserved under stochastic integration. Hence, since $\mathbb{E}\left[dM(t)\right] = 0$ the process

$$\left\{ U_t(\mathbf{X}) = \int_0^t g(u) dM(u), 0 \leq t \leq T \right\}$$

is a zero mean local martingale associated with the counting process $N(t)$, see Appendix A for further details. In addition, the integral in (??) is specified by the log-ratio $\log\left(\frac{\lambda_1(t)}{\lambda_2(t)}\right)$ and this is generally the unbounded function. To prevent this singularity it suffices to assume the class intensities $\lambda_1(t)$, $\lambda_2(t)$ that are bounded away from zero. Moreover, intensity functions are commonly bounded. All these restrictions can be formalized by the following assumption that will be used in the paper. Hence, assume that there exist positive numbers δ and C such that

$$\boldsymbol{A1:}\ 0 < \delta \leq \lambda_i(t) \leq C, \ \ i = 1, 2, \ \text{ for all } \ t \geq 0 . \tag{17}$$

We refer to [?], [?] for some weaker conditions for the existence of the aforementioned log-ratio.

In this section we present the preliminary results that characterize the Bayes rule specified by (??) and (??). This includes some bounds on the threshold α_T in (??) and the statistical properties of the stochastic term in (??). To do so, we recall that the Kullback-Leibler (KL) divergence [?] between densities $p(t)$ and $q(t)$ on $[0, T]$ is defined as follows

$$\mathbf{K}_T\left(p \parallel q\right) = \int_0^T \log\left(\frac{p(t)}{q(t)}\right) p(t) dt . \tag{18}$$

It is known that $\mathbf{K}_T\left(p \parallel q\right) \geq 0$ and $\mathbf{K}_T\left(p \parallel q\right) = 0$ if $p = q$.

The following lemma gives the upper and lower bounds for the threshold α_T in (??) in terms of the KL divergence between the class densities and the normalized square distance between the corresponding intensity factors. We will find these bounds useful in evaluating the Bayes risk.

Lemma 1. Let α_T be the threshold defined in (??). Then we have
(a) If $\mathbf{X} \in \omega_1$ then

$$\begin{aligned} -\frac{\left(\tau_1 - \tau_2\right)^2}{\tau_2} - \tau_1 \mathbf{K}_T\left(p_1 \parallel p_2\right) \\ \leq \alpha_T \leq -\tau_1 \mathbf{K}_T\left(p_1 \parallel p_2\right) \end{aligned}. \tag{19}$$

(b) If $\mathbf{X} \in \omega_2$ then

$$\begin{aligned} \tau_2 \mathbf{K}_T\left(p_2 \parallel p_1\right) \leq \alpha_T \\ \leq \frac{\left(\tau_1 - \tau_2\right)^2}{\tau_1} + \tau_2 \mathbf{K}_T\left(p_2 \parallel p_1\right) \end{aligned}, \tag{20}$$

where p_1, p_2, τ_1, τ_2 are defined in (??). The proof of Lemma ?? is given in Appendix B.

As the KL divergence is non-negative, then Lemma ??(a) yields $\alpha_T \leq 0$ if $\mathbf{X} \in \omega_1$, whereas Lemma ??(b) gives $\alpha_T \geq 0$ for $\mathbf{X} \in \omega_2$. Also it is seen that α_T lies in the interval of the length $\left(\tau_1 - \tau_2\right)^2 / \tau_2$ and $\left(\tau_1 - \tau_2\right)^2 / \tau_1$ if $\mathbf{X} \in \omega_1$

and $\mathbf{X} \in \omega_2$, respectively. It is also worth noting that $(\tau_1 - \tau_2)^2 = \left\{ \int_0^T (\lambda_1(t) - \lambda_2(t))\, dt \right\}^2$ represents the square of the difference of the average number of events on $[0, T]$ coming from classes ω_1 and ω_2. As a result, if $\tau_1 = \tau_2$ then $\alpha_T = -\tau_1 \mathbf{K}_T (p_1 \parallel p_2)$ for $\mathbf{X} \in \omega_1$ and $\alpha_T = \tau_2 \mathbf{K}_T (p_2 \parallel p_1)$ for $\mathbf{X} \in \omega_2$.

The next result concerns the stochastic part $U_T(\mathbf{X})$ defined in (??). This is given in the form of the stochastic integral of the log-ratio $\log\left(\frac{\lambda_1(t)}{\lambda_2(t)}\right)$ with respect to the increments of $M(t)$ such that $\mathbb{E}\left[U_T(\mathbf{X})\right] = 0$. In the following lemma we evaluate the basic statistical feature of this term by deriving its variance.

Lemma 2. Let us consider the stochastic part $U_T(\mathbf{X})$ of the Bayes rule in (??). Then,

(a) If $\mathbf{X} \in \omega_1$ then

$$\begin{aligned} &\mathbf{Var}\left[U_T(\mathbf{X})\right] \\ &= \tau_1 \int_0^T \left\{ \log\left(\frac{p_1(t)}{p_2(t)}\right) + log\left(\frac{\tau_1}{\tau_2}\right) \right\}^2 p_1(t)dt \end{aligned}. \tag{21}$$

(b) If $\mathbf{X} \in \omega_2$ then

$$\begin{aligned} &\mathbf{Var}\left[U_T(\mathbf{X})\right] \\ &= \tau_2 \int_0^T \left\{ \log\left(\frac{p_2(t)}{p_1(t)}\right) + log\left(\frac{\tau_2}{\tau_1}\right) \right\}^2 p_2(t)dt \end{aligned}. \tag{22}$$

The proof of Lemma ?? is given in Appendix B.

The formulas in Lemma ?? can be expressed in terms of the higher-order KL divergence between two class densities referred to as the KL variation [?]. Hence, let

$$\mathbf{V}_T (p \parallel q) = \int_0^T \log^2\left(\frac{p(t)}{q(t)}\right) p(t)dt \tag{23}$$

be the KL variation between densities $p(t)$ and $q(t)$ on $[0, T]$. Note that $\mathbf{V}_T (p \parallel q) = 0$ if $p = q$. Moreover, the following result describes the relationship between $\mathbf{V}_T (p \parallel q)$ and the standard KL divergence in (??).

Lemma 3. For any pair of probability densities p, q on $[0, T]$ we have

$$\mathbf{K}_T (p \parallel q) \leq \sqrt{\mathbf{V}_T (p \parallel q)}\,. \tag{24}$$

The bound in (??) results from the direct application of the Cauchy-Schwarz inequality. Returning back to the formula in (??) we can obtain that

$$\begin{aligned} \mathbf{Var}\left[U_T(\mathbf{X})\right] = \tau_1 \Big\{ &\mathbf{V}_T (p_1 \parallel p_2) \\ &+2\log\left(\frac{\tau_1}{\tau_2}\right) \mathbf{K}_T (p_1 \parallel p_2) + \log^2\left(\frac{\tau_1}{\tau_2}\right) \Big\} \end{aligned}. \tag{25}$$

The analogous formula can be written for (??).

It is an interesting question to examine the behavior of the stochastic term $U_T(\mathbf{X})$ for an increasing value of the observation interval T. In particular, we wish to derive an analog of the law of large numbers, i.e., the limit behavior of

$$\frac{1}{T} U_T(\mathbf{X}) = \frac{1}{T} \int_0^T g(t) dM(t) \tag{26}$$

as $T \to \infty$, where $g(t)$ is the log-ratio $\log\left(\frac{\lambda_1(t)}{\lambda_2(t)}\right)$. To give an answer to such questions we need to put some condition on the growth of the assumed class of intensity functions. Hence, suppose that there exists positive number d such that

$$\boldsymbol{A2:}\ \frac{1}{T} \int_0^T \lambda_i(u)du \to d, \ \ i = 1, 2 \ \text{ as } \ T \to \infty\,. \tag{27}$$

The meaning of this condition is that the average number of events from the each class increases linearly with T. It is worth noting that for intensity functions that are integrable on $[0, \infty)$ the condition in (??) holds with $d = 0$.
Based on the assumption $\boldsymbol{A2}$ we wish to evaluate the limit behavior of $\mathbf{Var}\left[U_T(\mathbf{X})\right]$ as $T \to \infty$. It is clear that such limit may not exist. Nevertheless, using the assumption $\boldsymbol{A1}$ we can find the upper and lower bounds for $\mathbf{Var}\left[U_T(\mathbf{X})\right]$. In fact, recalling (??) and (??) we have that if $\mathbf{X} \in \omega_1$

$$\begin{aligned} \mathbf{Var}\left[U_T(\mathbf{X})\right] &= \tau_1 \int_0^T \left\{ \log\left(\frac{\lambda_1(t)}{\lambda_2(t)}\right) \right\}^2 p_1(t)dt \\ &\leq \tau_1 \log^2\left(\frac{C}{\delta}\right) \end{aligned}. \tag{28}$$

On the other hand by (??) and Lemma ??, we get

$$\begin{aligned}\mathbf{Var}\left[U_T(\mathbf{X})\right] \geq \tau_1 \Big\{ & \mathbf{K}_T^2\left(p_1 \parallel p_2\right) \\ & +2\log\left(\frac{\tau_1}{\tau_2}\right)\mathbf{K}_T\left(p_1 \parallel p_2\right) + \log^2\left(\frac{\tau_1}{\tau_2}\right)\Big\}\end{aligned}.$$

The right-hand side of this inequality is equal to $\tau_1\left(\int_0^T \log\left(\frac{\lambda_1(t)}{\lambda_2(t)}\right)p_1(t)dt\right)^2$ and by (??) this is not smaller than $\tau_1 \log^2\left(\frac{\delta}{C}\right)$. Hence, if $\mathbf{X} \in \omega_1$ this gives the following bounds

$$\tau_1 \log^2\left(\frac{\delta}{C}\right) \leq \mathbf{Var}\left[U_T(\mathbf{X})\right] \leq \tau_1 \log^2\left(\frac{C}{\delta}\right). \tag{29}$$

Analogously, we can show that if $\mathbf{X} \in \omega_2$, then

$$\tau_2 \log^2\left(\frac{\delta}{C}\right) \leq \mathbf{Var}\left[U_T(\mathbf{X})\right] \leq \tau_2 \log^2\left(\frac{C}{\delta}\right). \tag{30}$$

The bounds in (??), (??) and the assumption in (??) lead to the following limit behavior of $\mathbf{Var}\left[U_T(\mathbf{X})\right]$.

Lemma 4. Let the assumptions ***A1***, ***A2*** hold. Then for $\mathbf{X} \in \omega_1$ or $\mathbf{X} \in \omega_2$ we have

$$\begin{aligned} d\log^2\left(\frac{\delta}{C}\right) &\leq \underline{\lim}_{T\to\infty}\mathbf{Var}\left[\frac{1}{\sqrt{T}}U_T\left(\mathbf{X}\right)\right] \\ &\leq \overline{\lim}_{T\to\infty}\mathbf{Var}\left[\frac{1}{\sqrt{T}}U_T\left(\mathbf{X}\right)\right] \leq d\log^2\left(\frac{C}{\delta}\right)\end{aligned}. \tag{31}$$

The question whether the inferior and superior limits in (??) are equal remains open. It should be noted that if (??) is in the form $\frac{1}{T}\int_0^T \lambda_i(u)du \to d_i$ $i = 1, 2$, then the result of Lemma ?? holds with d replaced by d_1 (if $\mathbf{X} \in \omega_1$) or d_2 (if $\mathbf{X} \in \omega_2$), respectively. To shed some light on the result in (??) let us consider the following simple example.

Example 1. Let us consider the classification problem with the intensity functions $\lambda_1(t)$ and $\lambda_2(t) = \mu\lambda_1(t)$ for some $\mu > 0$. Then we have $\tau_2 = \mu\tau_1$ and $p_2(t) = p_1(t)$. This implies that the condition in (??) reads as $\frac{1}{T}\int_0^T \lambda_1(u)du \to d$ and $\frac{1}{T}\int_0^T \lambda_2(u)du \to \mu d$. Then, a simple algebra gives the following analog of Lemma ??.
If $\mathbf{X} \in \omega_1$ then

$$\lim_{T\to\infty}\mathbf{Var}\left[\frac{1}{\sqrt{T}}U_T\left(\mathbf{X}\right)\right] = d\log^2(\mu), \tag{32}$$

whereas if $\mathbf{X} \in \omega_2$ then

$$\lim_{T\to\infty}\mathbf{Var}\left[\frac{1}{\sqrt{T}}U_T\left(\mathbf{X}\right)\right] = d\mu\log^2(\mu). \tag{33}$$

Note that the assumption ***A1*** is not required here. Also if $\mu = 1$ then the asymptotic constants are zero, i.e., this corresponds to the case $\lambda_1(t) = \lambda_2(t)$. Moreover, the asymptotic constants tend to infinity as $\mu \to \infty$.

An important consequence of Lemma ?? is the following weak law of large numbers for the average value of $U_T(\mathbf{X})$ defined in (??).

Theorem 1. Let the conditions of Lemma ?? hold. Then for $\mathbf{X}$ coming either from class ω_1 or class ω_2 we have

$$\frac{1}{T}U_T(\mathbf{X}) = \frac{1}{T}\int_0^T \log\left(\frac{\lambda_1(t)}{\lambda_2(t)}\right)dM(t) \to 0 \quad (P) \tag{34}$$

as $T \to \infty$.

The proof of this fact is a direct application of Lemma ?? and the Chebyshev inequality. In fact, let us consider the case $\mathbf{X} \in \omega_1$. Then, for any $\epsilon > 0$ we have

$$\mathbf{P}\left(\frac{1}{T}\left|U_T(\mathbf{X})\right| \geq \epsilon\right) \leq \frac{\mathbf{Var}\left[U_T(\mathbf{X})\right]}{T^2\epsilon^2}. \tag{35}$$

The right-hand side of (??) is equal to $\mathbf{Var}\left[\frac{1}{\sqrt{T}}U_T(\mathbf{X})\right]/T\epsilon^2$, where due to (??) the limit superior of $\mathbf{Var}\left[\frac{1}{\sqrt{T}}U_T(\mathbf{X})\right]$ is bounded by a finite constant. This confirms the claim of Theorem ??.

Our next goal is to strengthen the result of Theorem ?? by establishing the strong law of large numbers. This will result directly from the exponential inequality for the average of $U_T(\mathbf{X})$ defined in (??). Our main tools here are exponential inequalities for martingales of counting processes established recently in [?], see also [?] for earlier results. Hence, we employ the following adapted to our needs version of Theorem 5 in [?], see Appendix B for details.

Lemma 5. Let $N(t)$ be the counting process allowing the decomposition in (??). Let $U_T = \int_0^T g(t)dM(t)$ be the stochastic integral of the real-valued function $g(t)$ with respect to the martingale $M(t)$ increments. Suppose that
(a) $|g(t)| \le u_T$ for all $t \in [0, T]$.
(b) $\int_0^T g^2(t)\lambda(t)dt \le v_T$,
where u_T and v_T are some finite constants. Then, for each $\epsilon > 0$ we have

$$\mathbf{P}\left(|U_T| \ge \epsilon\right) \le 2\exp\left[-\frac{\epsilon^2}{2v_T + u_T\epsilon}\right]. \tag{36}$$

It is worth noting that this bound holds for any finite T.

Lemma ?? can be directly applied for the evaluation of the stochastic integral in (??). In fact, with $g(t) = \log\left(\frac{\lambda_1(t)}{\lambda_2(t)}\right)$ and by the assumption $\boldsymbol{A1}$, we have that $|g(t)| \le \log\left(\frac{C}{\delta}\right)$. Hence, the condition (a) in Lemma ?? is met with $u_T = \log\left(\frac{C}{\delta}\right)$ for all $T > 0$. By virtue of the property (??) in Appendix A the integral in the condition (b) of Lemma ?? reads as

$$\mathbf{Var}\left[U_T(\mathbf{X})\right] = T\mathbf{Var}\left[\frac{1}{\sqrt{T}}U_T(\mathbf{X})\right] = T\theta_T\,, \tag{37}$$

where due to (??) the limit superior of θ_T is bounded by a finite constant.

The preceding discussion gives the following exponential bound for the average value of $U_T(\mathbf{X})$ defined in (??). The bound is valid for any finite $T > 0$.

Lemma 6. Suppose that the assumption $\boldsymbol{A1}$ holds. Then for $\mathbf{X}$ coming either from class ω_1 or class ω_2 and every $\epsilon > 0$ we have

$$\mathbf{P}\left(\frac{1}{T}\left|U_T(\mathbf{X})\right| \ge \epsilon\right) \le 2\exp\left[-T\frac{\epsilon^2}{2\theta_T + u\epsilon}\right], \tag{38}$$

where $u = \log\left(\frac{C}{\delta}\right)$ and the factor θ_T is defined in (??).

The exponential bound in (??) and the Borel-Cantelli lemma yield the following strong version of Theorem ??. We should note, however, that the Borel-Cantelli lemma applies to a sequence of random variables, while the random variable $\xi_T = U_T(\mathbf{X})/T$ is a function of the continuous parameter T. Nevertheless, one can discretize ξ_T by finding a sequence of times T_n, such that $T_n \to \infty$ as $n \to \infty$ and then employ the standard Borel-Cantelli lemma. We refer to [?] for details for such discretization strategy.

Theorem 2. Let the assumptions $\boldsymbol{A1}$ and $\boldsymbol{A2}$ hold. Then for $\mathbf{X}$ coming either from class ω_1 or class ω_2 we have

$$\frac{1}{T}U_T(\mathbf{X}) = \frac{1}{T}\int_0^T \log\left(\frac{\lambda_1(t)}{\lambda_2(t)}\right)dM(t) \to 0 \quad (a.s.) \tag{39}$$

as $T \to \infty$.

B. The Bayes Risk

In this section we wish to evaluate the Bayes risk defined in (??). Our analysis will employ the results obtained in Section ??. Owing to (??) it suffices to consider the probability of misclassification $\mathbf{P}\left(W_T(\mathbf{X}) \ge \eta_T | \mathbf{X} \in \omega_2\right)$. The analysis of the probability $\mathbf{P}\left(W_T(\mathbf{X}) < \eta_T | \mathbf{X} \in \omega_1\right)$ is analogous. By virtue of (??) we can write

$$\begin{aligned}&\mathbf{P}\left(W_T(\mathbf{X}) \ge \eta_T | \mathbf{X} \in \omega_2\right)\\ &= \mathbf{P}\left(U_T(\mathbf{X}) \ge \alpha_T + \log\left(\frac{\pi_2}{\pi_1}\right)|\mathbf{X} \in \omega_2\right),\end{aligned} \tag{40}$$

where $U_T(\mathbf{X})$ is defined in (??) and α_T (under the fact that $\mathbf{X} \in \omega_2$) is given by

$$\alpha_T = \tau_1 - \tau_2 + \tau_2\log\left(\frac{\tau_2}{\tau_1}\right) + \tau_2\mathbf{K}_T(p_2 \parallel p_1)\,. \tag{41}$$

The first result reveals that the Bayes risk tends to zero as $T \to \infty$ under the assumptions $\boldsymbol{A1}$ and $\boldsymbol{A2}$. This is the direct consequence of the weak law of large numbers established in Theorem ??, see (??).
Hence, we have the following convergence result that also gives the upper bound for the Bayes risk.

Theorem 3. Let the assumptions $\boldsymbol{A1}$ and $\boldsymbol{A2}$ hold. Then, we have

$$\mathbf{R}_T^* \to 0 \;\text{ as }\; T \to \infty\,.$$

Furthermore,

$$\mathbf{R}_T^* \le \left(\pi_1 a_T + \pi_2 b_T\right)\frac{1}{T}\,, \tag{42}$$

for some finite constants a_T, b_T.

The proof of Theorem ?? is deferred to Appendix B, where also the explicit expressions for a_T and b_T are given. The bound in (??) is obtained by utilizing only the second moment of the stochastic integral $U_T(\mathbf{X})$ in (??).

Remark 1. Hence under the assumptions $\boldsymbol{A1}$ and $\boldsymbol{A2}$ the Bayes risk tends to zero with the rate $1/T$. The proof of Theorem ?? reveals also the following form of the asymptotic constant

$$c_1 = \frac{1}{d}\left(\frac{\log(C/\delta)}{\log(\delta/C)}\right)^2 . \tag{43}$$

Hence, for large T one can write $\mathbf{R}_T^* \prec c_1 \frac{1}{T}$.

By virtue of the result of Lemma ?? we can substantially improve the bound in (??). Hence, we have the following result.

Theorem 4. Let the assumptions $\boldsymbol{A1}$ and $\boldsymbol{A2}$ hold. Then, we have

$$\mathbf{R}_T^* \leq \pi_1 \exp\left[-A_T T\right] + \pi_2 \exp\left[-B_T T\right] , \tag{44}$$

for some finite constants A_T, B_T.

The proof of Theorem ?? is deferred to Appendix B, where also the explicit expressions for A_T and B_T are presented.

Remark 2. The proof of Theorem ?? shows that using the exponential inequality for the martingale process the Bayes risk tends to zero with the exponential rate and the following asymptotic constant

$$c_2 = d\frac{1}{3}\left(\frac{\log(\delta/C)}{\log(C/\delta)}\right)^2 , \tag{45}$$

where (δ, C) characterizes the assumption $\boldsymbol{A1}$, whereas d appears in the assumption $\boldsymbol{A1}$. Hence, for large T one can write $\mathbf{R}_T^* \prec \exp\left[-c_2 T\right]$. It is also worth noting that larger d in the assumption $\boldsymbol{A2}$ makes the bounds in (??) and (??) tighter. In fact, the constant c_1 in (??) decreases with d, whereas the constant c_2 in (??) increases with d.

Example 2. Consider the classification problem discussed in Example ??. Then, using the results in (??) and (??) and some algebra we can show the following counterpart of the result of Theorem ??

$$\mathbf{R}_T^* \prec \pi_1 \exp\left[-c_1(\mu) dT\right] + \pi_2 \exp\left[-c_2(\mu) dT\right] . \tag{46}$$

The asymptotic constants $c_1(\mu)$, $c_2(\mu)$ can be written in the explicit form and they obey the following properties

$$\lim_{\mu \to 1} c_1(\mu) = \lim_{\mu \to 1} c_2(\mu) = 0$$

and

$$\lim_{\mu \to \infty} c_1(\mu) = \lim_{\mu \to \infty} c_2(\mu) = \infty.$$

The former limit corresponds to the indistinguishable case, i.e., $\lambda_1(t) = \lambda_2(t)$. On the other hand, the latter limit exhibits that if $\mu \to \infty$ then $\mathbf{R}_T^* \to 0$. Again the assumption $\boldsymbol{A1}$ is not needed here.

Remark 3. In [?] the following upper bound for the Bayes risk is given

$$\mathbf{R}_T^* \leq \sqrt{\pi_1 \pi_2} \exp\left(-\beta(T)\right) ,$$

where $\beta(T) = \int_0^T \left[\frac{1}{2}\lambda_1(u) + \frac{1}{2}\lambda_2(u) - \sqrt{\lambda_1(u)\lambda_2(u)}\right] du$ is a positive factor. This is the classical Bhattacharya bound [?] extended to the classification problem for point processes. The behavior of $\beta(T)$ under the condition $\boldsymbol{A2}$ is an interesting open question. In the special case examined in Examples ??, ?? we can show that $\beta(T)$ behaves asymptotically as $c(\mu)dT$, where $c(\mu) = (\mu+1)/2 - \sqrt{\mu}$. Interestingly $c(\mu) \geq c_1(\mu), c_2(\mu)$, where $c_1(\mu), c_2(\mu)$ appear in our bound in (??).

Remark 4. The convergence of the Bayes risk $\mathbf{R}_T^*$ to zero is determined by the condition in $\boldsymbol{A2}$. This is due to the fact that the class intensity functions $\lambda_1(t), \lambda_2(t)$ grow with increasing T. If $\boldsymbol{A2}$ does not hold, e.g, if $\lambda_1(t), \lambda_2(t)$ are compactly supported then the convergence of $\mathbf{R}_T^*$ to zero is impossible. In this case in order to enforce the grow of $\lambda_1(t), \lambda_2(t)$ one could use the multiplicative model due to Aalen [?], i.e., we consider

$$\lambda_i(t) = d\gamma_i(t), \;\; i = 1, 2 , \tag{47}$$

where $\gamma_i(t) \;\; i = 1, 2$ are fixed functions and d is a parameter that is allowed to grow. It is an interesting alternative to derive the results obtained in this paper under the multiplicative class intensity model in (??).

In the following example we give some numerical illustration of the aforementioned results.

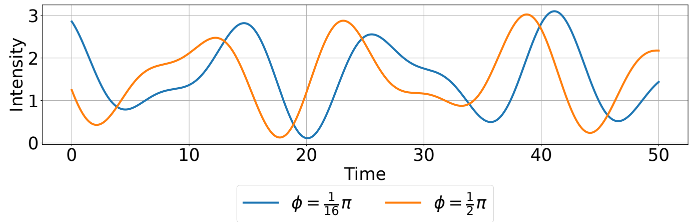


Figure 1. Intensity functions $\lambda_1(t) = \lambda\left(t; \frac{\pi}{16}\right)$ and $\lambda_2(t) = \lambda\left(t; \frac{\pi}{2}\right)$.

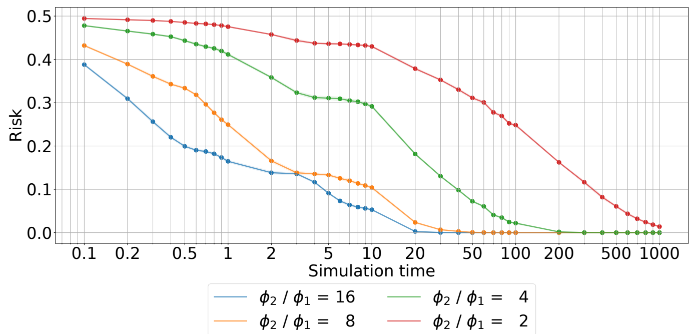


Figure 2. The Bayes risk $\mathbf{R}_T^*$ versus T for a two-class classification problem.

Example 3. Let us model the class intensities in the following form

$$\begin{aligned}\lambda(t;\phi) = 1.6 + \cos\left(\frac{\pi}{4\sqrt{3}}t + \phi\right) \\ + 0.5\cos\left(\frac{\pi}{3\sqrt{2}}t + \frac{\pi}{4} + \phi\right)\end{aligned}. \tag{48}$$

Various choices of ϕ define $\lambda_1(t)$, $\lambda_2(t)$. Figure ?? depicts $\lambda_1(t) = \lambda\left(t; \frac{\pi}{16}\right)$ and $\lambda_2(t) = \lambda\left(t; \frac{\pi}{2}\right)$.

Figure ?? illustrates the fact that the Bayes risk tends to zero as T gets larger. The model of class intensities defined in (??) is parametrized by ϕ, i.e., we set $\lambda_1(t) = \lambda_1(t;\phi_1)$ and $\lambda_2(t) = \lambda_1(t;\phi_2)$. The slowest decay of $\mathbf{R}_T^*$ is seen for very close intensities, i.e., when $\phi_2/\phi_1 = 2$ (in red), whereas the fast rate of convergence is observed for distant intensities, i.e., when $\phi_2/\phi_1 = 16$ (in blue). Nevertheless, since $\lambda(t;\phi)$ in (??) meets the assumptions $\boldsymbol{A1}$ and $\boldsymbol{A2}$ we can observe the exponential rate of convergence.

IV. Nonparametric Classification Rules

A. Plug-in Classifiers

In practice one does not know the true class intensities functions and must rely on some training data in order to form a data-driven classification rule. In this paper we apply the plug-in strategy to design a classifier, i.e. the classifier that is the empirical counterpart of the optimal Bayes rule in (??) or equivalently in (??). We have already pointed out that the single-sample based intensity function estimate cannot be consistent unless there is a certain mechanism that makes the intensity function increase, e.g., the multiplicative model in (??). In this paper we consider the intensity model based on the increasing number of replicates of the class spiking processes. Hence, contrary to the results of Section ?? the observation interval $[0,T]$ is kept constant.

Hence, let $\mathbf{D}_L = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_L, Y_L)\}$ be the learning sequence being a sample of L independent observations of the labeled spiking processes $(\mathbf{X}, Y)$. Here $\mathbf{X}_j$ is the variable-length vector, i.e., $\mathbf{X}_j = \left[t_1^{[j]}, \ldots, t_{N^{[j]}}^{[j]}; N^{[j]}\right]$ and $Y_j \in \{\omega_1, \omega_2\}$, where $N^{[j]} = N^{[j]}(T)$. Hence, all data are measured in the fixed time window $[0,T]$. Let L_1, L_2 be the number of training data of classes ω_1 and ω_2, respectively.

We wish to form the plug-in classification rule based on the optimal decision given in (??). This requires estimating the class intensity functions $\lambda_1(t)$, $\lambda_2(t)$, or equivalently the shape densities $p_1(t)$, $p_2(t)$ and the corresponding intensity factors τ_1, τ_2. It is known that the prior probabilities can be estimated by $\widehat{\pi}_1 = L_1/L$ and $\widehat{\pi}_2 = L_2/L$. In order to estimate $\{(\tau_i, p_i(t)), i = 1,2\}$ one can begin with the use of the single sample $\mathbf{X}_j$. Note that $\mathbb{E}\left[N^{[j]}|Y_j = \omega_i\right] = \tau_i$ and one can form the unbiased estimate of τ_i as $\widehat{\tau}_i^{[j]} = N^{[j]}$. However, $\mathbf{Var}\left[N^{[j]}|Y_j = \omega_i\right] = \tau_i$ and this is an inconsistent estimate of τ_i. The latter fact results from the local Poisson behavior of the spiking process, see Appendix A. Nevertheless, the aggregation of $\{\widehat{\tau}_i^{[j]}\}$ leads to consistent estimate of τ_i for the increased size of the training set. Hence, let

$$\widehat{\tau}_i = \frac{1}{L_i} \sum_{j=1}^{L} N^{[j]} \mathbf{1}(Y_j = \omega_i) \tag{49}$$

be an estimate of $\tau_i, \quad i = 1, 2$. In the analogous way we can deal with the problem of estimating $p_i(t)$. Let $\widehat{p}_i^{[j]}(t)$ be a certain nonparametric estimate of $p_i(t)$ based on the single sample $\mathbf{X}_j$ from the class ω_i. Then, the aggregated estimate of $p_i(t)$ takes the following form

$$\widehat{p}_i(t) = \frac{1}{L_i} \sum_{j=1}^{L} \widehat{p}_i^{[j]}(t) \mathbf{1}(Y_j = \omega_i), \;\; i = 1, 2\,. \tag{50}$$

Plugging (??) and (??) into (??) gives us the following empirical classification rule $\widehat{\psi}_{L,T}$: classify $\mathbf{X} = [t_1, \ldots, t_N; N] \in \omega_1$ if

$$\widehat{W}_{L,T}(\mathbf{X}) \geq \widehat{\eta}_{L,T}\,, \tag{51}$$

where $\widehat{W}_{L,T}(\mathbf{X}) = \sum_{i=1}^{N} \log\left(\frac{\widehat{p}_1(t_i)}{\widehat{p}_2(t_i)}\right)$, $\widehat{\eta}_{L,T} = \widehat{\tau}_1 - \widehat{\tau}_2 + N \log\left(\frac{\widehat{\tau}_2}{\widehat{\tau}_1}\right) + \log\left(\frac{L_2}{L_1}\right)$. In Section ?? we propose a concrete kernel-type estimate of the shape densities.

In this section we present a general result on the convergence of the rule $\widehat{\psi}_{L,T}$ to the Bayes decision $\psi_T^\star$. This result is in the spirit of the Bayes risk consistency theorem established in [?] in the context of the standard fixed dimension data sets. Let us first consider the pointwise behavior of the rule $\widehat{\psi}_{L,T}$ in (??). Hence, let $\mathbf{P}\left(\widehat{\psi}_{L,T}(\mathbf{x}) = \psi_T^\star(\mathbf{x})\right)$ be the probability that the empirical rule makes the same decisions as the optimal Bayes rule for a fixed test vector $\mathbf{x}$. Our first result reveals that this probability tends to one if the size of the training set tends to infinity.

Theorem 5. Suppose that for $i = 1, 2$ and $L \to \infty$ the following property holds

$$\widehat{p}_i(t) \to p_i(t) \;\; (P) \;\; \text{uniformly on} \;\; [0, T]\,. \tag{52}$$

Then,

$$\mathbf{P}\left(\widehat{\psi}_{L,T}(\mathbf{x}) = \psi_T^\star(\mathbf{x})\right) \to 1$$

as $L \to \infty$. The proof of Theorem ?? is given in Appendix C. This result assures that $\widehat{\psi}_{L,T}$ converges to $\psi_T^\star$ as long as one can construct uniformly consistent estimates of $p_i(t), \quad i = 1, 2$. Clearly, the uniform convergence of estimates of the class intensity functions $\lambda_i(t)$ also implies the local consistency result of Theorem ??.

The proof of Theorem ?? reveals also that the 0-1 distance between $\widehat{\psi}_{L,T}(\mathbf{x})$ and $\psi_T^\star(\mathbf{x})$ tends to zero. Hence, we have

$$\rho(\widehat{\psi}_{L,T}(\mathbf{x}), \psi_T^\star(\mathbf{x})) \to 0 \;\; (P) \tag{53}$$

as $L \to \infty$, where

$$\rho(\widehat{\psi}_{L,T}(\mathbf{x}), \psi_T^\star(\mathbf{x})) = \mathbf{1}\left(\widehat{\psi}_{L,T}(\mathbf{x}) \neq \psi_T^\star(\mathbf{x})\right)\,.$$

The condition in (??) of Theorem ?? assures that the decision function $\widehat{W}_{L,T}(\mathbf{x})$ in (??) tends to the optimal decision function $W_T(\mathbf{x})$ in (??). This is the convergence needed in the proof of Theorem ?? and is summarized in the following lemma.

Lemma 7. Let the class intensities $\lambda_1(t), \lambda_2(t)$ be uniformly continuous on $[0, \infty)$ such that restricted to $[0, T]$ satisfy the assumption ***A1***. Let (??) hold. Then, we have

$$\widehat{W}_{L,T}(\mathbf{x}) \to W_T(\mathbf{x}) \;\; (P)\,, \tag{54}$$

as $L \to \infty$.

The proof of Lemma ?? is postponed to Appendix C. The convergence in (??) is uniform with respect to $\mathbf{x}$.

The classification rule $\widehat{\psi}_{L,T}$ in (??) is also characterized by the threshold value $\widehat{\eta}_{L,T}$. Note that $\pi_1 = L_1/L, \quad \pi_2 = L_2/L$ are weakly consistent estimates of the prior probabilities π_1, π_2. Also the aggregated estimate τ_i in (??) of the intensity factor τ_i is weakly consistent. Hence, the preceding discussion gives the following consistency result

$$\widehat{\eta}_{L,T} \to \eta_T \;\; (P) \tag{55}$$

as $L \to \infty$.

The local consistency of $\widehat{\psi}_{L,T}$ leads to the global convergence characterized by the conditional risk. Hence, let

$\mathbf{P}(\widehat{\psi}_{L,T}(\mathbf{X}) \neq Y|\mathbf{D}_L) = \mathbb{E}\left[\mathbf{1}\left(\widehat{\psi}_{L,T}(\mathbf{X}) \neq Y\right)|\mathbf{D}_L\right]$ be the conditional risk associated with the rule $\widehat{\psi}_{L,T}$. Since $\mathbf{R}_T^\star = \mathbb{E}\left[\mathbf{1}\left(\psi_T^\star(\mathbf{X}) \neq Y\right)\right]$, then one can write

$$\begin{aligned} 0 &\leq \mathbf{R}_{L,T} - \mathbf{R}_T^\star \\ &= \mathbb{E}\left[\mathbf{1}\left(\widehat{\psi}_{L,T}(\mathbf{X}) \neq Y\right) - \mathbf{1}\left(\psi_T^\star(\mathbf{X}) \neq Y\right)|\mathbf{D}_L\right] . \end{aligned}$$

Recalling the definition of the distance in (??) the above is bounded by

$$\mathbb{E}\left[\rho(\widehat{\psi}_{L,T}(\mathbf{X}), \psi_T^\star(\mathbf{X}))|\mathbf{D}_L\right] .$$

Owing to (??) and Lebesgue's dominated convergence theorem we obtain the main result of this section.

Theorem 6. Consider the class of plug-in classifiers defined in (??). Suppose that the conditions of Theorem ?? hold. Then, we have the following Bayes risk consistency result

$$\mathbf{R}_{L,T} \to \mathbf{R}_T^\star \;\; (P) \tag{56}$$

as $L \to \infty$.

B. Kernel Classifiers

It is known [?], [?] that the intensity function of a point process can be efficiently estimated by a class of kernel methods [?], [?]. In particular, the standard single sample kernel estimate of $\lambda_i(t)$ takes the form

$$\widehat{\lambda}_i^{[j]}(t) = \sum_{l=1}^{N^{[j]}} K_h\left(t - t_l^{[j]}\right) , \tag{57}$$

where the sample $\mathbf{X}_j = \left[t_1^{[j]}, \ldots, t_{N^{[j]}}^{[j]}; N^{[j]}\right]$ comes from the class ω_i.

Here $K_h(t) = h^{-1}K(t/h)$, where the kernel $K(t)$ is assumed to be a compactly supported on $[-1, 1]$, symmetric probability density function. For instance, one can choose the so-called Epanechnikov kernel

$$K(t) = \frac{3}{4}\left(1 - t^2\right)\mathbf{1}\left(|t| \leq 1\right) .$$

The crucial tuning parameter h is called the bandwidth as it controls the level of smoothing via the scaled kernel $K_h(t)$.

The parameter τ_i can be estimated (from a single sample) by $\widehat{\tau}_i^{[j]} = N^{[j]}$. Therefore (??) yields the following estimate of the shape density

$$\widehat{p}_i^{[j]}(t) = \frac{1}{N^{[j]}}\sum_{l=1}^{N^{[j]}} K_h\left(t - t_l^{[j]}\right) .$$

As we have already pointed in Section ?? the estimates $\widehat{\lambda}_i^{[j]}(t)$, $\widehat{p}_i^{[j]}(t)$ cannot be consistent by merely increasing T. To overcome this problem one can utilize the observed multiple training vectors and aggregate the single-sample estimates $\left\{\widehat{\lambda}_i^{[j]}\right\}$, $\left\{\widehat{p}_i^{[j]}\right\}$. This leads to the following aggregated kernel estimate of $p_i(t)$

$$\widehat{p}_i(t) = \frac{1}{L_i}\sum_{j=1}^{L} \widehat{p}_i^{[j]}(t)\mathbf{1}\left(Y_j = \omega_i\right) . \tag{58}$$

Moreover, the aggregated estimate $\widehat{\tau}_i$ of τ_i is defined in (??). Plugging $\widehat{p}_i(t)$ and $\widehat{\tau}_i, \;\; i = 1, 2$ into (??) we obtain the kernel classification rule. The aggregated kernel estimate $\widehat{\lambda}_i(t)$ of $\lambda_i(t)$ is defined in the analogous way, see (??).

Theorem ?? and Theorem ?? reveal that the sufficient condition for the Bayes risk consistency is the convergence property in (??). Note that the statistical behavior of $\widehat{p}_i(t)$ and $\widehat{\lambda}_i(t)$ is the same and therefore we can verify the requirement in (??) for the kernel intensity estimate. Hence, with some abuse of the notation let $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_L\}$ be the data set from the fixed class (ω_1 or ω_2) of the counting process $N(t)$ characterized by the class intensity function $\lambda(t)$. Thus, one observes the L copies $\left\{N^{[j]}(t)\right\}$ of the counting process $N(t)$, where $N^{[j]}(t)$ is represented by the feature vector $\mathbf{X}_j = \left[t_1^{[j]}, \ldots, t_{N^{[j]}}^{[j]}; N^{[j]}\right]$ with $N^{[j]} = N^{[j]}(T)$. The local martingale decomposition in (??) for $N^{[j]}(t)$ reads

$$dN^{[j]}(t) = \lambda(t)dt + dM^{[j]}(t), \;\; j = 1, \ldots, L .$$

This gives the analogous decomposition for the aggregated counting process, i.e., we have

$$d\overline{N}_L(t) = \lambda(t)dt + d\overline{M}_L(t) , \tag{59}$$

where

$$d\overline{N}_L(t) = \frac{1}{L}\sum_{j=1}^{L} dN^{[j]}(t)\,,$$

$$d\overline{M}_L(t) = \frac{1}{L}\sum_{j=1}^{L} dM^{[j]}(t)\,.$$

It is important to note that the aggregated residual process $d\overline{M}_L(t)$ meets all the properties listed in Appendix A. Hence, $\mathbb{E}\left[d\overline{M}_L(t)\right] = 0$ and the properties in (??) and (??) are as follows

$$\begin{aligned}\mathbf{Var}\left[d\overline{M}_L(t)\right] &= \frac{1}{L}\lambda(t)dt\,,\\ \mathbf{Var}\left[\int_0^T g(u)d\overline{M}_L(u)\right] &= \frac{1}{L}\int_0^T g^2(u)\lambda(u)du\,.\end{aligned} \tag{60}$$

The single-sample kernel estimate of $\lambda(t)$ is as in (??), whereas its aggregated version takes the form

$$\widehat{\lambda}(t) = \frac{1}{L}\sum_{j=1}^{L}\widehat{\lambda}^{[j]}(t)\,. \tag{61}$$

This due to (??) can be written in the convenient stochastic integral form

$$\widehat{\lambda}(t) = \int_0^T K_h\left(t-s\right)d\overline{N}_L(s)\,. \tag{62}$$

Employing this identity along with (??) and the aforementioned properties of $d\overline{M}_L(t)$ (see (??)) yield the following identities for the bias and the variance of $\widehat{\lambda}(t)$

$$\mathbb{E}\left[\widehat{\lambda}(t)\right] = \int_0^T \frac{1}{h}K\left(\frac{t-s}{h}\right)\lambda(s)ds\,, \tag{63}$$

$$\mathbf{Var}\left[\widehat{\lambda}(t)\right] = \frac{1}{Lh}\int_0^T \frac{1}{h}K^2\left(\frac{t-s}{h}\right)\lambda(s)ds\,. \tag{64}$$

These formulas and the standard analysis developed in the context of kernel estimates [?], [?] reveal that if

$$h(L)\to 0 \ \text{ and } \ Lh(L)\to\infty$$

then

$$\widehat{\lambda}(t)\to\lambda(t) \ \ (P) \ \text{ as } \ L\to\infty \tag{65}$$

at $t\in(0,T)$ where $\lambda(t)$ is continuous. This is the pointwise convergence that holds at interior points of $[0,T]$. It is known [?], [?] that the convergence fails at the boundary points near $t=0,\quad t=T$. This enforces us to confine the required uniform convergence to the interval $[\epsilon, T-\epsilon]$ for arbitrarily small $\epsilon>0$. Yet another option is to introduce the boundary modified kernels [?], [?] that are able to restore the convergence property at the boundary points. The following lemma gives the sufficient conditions for the uniform convergence property of the estimate $\widehat{\lambda}(t)$ in (??).

Lemma 8. Let $\lambda(t)$ be Lipschitz continuous on $[0,\infty)$. Let the kernel function $K(t)$ be Lipschitz continuous on $[-1,1]$. Suppose that

$$h(L)\to 0 \ \text{ and } \ Lh^3(L)\to\infty \ \text{ as } \ L\to\infty\,. \tag{66}$$

Then for arbitrarily small $\epsilon>0$

$$\sup_{t\in[\epsilon,T-\epsilon]}\left|\widehat{\lambda}(t)-\lambda(t)\right|\to 0 \ \ (P) \ \text{ as } \ L\to\infty\,. \tag{67}$$

It is worth noting that the uniform convergence holds under the condition $Lh^3(L)\to\infty$. This is the stronger restriction than the one required for the pointwise convergence, where one needs that $Lh(L)\to\infty$. We conjecture that (??) can be replaced by the weaker condition $Lh(L)/\log(L)\to\infty$ This is the case for the uniform convergence of the kernel density estimate where advanced tools from the empirical processes theory have been utilized [?], [?]. Our proof is based on more elementary techniques. The proof of Lemma ?? is given in Appendix D. The result of Lemma ?? applies directly to the shape densities and by using Theorem ?? and Theorem ?? we can formulate the following Bayes risk consistency result for the kernel classifier.

Theorem 7. Let the class intensities $\lambda_1(t)$, $\lambda_2(t)$ satisfy the conditions of Lemma ??. If (??) holds, then the kernel classification rule is Bayes risk consistent, i.e., we have

$$\mathbf{R}_{L,T} \to \mathbf{R}_T^\star \ \ (P)$$

as $L \to \infty$.

The convergence in Theorem ?? is an important property of the kernel classifier. Nevertheless, the issue of the rate of convergence would also be essential. This question is left for further research.

The selection of the bandwidth h is the most important issue in determining the the finite sample accuracy of the kernel classification rule. The standard analysis applied to the expression in (??) and (??) shows that if $\lambda(t)$ has two continuous derivatives for $t \in (0,T)$ then

$$\mathbb{E}\left[\widehat{\lambda}(t)\right] = \lambda(t) + \mathcal{O}\left(h^2\right)$$

and

$$\mathbf{Var}\left[\widehat{\lambda}(t)\right] = \mathcal{O}\left(\frac{1}{Lh}\right).$$

This leads to the following asymptotical formula for the mean squared error

$$\mathbb{E}\left[\widehat{\lambda}(t) - \lambda(t)\right]^2 = \mathcal{O}\left(\frac{1}{Lh}\right) + \mathcal{O}\left(h^4\right), \ \ t \in (0,T).$$

The minimum of the error yields the asymptotically optimal choice of the bandwidth, i.e., $h^\star = cL^{-1/5}$ for some positive constant c. This is the asymptotically optimal choice of h that optimizes the kernel intensity estimate. An optimal bandwidth for the kernel classifier may be quite different as it is seen from the restriction in (??). See also [?] for the general theory of plug-in nonparametric classifiers.
In practical applications one can specify the bandwidth using some resampling techniques like cross-validation [?], [?]. In our experimental studies we choose separate bandwidth for each class. This is done by finding the maximum of the cross-validated log-likelihood of the kernel estimate of the shape densities. Hence, let $\widehat{p}_i(t;h)$ be the kernel estimate in (??) specified by the bandwidth h. Then, the likelihood function of $\widehat{p}_i(t;h)$ specified by test data is given by

$$\mathbf{CV}(h) = \prod_{l=1}^{p}\prod_{r=1}^{N^{[l]}} \widetilde{p}_i(t_r^{[l]};h), \tag{68}$$

where $t_r^{[l]}$ represents the r-th observation of the l-th test sample. We use the test sample of size q (per class). Also $\widetilde{p}_i(t;h)$ is the version of $\widehat{p}_i(t;h)$ in (??) determined from the $L_i - q$ size training set. Then, the bandwidth is selected as the one that maximizes $\mathbf{CV}(h)$ in (??). This is equivalent to the following choice

$$\widehat{h}_i = \arg\max_h \sum_{l=1}^{p}\sum_{r=1}^{N^{[l]}} \log\left(\widetilde{p}_i(t_r^{[l]};h)\right).$$

V. Simulation Results

In order to assess the proposed methodology, we conduct a simulated data study. We limit the scope of our experiments to time-dependent intensity functions defined in (??), and use these in simulations in order to gain insight into the behavior of $\mathbf{R}_{L,T}$ with respect to the training set size L and the observation window size T.

In all experiments the kernel classifier is given by (??) with the estimated $\widehat{\tau}_i$, $\widehat{p}_i(t;h)$ specified by (??) and (??), respectively. The Gaussian kernel is employed, whereas the bandwidth is selected by the log-likelihood method in (??). When selecting the bandwidth, we consider a grid of ten evenly logarithmically spaced points $h \in \langle 10^{-1}, 10^{1}\rangle$. Additionally, we employ a 5-fold cross validation in order to avoid biasing the selected bandwidth $\widehat{h}_i$ with the test data. Finally, we denote $\mathbb{E}\left[\mathbf{R}_{L,T}\right]$ as an empirically evaluated risk averaged over ten simulation runs with a testing set size of 10^4.

We shall focus on the simulation results obtained for the intensity function specified by (??). Unless noted otherwise, we refer to the intensity function pair parametrized by $\phi_1 = \pi/16$ and $\phi_2 = \pi/4$.

Figure ?? depicts the average risk versus T for the size of training data ranging from $L = 10$ to $L = 200$. The Bayes risk $\mathbf{R}_T^*$ is also plotted for comparison. The convergence of $\mathbb{E}[\mathbf{R}_{L,T}]$ to zero analogous as it was observed for the Bayes risk (see Figure ??) is seen. Also the small value of the difference $\mathbb{E}[\mathbf{R}_{L,T}] - \mathbf{R}_T^*$ for all T should be noted. We also observe the small variability of the risk with respect to the training data size L. The vertical dashed line at $T = 10$ denotes the simulation space slice in subsequent analysis, i.e., with the value of T fixed.

Next, we analyze the value of the optimal bandwidth selected according to the log-likelihood method versus T. For brevity, in Figure ??a we show only the results for $\widehat{h}_1$, noting that the curves obtained for $\widehat{h}_2$ are analogous. We observe

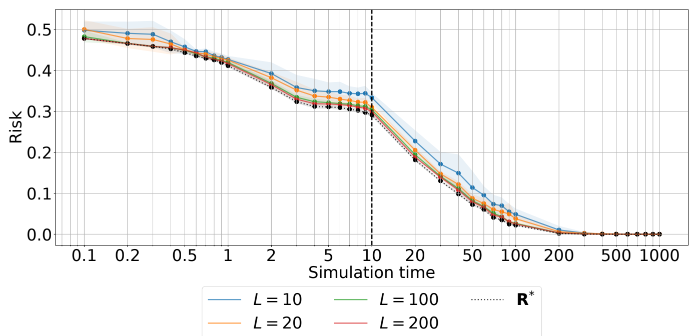


Figure 3. The average risk $\mathbb{E}\left[\mathbf{R}_{L,T}\right]$ versus T for different values of L. The vertical dashed line at $T = 10$ denotes the simulation space slice presented in Figure ??b-??.

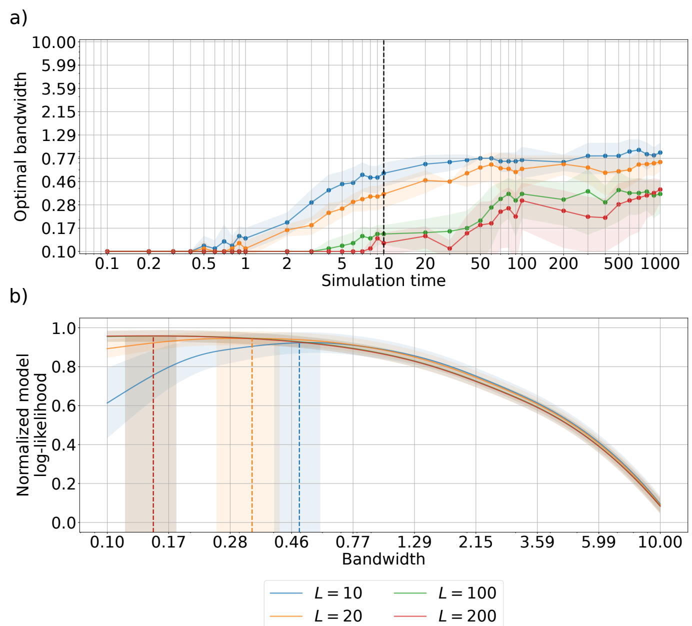


Figure 4. a) The average empirical bandwidth $\mathbb{E}\left[\widehat{h}_1\right]$ versus T for different values of L. The vertical dashed line at $T = 10$ denotes the simulation space slice presented in the lower subfigure. b) The average normalized model log-likelihood on test data versus h for different values of L. The vertical dashed lines denote function maxima. Note that the curves for $L = 100$ and $L = 200$ overlap.

an increase in $\widehat{h}_1$ with T, which aligns with the notion that as the observation window increases, the distribution of events in time becomes sparser, yielding the larger bandwidth. On the other hand, the obtained results also show that $h(L) \to 0$ as L increases. Another way to view this property is to analyze the model log-likelihood versus h for fixed T (Figure ??b).

Finally, Figure ?? shows the convergence of the empirical kernel rule risk to the Bayes risk for different values of the intensity function pair parameters ϕ_1, ϕ_2 versus L and fixed $T = 10$. Clearly, as the difficulty of the problem increases, i.e., when the two intensity functions become more similar to one another, the rate of convergence decreases. Also note that the Bayes risk is higher for more difficult classification problems.

Let us briefly examine a counter-example when the proposed algorithm fails to converge. Consider the following Gaussian type intensity function

$$\lambda(t; a, b) = a \exp\left[-b\,(t - 0.5)^2\right], \tag{69}$$

which does not satisfy the assumptions ***A1*** and ***A2***. While the intensity function has an infinite support, in practice it is extremely unlikely for events to occur outside of some narrow time interval. In Figure ?? we consider the classification problem with $\lambda_1(t) = \lambda\,(t; 300, 20)$ and $\lambda_2(t) = \lambda\,(t; 600, 40)$. For such specified intensity functions we can evaluate that $\int_0^T \lambda_1(t)dt < 119$ and $\int_0^T \lambda_2(t)dt < 169$ for all T. Hence, the average number of events from each class is finite and consequently the condition A2 does not hold. Note that the empirical risk does not converge to the Bayes risk that takes very small values for $T > 0.5$. Note also that the risk $\mathbb{E}\left[\mathbf{R}_{L,T}\right]$ is the smallest around the maximum of (??) at $t = 0.5$. Afterwards $\mathbf{R}_{L,T}$ slightly increases and reaches a plateau because no new events can be observed.

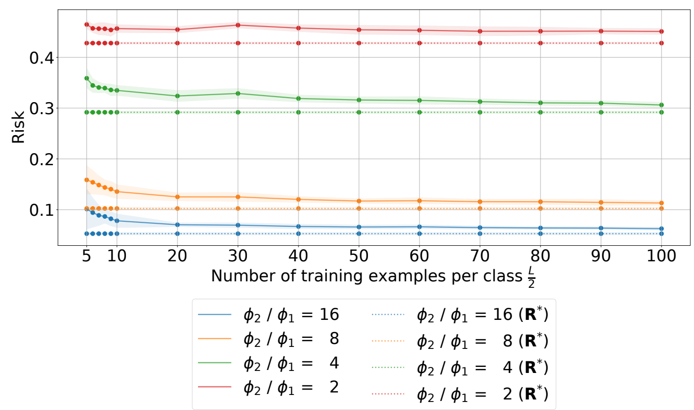


Figure 5. The average risk $\mathbb{E}[\mathbf{R}_{L,T}]$ versus L at given T for different values intensity function pairs parametrized by ϕ_1, ϕ_2. The horizontal dashed lines denote estimated Bayes risk $\mathbf{R}_T^*$ for the associated intensity function pair.

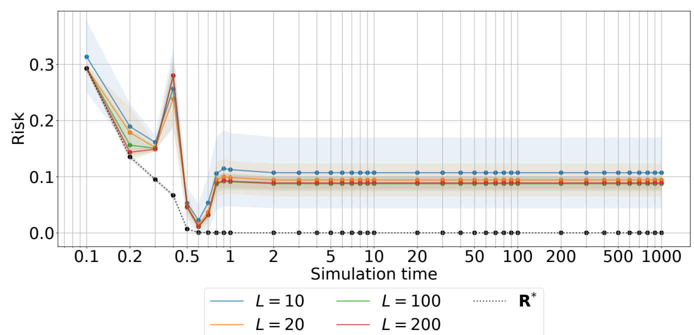


Figure 6. The average risk $\mathbb{E}\left[\mathbf{R}_{L,T}\right]$ versus T for different values of L for the Gaussian type intensity function.

VI. Concluding Remarks

In this paper we have developed the rigorous asymptotic analysis for the classification problem applied to spike trains data characterized by non-random intensity functions. The optimal Bayes rule was derived and its finite and asymptotic (with respect to the length of the observation interval) properties were established. This includes the exponential bound for the Bayes risk. Our asymptotic theory is relied on the martingale representation of counting processes. We then introduced a general class of plug-in empirical classification rules and formulated the sufficient conditions for their convergence (as the amount of data grows) to the Bayes risk. This optimality property is confirmed and verified for the plug-in kernel classifier derived from the aggregated data.

There are various ways to extend and generalize the results obtained in this paper. First of all, the log transformed version of the Bayes rule in (??) holds for a general class of point processes such as the Hawkes self-excited process [?] and multivariate or marked point processes [?]. Hence, the extension of our results to this type of point processes is a natural topic for future research. The two-class classification problem studied in this paper has straightforward generalization to the multi-class situation with the class labels denoted as $\{\omega_1, \ldots, \omega_c\}$. In fact, the Bayes rule in (??) for the $c-$class classification problem reads as

$$\mathbf{X} \in \omega_i \ \text{ if } \ \sum_{s=1}^{N} \log\left(\frac{\lambda_i(t_s)}{\lambda_k(t_s)}\right) \geq \gamma_{ik}\,, \qquad \text{for all } \ k = 1, \ldots, c, \ \ k \neq i$$

where $\gamma_{ik} = \int_0^T (\lambda_i(u) - \lambda_k(u))\, du + \log(\pi_k/\pi_i)$. Here $\{\lambda_i(t)\}$ are class intensity functions and $\{\pi_i\}$ are prior probabilities. Utilizing the martingale decomposition (see (??)) for point processes would allow us to generalize our asymptotic results to the multi-class case. Also designing nonparametric plug-in classification rules with the desirable asymptotic optimality property would be of a great practical topic for further research.

Appendix A

The asymptotic theory of the classification problem examined in this paper is based on martingale methods. This appendix gives brief summary of the essential facts concerning the counting processes theory and their martingale representation, see [?] for the full account of this theory. Hence, let $N(t)$ be a spike train process which can be consider as a counting process of the occurrences in the interval $[0,t]$ such that $N(0)=0$. By $dN(t)$ we denote the increment of $N(t)$ over the small interval $[t, t+dt)$. The evolution of $N(t)$ in time is completely characterized by the local intensity function $\lambda(t)$. This is defined as

$$\mathbb{E}\left[dN(t)|\mathbf{F}_t\right] = \lambda(t)dt\,, \tag{70}$$

where $\mathbf{F}_t$ denotes the history of $N(t)$ in the interval $[0,t)$. Note that $\lambda(t)$ is generally random due to the dependence on the values of $N(t)$ prior to the time t. The formula in (??) implies that the residual process

$$dM(t) = dN(t) - \lambda(t)dt \tag{71}$$

satisfies the property

$$\mathbb{E}\left[dM(t)|\mathbf{F}_t\right] = 0\,. \tag{72}$$

This confirms the fact that the process

$$M(t) = N(t) - \int_0^t \lambda(u)du \tag{73}$$

is a zero mean local martingale.

The formula in (??) can be written as

$$dN(t) = \lambda(t)dt + dM(t)\,. \tag{74}$$

This can be viewed as the local signal plus noise decomposition of $N(t)$. Moreover, the noise process $dM(t)$ in (??) is a zero mean martingale that has uncorrelated but nonstationary increments [?]. Based on these facts it can be shown that $dM(t)$ has the following second order property

$$\mathbf{Var}\left[dM(t)|\mathbf{F}_t\right] = \lambda(t)dt\,. \tag{75}$$

Also $\mathbf{Var}\left[dM(t)|\mathbf{F}_t\right] = \mathbf{Var}\left[dN(t)|\mathbf{F}_t\right]$.

The fact that $dM(t)$ has uncorrelated increments and that it reveals a piecewise constant sample paths allow us to define the stochastic Stieltjes type integral with respect to $dM(t)$. Hence, let

$$\mathbf{I}(t) = \int_0^t g(u)dM(u)$$

define the stochastic integral of the measurable function $g(t)$ with respect to the increments of the martingale $M(t)$. It is known [?] that the martingale property is preserved under stochastic integration. Since $\mathbb{E}\left[\mathbf{I}(t)\right] = 0$ the integral $\mathbf{I}(t)$ is a zero mean martingale with respect to the history of the counting process $N(t)$. The variance of $\mathbf{I}(t)$ is given by

$$\mathbf{Var}\left[\mathbf{I}(t)|\mathbf{F}_t\right] = \int_0^t g^2(u)\lambda(u)du\,. \tag{76}$$

The uncorrelated increments property of the martingale process allows us to establish the following generalized version of (??)

$$\begin{aligned}&\mathbf{Cov}\left[\int_0^t g_1(u)dM(u),\, \int_0^t g_2(u)dM(u),\, |\mathbf{F}_t\right] \\ &= \int_0^t g_1(u)g_2(u)\lambda(u)du\end{aligned}\,, \tag{77}$$

where $g_1(t)$, $g_2(t)$ are measurable functions.

Appendix B

To prove the results of this section we need the following elementary inequalities

$$\frac{x-1}{x} \le \log(x) \le x-1,\;\; x>0\,. \tag{78}$$

The tighter version of this inequality for $x \ge 1$ reads as follows

$$2\frac{x-1}{x+1} \le \log(x) \le \frac{x^2-1}{2x},\;\; x \ge 1\,. \tag{79}$$

Proof of Lemma ??. Let $\mathbf{X} \in \omega_1$. Then, the formula for the threshold value α_T in (??) becomes

$$\alpha_T = \tau_1 - \tau_2 + \tau_1 \log\left(\frac{\tau_2}{\tau_1}\right) - \tau_1 \int_0^T \log\left(\frac{p_1(t)}{p_2(t)}\right) p_1(t)dt\,. \tag{80}$$

Here $\mathbf{K}_T\left(p_1 \parallel p_2\right) = \int_0^T \log\left(\frac{p_1(t)}{p_2(t)}\right) p_1(t)dt$ is the Kullback-Leibler divergence between the densities $p_1(t)$ and $p_2(t)$. Then, by virtue of (??) we have

$$\begin{aligned}\alpha_T &\le \tau_1 - \tau_2 + \tau_1\left\{\frac{\tau_2}{\tau_1} - 1\right\} - \tau_1\mathbf{K}_T(p_1 \parallel p_2) \\ &= -\tau_1\mathbf{K}_T(p_1 \parallel p_2)\end{aligned}.$$

As $\mathbf{K}_T(p_1 \parallel p_2) \ge 0$ we conclude that $\alpha_T \le 0$. Concerning the lower bound for α_T in (??) we again use (??). Hence,

$$\begin{aligned}\alpha_T &\ge \tau_1 - \tau_2 + \tau_1\left\{1 - \frac{\tau_1}{\tau_2}\right\} - \tau_1\mathbf{K}_T(p_1 \parallel p_2) \\ &= \frac{(\tau_1 - \tau_2)^2}{\tau_2} - \tau_1\mathbf{K}_T(p_1 \parallel p_2)\end{aligned}.$$

This confirms the inequalities in Lemma ?? (a). The case when $\mathbf{X} \in \omega_2$ can be proved in the analogous way by noting that α_T is now equal to

$$\alpha_T = \tau_1 - \tau_2 + \tau_2\log\left(\frac{\tau_2}{\tau_1}\right) + \tau_2\mathbf{K}_T(p_2 \parallel p_1)\,.$$

Then, the application of (??) gives the result in Lemma ?? (b). □

Proof of Lemma ??. The result of Lemma ?? is implied by the straightforward application of the identity in (??) to the stochastic integral

$$\int_0^T \log\left(\frac{\lambda_1(t)}{\lambda_2(t)}\right) dM(t)\,.$$

Here $M(t)$ is the local martingale corresponding to the intensity $\lambda_1(t)$ or $\lambda_2(t)$ depending whether $\mathbf{X} \in \omega_1$ or $\mathbf{X} \in \omega_2$, respectively. □

Proof of Lemma ??. The result in (??) of Lemma ?? is the version of Theorem 5 in [?] that says that under the conditions (a) and (b) of Lemma ?? we have

$$\mathbf{P}\left(|U_T| \ge \epsilon\right) \le 2\exp\left[-\frac{v_T}{u_T^2} J\left(\epsilon\frac{u_T}{v_T}\right)\right]\,, \tag{81}$$

where $J(x) = (1+x)\log(1+x) - x,\ x \ge 0$. Using the inequalities in (??) we can easily obtain that

$$\frac{x^2}{2+x} \le J(x) \le \frac{x^2}{2}\,.$$

The application of the above lower bound in (??) leads to the version of (??) given in (??). □

Proof of Theorem ??. We will prove the result in (??). This clearly implies the convergence $\mathbf{R}_T^* \to 0$ as $T \to \infty$. By virtue of (??) it suffices to consider the probability of misclassification $\mathbf{P}\left(W_T(\mathbf{X}) \ge \eta_T | \mathbf{X} \in \omega_2\right)$. As it has been observed in (??) this probability is equivalent to the following probability

$$\mathbf{P}\left(\frac{1}{T}U_T(\mathbf{X}) \ge \frac{1}{T}\alpha_T + \frac{1}{T}\kappa | \mathbf{X} \in \omega_2\right)\,, \tag{82}$$

where $\kappa = \log(\pi_2/\pi_1)$. Since $\mathbf{X} \in \omega_2$ then

$$\alpha_T = \tau_1 - \tau_2 + \tau_2\log\left(\frac{\tau_2}{\tau_1}\right) + \tau_2\mathbf{K}_T(p_2 \parallel p_1)\,. \tag{83}$$

Then, by the Chebyshev inequality the probability in (??) is bounded by

$$b_T\frac{1}{T}\,, \tag{84}$$

where

$$b_T = \frac{\mathbf{Var}\left[\frac{1}{\sqrt{T}}U_T(\mathbf{X})\right]}{\left[T^{-1}\alpha_T + T^{-1}\kappa\right]^2}\,. \tag{85}$$

By the assumptions ***A1*** and ***A2*** we have

$$\overline{\lim}_{T\to\infty}\alpha_T/T \le d\log\left(\frac{C}{\delta}\right) \tag{86}$$

and also

$$\underline{\lim}_{T\to\infty}\alpha_T/T \geq d\log\left(\frac{\delta}{C}\right)\,. \tag{87}$$

Then by the result of Lemma ?? and (??), we get

$$\overline{\lim}_{T\to\infty}b_T \leq \frac{d\log^2\left(\frac{C}{\delta}\right)}{d^2\log^2\left(\frac{\delta}{C}\right)} = \frac{1}{d}\left(\frac{\log(C/\delta)}{\log(\delta/C)}\right)^2\,. \tag{88}$$

Hence, the probability of misclassification $\mathbf{P}\left(W_T(\mathbf{X}) \geq \eta_T|\mathbf{X}\in\omega_2\right)$ is bounded by b_T/T, where the superior limit of b_T is given in (??). In the analogous way one can show that the probability of misclassification $\mathbf{P}\left(W_T(\mathbf{X}) < \eta_T|\mathbf{X}\in\omega_1\right)$ is bounded by a_T/T, where the superior limit of a_T is also given by (??). This concludes the proof of Theorem ??. □

Proof of Theorem ??. Consider again $\mathbf{P}\left(W_T(\mathbf{X}) \geq \eta_T|\mathbf{X}\in\omega_2\right)$ or equivalently the probability in (??). We wish to use the exponential inequality in (??) of Lemma ??. Then, the probability in (??) is bounded by

$$\exp\left[-T\frac{\epsilon_T^2}{2\theta_T + u\epsilon_T}\right]\,, \tag{89}$$

where $u = \log\left(\frac{C}{\delta}\right)$ characterizes the assumption $\boldsymbol{A1}$, $\epsilon_T = \frac{1}{T}\alpha_T + \frac{1}{T}\kappa$, and $\theta_T = \mathbf{Var}\left[\frac{1}{\sqrt{T}}U_T(\mathbf{X})\right]$. This defines the exponential factor

$$B_T = \frac{\epsilon_T^2}{2\theta_T + u\epsilon_T}\,.$$

Owing to Lemma ??, (??) and (??), the limit inferior of B_T is not smaller than

$$\begin{aligned}\underline{\lim}_{T\to\infty}B_T &\geq \frac{d^2\log^2(\delta/C)}{2d\log^2(C/\delta) + ud\log(C/\delta)}\\ &= d\frac{1}{3}\left(\frac{\log(\delta/C)}{\log(C/\delta)}\right)^2\end{aligned}\,.$$

This combined with (??) gives the required bound. Since the analogous analysis can be carried out for the probability of misclassification $\mathbf{P}\left(W_T(\mathbf{X}) < \eta_T|\mathbf{X}\in\omega_1\right)$ therefore the proof of Theorem ?? has been completed. □

Appendix C

Proof of Theorem ??. The proof of Theorem ?? is in the spirit of the proof of Theorem 1 in [?]. Hence, the consistency results established in (??) and (??) imply that for the selected $\delta > 0$ there exists l_0 such that for $L > l_0$ and $\epsilon > 0$ we have

$$\begin{aligned}&\mathbf{P}\left(\left|\widehat{W}_{L,T}(\mathbf{x}) - W_T(\mathbf{x})\right| < \epsilon\right) > 1 - \delta/2\,,\\ &\mathbf{P}\left(\left|\widehat{\eta}_{L,T} - \eta_T\right| < \epsilon\right) > 1 - \delta/2\,.\end{aligned} \tag{90}$$

Let $\psi_T^\star(\mathbf{x}) = \omega_1$, i.e., we have $W_T(\mathbf{x}) > \eta_T$. Then,

$$\mathbf{P}\left(\widehat{\psi}_{L,T}(\mathbf{x}) = \psi_T^\star(\mathbf{x})\right) = \mathbf{P}\left(\widehat{W}_{L,T}(\mathbf{x}) > \widehat{\eta}_{L,T}\right)\,.$$

The right-hand side of this equality is not smaller than

$$\mathbf{P}\left(\left|\left(\widehat{W}_{L,T}(\mathbf{x}) - \widehat{\eta}_{L,T}\right) - \left(W_T(\mathbf{x}) - \eta_T\right)\right| < 2\epsilon\right) \tag{91}$$

for $0 < \epsilon < \frac{1}{2}\left(W_T(\mathbf{x}) - \eta_T\right)$. Moreover, (??) is bounded from below by

$$\mathbf{P}\left(\left|\widehat{W}_{L,T}(\mathbf{x}) - W_T(\mathbf{x})\right| < \epsilon,\ |\widehat{\eta}_{L,T} - \eta_T| < \epsilon\right)\,. \tag{92}$$

In turn by the elementary inequality $\mathbf{P}(A\cap B) \geq \mathbf{P}(A) + \mathbf{P}(B) - 1$, the lower bound for (??) is

$$\mathbf{P}\left(\left|\widehat{W}_{L,T}(\mathbf{x}) - W_T(\mathbf{x})\right| < \epsilon\right) + \mathbf{P}\left(|\widehat{\eta}_{L,T} - \eta_T| < \epsilon\right) - 1\,.$$

Recalling (??) we have shown that for $L > l_0$

$$\mathbf{P}\left(\widehat{\psi}_{L,T}(\mathbf{x}) = \psi_T^\star((x))\right) > 1 - \delta\,.$$

Since we can choose an arbitrary small δ, this confirms the claimed convergence. □

Proof of Lemma ??. The proof will be based on the following version of Helly's thorem [?] for the Stieltjes integral.

Let
$$f_L(x) \to f(x) \text{ uniformly on } [a,b] \text{ as } L \to \infty\,.$$
If $g(x)$ is a function of bounded variation on $[a,b]$ then
$$\int_a^b f_L(x)dg(x) \to \int_a^b f(x)dg(x) \text{ as } L \to \infty\,. \tag{93}$$
Consider the optimal decision function $W_T(\mathbf{x})$ in (??) and its empirical counterpart $\widehat{W}_{L,T}(\mathbf{x})$ in (??). Then, we can write (see (??))
$$\begin{aligned}&\widehat{W}_{L,T}(\mathbf{x}) - W_{L,T}(\mathbf{x})\\ &= \int_0^T \left[\log\left(\frac{\widehat{p}_1(t)}{\widehat{p}_2(t)}\right) - \log\left(\frac{p_1(t)}{p_2(t)}\right)\right] dN(t)\end{aligned}. \tag{94}$$
We wish to prove that $\left|\widehat{W}_{L,T}(\mathbf{x}) - W_{L,T}(\mathbf{x})\right|$ (P) as $L \to \infty$. Owing to Helly's theorem it suffices to show that
$$\begin{aligned}&\left|\log\left(\frac{\widehat{p}_1(t)}{\widehat{p}_2(t)}\right) - \log\left(\frac{p_1(t)}{p_2(t)}\right)\right| \to 0 \;\; (P)\\ &\text{uniformly on } [0,T] \text{ as } L \to \infty\end{aligned}. \tag{95}$$
Observe that the left-hand side of (??) is equal to $\left|\log\left(\frac{\widehat{p}_1(t)p_2(t)}{\widehat{p}_2(t)p_1(t)}\right)\right|$. Then, using (??) this is bounded by
$$\begin{aligned}&\left|\frac{\widehat{p}_1(t)p_2(t) - \widehat{p}_2(t)p_1(t)}{\widehat{p}_2(t)p_1(t)}\right|\\ &= \left|\frac{(\widehat{p}_1(t) - p_1(t))\,p_2(t) + (p_2(t) - \widehat{p}_2(t))\,p_1(t)}{(\widehat{p}_2(t) - p_2(t))\,p_1(t) + p_1(t)p_2(t)}\right|\end{aligned}.$$
This is not greater than
$$\frac{|\widehat{p}_1(t) - p_1(t)|\,p_2(t) + |\widehat{p}_2(t) - p_2(t)|\,p_1(t)}{p_1(t)p_2(t)}\,.$$
By the assumption $\boldsymbol{A1}$ limited to the interval $[0,T]$ and the fact that $p_i(t) = \lambda_i(t)/\tau_i$ the above expression does not exceed
$$\left(\frac{C}{\delta}\right)^3 T\left\{|\widehat{p}_1(t) - p_1(t)| + |\widehat{p}_2(t) - p_2(t)|\right\}\,.$$
This by recalling the assumption in (??) proves (??). The proof of Lemma ?? has been completed. □

Appendix D

Proof of Lemma ??. We wish to show that
$$\sup_{t\in T_\epsilon}\left|\widehat{\lambda}(t) - \lambda(t)\right| \to 0 \;\; (P) \text{ as } L \to \infty\,. \tag{96}$$
where $T_\epsilon = [\epsilon, T-\epsilon]$ for small $\epsilon > 0$. We begin with the standard bounding into the variance and bias terms
$$\left|\widehat{\lambda}(t) - \lambda(t)\right| \le \left|\widehat{\lambda}(t) - \mathbb{E}\left[\widehat{\lambda}(t)\right]\right| + \left|\mathbb{E}\left[\widehat{\lambda}(t)\right] - \lambda(t)\right|\,. \tag{97}$$
Owing to (??) the bias term is equal to
$$\mathbb{E}\left[\widehat{\lambda}(t)\right] - \lambda(t) = \int_{-t/h}^{(T-t)/h} K(s)\lambda\left(t+hs\right)ds - \lambda(t)$$
for $t \in T_\epsilon$. Since $K(t)$ and $\lambda(t)$ are positive and $K(t)$ is a density function supported on $[-1,1]$ then we have
$$\begin{aligned}\left|\mathbb{E}\left[\widehat{\lambda}(t)\right] - \lambda(t)\right| &= \int_{-1}^{1} K(s)\left|\lambda\left(t+hs\right) - \lambda(t)\right|ds\\ &\le M_\lambda h\int_{-1}^{1} K(s)|s|ds\end{aligned}$$
uniformly in $t \in T_\epsilon$, where M_λ is the Lipschitz constant of $\lambda(t)$.
Let us consider the stochastic part in (??). As the interval T_ϵ is compact, one can define a finite partition of T_ϵ into disjoint equal size intervals, i.e., $T_\epsilon = \bigcup_{j=1}^{q(L)} \mathbf{U}_j$, where the size of $\mathbf{U}_j$ is denoted as $\Delta(L)$. Clearly the number of intervals

is of order $T_\epsilon/\Delta(L)$. Let $u_j \in \mathbf{U}_j$ be the middle point of $\mathbf{U}_j$. Then, the uniform norm of the stochastic term in (??) can be bounded as follows

$$\begin{aligned}
&\sup_{t\in T_\epsilon}\left|\widehat{\lambda}(t)-\mathbb{E}\left[\widehat{\lambda}(t)\right]\right| \\
&\le \max_{1\le j\le q(L)}\sup_{t\in T_\epsilon\cap\mathbf{U}_j}\left|\widehat{\lambda}(t)-\widehat{\lambda}(u_j)\right| \\
&+\max_{1\le j\le q(L)}\sup_{t\in T_\epsilon\cap\mathbf{U}_j}\left|\mathbb{E}\left[\widehat{\lambda}(t)\right]-\mathbb{E}\left[\widehat{\lambda}(u_j)\right]\right| . \\
&+\max_{1\le j\le q(L)}\left|\widehat{\lambda}(u_j)-\mathbb{E}\left[\widehat{\lambda}(u_j)\right]\right| \\
&= A_1+A_2+A_3.
\end{aligned} \tag{98}$$

Consider first the term A_1. By virtue of (??) we have

$$\widehat{\lambda}(t)-\widehat{\lambda}(u_j)=\int_0^T\left[K_h(t-s)-K_h(u_j-s)\right]d\overline{N}_L(s)$$

for $t,u_j\in\mathbf{U}_j$. Noting that $|t-u_j|\le\Delta(L)$ and using the fact that $K(t)$ is Lipschitz we get

$$\left|\widehat{\lambda}(t)-\widehat{\lambda}(u_j)\right|\le M_K\frac{\Delta(L)}{h^2}\int_0^T d\overline{N}_L(s)\,. \tag{99}$$

Note that $\int_0^T d\overline{N}_L(s)=\overline{N}_L(T)$ and we know, see (??), that $\mathbb{E}\left[\overline{N}_L(T)\right]=\int_0^T\lambda(t)dt$ and $\mathbf{Var}\left[\overline{N}_L(T)\right]=\frac{1}{L}\int_0^T\lambda(t)dt$. This proves that

$$\left|\widehat{\lambda}(t)-\widehat{\lambda}(u_j)\right|=\mathcal{O}\left(\frac{\Delta(L)}{h^2}\right)\quad(a.s.) \tag{100}$$

uniformly in $t\in T_\epsilon$. Concerning the term A_2 in (??) we can use (??). Then, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\widehat{\lambda}(t)\right]-\mathbb{E}\left[\widehat{\lambda}(u_j)\right] \\
&=\int_0^T\left[K_h(t-s)-K_h(u_j-s)\right]\lambda(s)ds
\end{aligned}.$$

This gives

$$A_2=\mathcal{O}\left(\frac{\Delta(L)}{h^2}\right)\,. \tag{101}$$

Hence, we have shown that the terms A_1 and A_2 are of order $\mathcal{O}\left(\frac{\Delta(L)}{h^2}\right)$, where $\Delta(L)$ is to be selected.

Finally, let us consider the term A_3 in (??). First we note that for $\delta>0$

$$\begin{aligned}
&\mathbf{P}\left(\max_{1\le j\le q(L)}\left|\widehat{\lambda}(u_j)-\mathbb{E}\left[\widehat{\lambda}(u_j)\right]\right|\ge\delta\right) \\
&\le q(L)\sup_{t\in T_\epsilon}\mathbf{P}\left(\left|\widehat{\lambda}(t)-\mathbb{E}\left[\widehat{\lambda}(t)\right]\right|\ge\delta\right)
\end{aligned}. \tag{102}$$

By virtue of (??) and (??) we have

$$\widehat{\lambda}(t)-\mathbb{E}\left[\widehat{\lambda}(t)\right]=\int_0^T K_h(t-s)d\overline{M}_L(s)\,.$$

This, (??) and Chebyshev inequality yield

$$\mathbf{P}\left(\left|\widehat{\lambda}(t)-\mathbb{E}\left[\widehat{\lambda}(t)\right]\right|\ge\delta\right)\le\frac{1}{L}\int_0^T K_h^2(t-s)\lambda(s)ds/\delta^2\,.$$

Note that the right-hand side of this inequality is of order $\mathcal{O}(1/Lh)$ uniformly in $t\in T_\epsilon$. This, (??) and the fact that $q(L)=\mathcal{O}(1/\Delta(L))$ lead to the following uniform bound

$$\mathbf{P}\left(A_3\ge\delta\right)=\mathcal{O}\left(1/\Delta(L)Lh\right)$$

or equivalently $A_3=\mathcal{O}_P\left(1/\sqrt{\Delta(L)Lh}\right)$. Hence, balancing $A_3=\mathcal{O}_P\left(1/\sqrt{\Delta(L)Lh}\right)$ versus $A_1,A_2=\mathcal{O}(\Delta(L)/h^2)$ gives the choice $\Delta(L)=h/L^{1/3}$. This yields the convergence in (??) if

$$h(L)\to 0\ \text{ and }\ Lh^3(L)\to\infty\ \text{ as }\ L\to\infty\,.$$

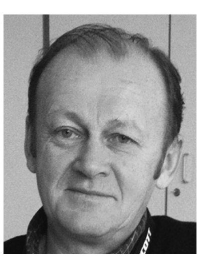
The proof of Lemma ?? has been completed. □

Acknowledgment

This work was supported by the Polish National Center of Science under Grant DEC-2017/27/B/ST7/03082 and NSERC Grant 319732.

References

[1] W. Gerstner, W. M. Kistler, R. Naud, and L. Paninski, Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition. Cambridge University Press, 2014.
[2] H. Jang, O. Simeone, B. Gardner, and A. Gruning, “An introduction to probabilistic spiking neural networks: Probabilistic models, learning rules, and applications,” IEEE Signal Processing Magazine, vol. 36, no. 6, pp. 64–77, 2019.
[3] O. Shchur, A. C. Türkmen, T. Januschowski, and S. Günnemann, “Neural temporal point processes: A review,” arXiv preprint arXiv:2104.03528, 2021.
[4] I. Bar-David, “Communication under the Poisson regime,” IEEE Transactions on Information Theory, vol. 15, pp. 31–37, 1969.
[5] D. Guo, S. Shamai, and S. Verdu, “Mutual information and conditional mean estimation in Poisson channels,” IEEE Transactions on Information Theory, vol. 54, pp. 1837–1849, 2008.
[6] N. Merhav, “Optimal correlators for detection and estimation in optical receivers,” IEEE Transactions on Information Theory, vol. 67, pp. 5200–5210, 2021.
[7] D. J. Daley and D. Vere-Jones, An Introduction to the Theory of Point Processes. Springer, 2003.
[8] P. K. Andersen, O. Borgan, R. D. Gill, and N. Keiding, Statistical Models Based on Counting Processes. Springer, 2012.
[9] L. Devroye, L. Györfi, and G. Lugosi, A Probabilistic Theory of Pattern Recognition. Springer, 1996.
[10] A. Cholaquidis, L. Forzani, P. Llop, and L. Moreno, “On the classification problem for Poisson point processes,” Journal of Multivariate Analysis, vol. 153, pp. 1–15, 2017.
[11] X. Rong and V. Solo, “On the error rate for classifying point processes,” in 60th IEEE Conference on Decision and Control (CDC). IEEE, 2021, pp. 120–125.
[12] M. Mazza, S. Cresci, M. Avvenuti, W. Quattrociocchi, and M. Tesconi, “RTbust: Exploiting temporal patterns for botnet detection on twitter,” in Proceedings of the 10th ACM conference on web science, 2019, pp. 183–192.
[13] H. Fanaee-T and J. Gama, “Event labeling combining ensemble detectors and background knowledge,” Progress in Artificial Intelligence, vol. 2, no. 2, pp. 113–127, 2014.
[14] W. Greblicki, “Asymptotically optimal pattern recognition procedures with density estimates,” IEEE Transactions on Information Theory, vol. 24, no. 2, pp. 250–251, 1978.
[15] P. Diggle and J. S. Marron, “Equivalence of smoothing parameter selectors in density and intensity estimation,” Journal of the American Statistical Association, vol. 83, no. 403, pp. 793–800, 1988.
[16] O. Aalen, “Nonparametric inference for a family of counting processes,” The Annals of Statistics, pp. 701–726, 1978.
[17] M. Pawlak, M. Pabian, and D. Rzepka, “Asymptotically optimal nonparametric classification rules for spike train data,” in ICASSP 2023-IEEE International Conference on Acoustics, Speech and Signal Processing, 2023, pp. 1–5.
[18] Á. Gajardo and H.-G. Müller, “Cox point process regression,” IEEE Transactions on Information Theory, vol. 68, no. 2, pp. 1133–1156, 2022.
[19] L. Birgé and P. Massart, “Rates of convergence for minimum contrast estimators,” Probability Theory and Related Fields, vol. 97, pp. 113–150, 1993.
[20] S. Van de Geer, “Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes,” The Annals of Statistics, pp. 1779–1801, 1995.
[21] S. Ghosal and A. Van der Vaart, Fundamentals of Nonparametric Bayesian Inference. Cambridge University Press, 2017.
[22] R. Le Guével, “Exponential inequalities for the supremum of some counting processes and their square martingales,” Comptes Rendus. Mathématique, vol. 359, no. 8, pp. 969–982, 2021.
[23] D. Vere-Jones, “On the estimation of frequency in point-process data,” Journal of Applied Probability, vol. 19, pp. 383–394, 1982.
[24] M. P. Wand and M. C. Jones, Kernel Smoothing. CRC Press, 1994.
[25] W. Greblicki and M. Pawlak, Nonparametric System Identification. Cambridge University Press, 2008.
[26] M. D. Cattaneo, M. Jansson, and X. Ma, “Simple local polynomial density estimators,” Journal of the American Statistical Association, vol. 115, no. 531, pp. 1449–1455, 2020.
[27] E. Giné and A. Guillou, “Rates of strong uniform consistency for multivariate kernel density estimators,” Annales de l’Institut Henri Poincare (B) Probability and Statistics, vol. 38, no. 6, pp. 907–921, 2002.
[28] E. Masry, “Multivariate regression estimation local polynomial fitting for time series,” Stochastic Processes and their Applications, vol. 65, no. 1, pp. 81–101, 1996.
[29] J.-Y. Audibert and A. B. Tsybakov, “Fast learning rates for plug-in classifiers,” The Annals of Statistics, vol. 35, pp. 608–633, 2007.
[30] R. Lima, “Hawkes processes modeling, inference, and control: an overview,” SIAM Review, vol. 65, pp. 331–374, 2023.
[31] T. Apostol, Mathematical Analysis. Addison-Wesley, 1974.



Mirosław Pawlak received the Ph.D. (under the supervision of Prof. Greblicki) and D.Sc. degrees in computer engineering from Wrocław University of Technology, Wrocław, Poland, in 1984 and 2006, respectively. He is currently a Professor with the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB, Canada. He has held a number of visiting positions in North American, Australian, and European Universities. He was with the University of Ulm and University of Goettingen as an Alexander von Humboldt Foundation Fellow. Among his publications in these areas are the books Image Analysis by Moments (Wrocław Univ. Technol. Press, 2006), and Nonparametric System Identification (Cambridge Univ. Press, 2008), coauthored with Prof. Włodzimierz Greblicki. His research interests include statistical signal processing, machine learning, and nonparametric modeling. Dr. Pawlak has been an Associate Editor for the Journal of Pattern Recognition and Applications, Pattern Recognition, International Journal on Sampling Theory in Signal and Image Processing, Opuscula Mathematica and Statistics in Transition-New Series.

Mateusz Pabian received the M.Sc. degree in biomedical engineering from the AGH University of Science and Technology, Kraków, Poland, in 2019, where he is currently pursuing the Ph.D. degree at the Department of Measurement and Electronics. In 2017 he has completed an internship at NTT Communication Science Laboratories as part of the Vulcanus in Japan 2016-2017 Programme, which focused on experiment design, signal acquisition and analysis of the auditory evoked potentials. From 2017 to 2022 he was a Machine Learning Researcher at Comarch Healthcare, where he has been involved in the development of computer-aided medical diagnostics systems. Since 2022 he has been working as Model Developer at UBS in Kraków, Poland. His research interests include machine learning and event-based systems, particularly spiking neural networks.

Dominik Rzepka received his M.Sc. and Ph.D. degree in electrical engineering from the AGH University of Science and Technology, Kraków, Poland, in 2009 and 2018 respectively. From 2007 to 2011, he was with the Wireless Sensor and Control Networks Group, AGH University of Science and Technology, where he was involved in the design of the low-power algorithms for the processing of radio signals and software-defined radio. In 2011, he joined the Event-Based Control and Signal Processing Group at AGH University of Science and Technology, where he currently works on methods of signals reconstruction from event-triggered samples and on neuromorphic machine learning. He was a Visiting Student and Postdoc Researcher in the University of Manitoba, Winnipeg, Canada, from 2014 to 2023, and in The City College of New York, USA, in 2015. Since 2015, he is working as Signal Processing and Machine Learning Researcher in Comarch Healthcare and Fitech, developing algorithms for diagnostics and quality assurance systems. His research interests include signal processing and machine learning in biomedicine, wireless communication and industrial inspection, and event-based systems.