

Nonparametric maximum likelihood estimation of intensities for covariate-driven Poisson processes

Atsutomo Yara¹ Yoshikazu Terada^{1,2}

¹The University of Osaka ²RIKEN AIP

Abstract

We study nonparametric maximum likelihood estimation of intensity functions for Poisson processes with covariates. The observations are independent pairs consisting of a covariate vector and a point pattern on a fixed domain. We establish an oracle inequality for the squared Hellinger loss averaged over the covariate distribution, without imposing a common positive lower bound on the candidate intensities. For deep ReLU networks, this inequality yields convergence rates that depend on the compositional structure of the intensity and its behavior near zero. On a Euclidean cube, the upper bound matches a minimax lower bound up to logarithmic factors when no additional condition is imposed on the intensity near zero. We also obtain convergence rates that depend on the intrinsic dimension when the observation domain is a known smooth manifold. Numerical examples compare neural network and kernel estimators. An application to stimulus-evoked spike trains from one recording session shows lower average negative log-likelihood on the test data for the neural network estimators than for the conditional kernel estimator.

1 Introduction

Point patterns observed under different experimental conditions often exhibit systematic variation in both the number and the locations of events. For example, in visual neural recordings, the position and orientation of a stimulus may affect spike counts and the distribution of spike times. A Poisson intensity function that depends on these covariates provides a model for both aspects of the response.

We consider independent observations, each consisting of a covariate vector and a point pattern on a fixed domain. The sample size is the number of independent point patterns, rather than the total number of events or the size of an expanding observation window. Our aim is to estimate the conditional intensity function nonparametrically and to derive convergence rates for maximum likelihood estimators based on deep neural networks.

Intensity estimation for Poisson processes with covariates has been studied using model selection and Hellinger-type loss (Sart, 2015). For neural network estimators, compositional structure is known to yield faster convergence rates in nonparametric regression (Schmidt-Hieber, 2020; Kohler and Langer, 2021). In the present setting, an additional difficulty arises when candidate intensities are not bounded away from zero: the excess risk associated with the negative log-likelihood may be infinite. We therefore analyze the squared Hellinger loss directly. The resulting oracle inequality requires a suitable approximating intensity for which the ratio of the true intensity to the approximation is bounded, but it does not require a common positive lower bound on all candidate intensities.

We apply this inequality to intensities with a compositional structure and to smooth intensities defined on a known manifold. The convergence rates depend on the dimensions of the component functions or the intrinsic dimension of the manifold, together with a condition on the measure of the sets where the intensity is small. We present the main theoretical results, numerical examples, and an application to neural spike trains. Detailed proofs are omitted.

2 Model and maximum likelihood estimation

2.1 Poisson process with covariates

Let $A \subset \mathbb{R}^{d_x}$ be compact, and let ν be a known finite reference measure on A . We observe independent and identically distributed pairs (Z_i, X_i) , $i = 1, \dots, n$, where $Z_i \in [0, 1]^{d_z}$ has distribution P_Z and

$$X_i \mid Z_i = z \sim \text{PP}\{S_0(x \mid z) d\nu(x)\}, \quad \lambda_0(B \mid z) = \int_B S_0(x \mid z) d\nu(x). \quad (1)$$

Here PP denotes a Poisson point process, S_0 is bounded, nonnegative, and measurable, and $B \subset A$ is measurable. For $d_z > 0$, we assume that P_Z has a positive density on $[0, 1]^{d_z}$; $d_z = 0$ denotes the case without covariates. The covariates describe the conditions under which an entire point pattern is observed. Thus, the conditioning in (1) is on observed covariates, not on the history of the point process.

Write $N_i = X_i(A)$ and let $x_{i1}, \dots, x_{iN_i}$ denote the observed event locations. For a class $\mathcal{F}_n$ of nonnegative intensity functions, define

$$\ell_n(S) = \frac{1}{n} \sum_{i=1}^n \left\{ - \sum_{j=1}^{N_i} \log S(x_{ij} | Z_i) + \int_A S(x | Z_i) d\nu(x) \right\}, \quad \widehat{S}_n \in \arg \min_{S \in \mathcal{F}_n} \ell_n(S). \quad (2)$$

The criterion ℓ_n is the negative log-likelihood, up to terms independent of S (Leskelä, 2024). We use $\log 0 = -\infty$ and interpret an empty sum as zero. The theoretical results concern a measurable exact minimizer of (2), when one exists.

2.2 Hellinger loss

For finite intensity measures with densities S_1, S_2 relative to ν , define

$$H^2(S_1, S_2) = \frac{1}{2} \int_A (\sqrt{S_1} - \sqrt{S_2})^2 d\nu, \quad \widetilde{H}^2(S_1, S_2) = \mathbb{E}_Z[H^2(S_1(\cdot | Z), S_2(\cdot | Z))]. \quad (3)$$

The first expression is the squared Hellinger distance between the finite intensity measures; the second averages this distance over the covariate distribution. The distance does not depend on the choice of dominating measure and accounts for differences in total intensity as well as in the distribution of event locations. If P_λ denotes the law of a Poisson process with intensity measure λ , then

$$H^2(P_{\lambda_1}, P_{\lambda_2}) = 1 - \exp\{-H^2(\lambda_1, \lambda_2)\} \leq H^2(\lambda_1, \lambda_2).$$

Thus, convergence under (3) implies convergence of the corresponding conditional Poisson laws in squared Hellinger distance, averaged over P_Z (Leskelä, 2024).

For a candidate intensity S , the excess risk associated with the negative log-likelihood is

$$R(S) = \mathbb{E}_Z \left[\int_A \left\{ S_0 \log \frac{S_0}{S} + S - S_0 \right\} d\nu \right] \geq 2\widetilde{H}^2(S, S_0).$$

This risk is infinite if S is zero on a set with positive true intensity measure for a set of covariate values with positive P_Z -probability. Direct analysis of (3) avoids the need to assume that the excess risk is finite for every candidate intensity.

3 An oracle inequality under Hellinger loss

Assume that every $S \in \mathcal{F}_n$ satisfies $0 \leq S \leq F < \infty$. We compare $\widehat{S}_n$ with an intensity $S_n^* \in \mathcal{F}_n$ such that

$$S_0(x | z)/S_n^*(x | z) \leq c_0^2 \quad \text{for all } n, x, z, \quad (4)$$

where $c_0 > 0$ is fixed. We use $0/0 = 0$ and interpret a positive number divided by zero as ∞ . Condition (4) is imposed only on S_n^* , not on every member of $\mathcal{F}_n$. To obtain a useful risk bound, S_n^* must also approximate S_0 well.

Let $\mu = \nu \otimes P_Z$ and define

$$\mathcal{Q}_n(\delta) = \left\{ \sqrt{\frac{S + S_n^*}{2}} : S \in \mathcal{F}_n, \widetilde{H}^2\left(\frac{S + S_n^*}{2}, S_n^*\right) \leq \delta^2 \right\}.$$

Write $N_{2,B}(u, \mathcal{Q}_n(\delta), \mu)$ for the $L^2(\mu)$ bracketing number with bracket width at most u . For a sufficiently small fixed $c_* > 0$, depending only on c_0, F , and $\nu(A)$, set

$$J_n(\delta) = \left\{ \int_{c_*\delta^2}^\delta \sqrt{\log N_{2,B}\{u, \mathcal{Q}_n(\delta), \mu\}} du \right\} \vee \delta.$$

Theorem 1 (Oracle inequality). *Suppose that (4) holds and $0 \leq S \leq F$ for every $S \in \mathcal{F}_n$. Let $\Psi_n \geq J_n$ be such that $\Psi_n(\delta)/\delta^2$ is nonincreasing. If*

$$\sqrt{n} \delta_n^2 \geq C \Psi_n(\delta_n),$$

then, for all $\delta \geq \delta_n$,

$$\mathbb{P} \left(\tilde{H}^2(\hat{S}_n, S_0) > C(\delta^2 + \tilde{H}^2(S_n^*, S_0)) \right) \leq C \exp(-cn\delta^2). \quad (5)$$

Consequently,

$$\mathbb{E} \tilde{H}^2(\hat{S}_n, S_0) \leq C \{ \delta_n^2 + \tilde{H}^2(S_n^*, S_0) + n^{-1} \}. \quad (6)$$

The positive constants depend only on c_0 , F , and the fixed domain and reference measure. The expectation is taken over the observed sample.

The bound consists of an estimation term determined by the local bracketing entropy and an approximation term determined by S_n^* . It applies to any candidate class for which an approximation satisfying (4) and a suitable bracketing bound are available.

Proof sketch. The proof follows the likelihood approach of van de Geer (2000). Comparing the empirical likelihood at $\hat{S}_n$ and S_n^* and using concavity of the logarithm gives a basic inequality involving their midpoint and a localized empirical process. For point patterns, the log-likelihood contains a random, unbounded number of summands. Conditional on $Z_i = z$ and $N_i = m > 0$, the m event locations are independent with density proportional to $S_0(\cdot | z)$ (Last and Penrose, 2017). This representation, Poisson moment bounds, and (4) yield moment bounds for the localized log-likelihood ratios. A bracketing argument and a peeling argument give (5); integration of the tail bound gives (6).

4 Convergence rates for deep neural network estimators

4.1 Network class and the small value bound condition

Let $\sigma(t) = t \vee 0$. Denote by $\mathcal{F}_\sigma(L, r, B, F)$ the class of fully connected networks with input $(x, z) \in \mathbb{R}^{d_x + d_z}$, L hidden layers of width r , and ReLU activation in both the hidden and output layers. The weights and hidden-layer biases are bounded in absolute value by B , and the output is bounded above by F on the domain of interest. The ReLU output ensures that the estimated intensity is nonnegative.

To describe the behavior of the intensity near zero, suppose that

$$\mu\{(x, z) : S_0(x | z) \leq t\} \leq C_{\text{svb}} t^\alpha, \quad t > 0, \quad (7)$$

for some $\alpha \geq 0$, with a common constant C_{svb} over the class of true intensities under consideration. We refer to (7) as the small value bound condition, adapted from Bos and Schmidt-Hieber (2022) to intensity functions. It imposes no restriction when $\alpha = 0$ and $C_{\text{svb}} \geq \nu(A)$. For $\alpha > 0$, it bounds the measure of sets on which the intensity is small and excludes a zero set of positive μ -measure. A common positive lower bound on the intensities implies (7) for every fixed α , with a suitable constant.

Put $a = 1 + (\alpha \wedge 1)$. Given a uniform approximation $\tilde{S}_n$ with error of order ξ_n , define

$$S_n^* = (\tilde{S}_n \vee \xi_n) \wedge F = \xi_n + \sigma(\tilde{S}_n - \xi_n) - \sigma(\tilde{S}_n - F). \quad (8)$$

For $0 < \xi_n \leq \min\{1, F\}$ and $0 \leq S_0 \leq F$, truncation to $[\xi_n, F]$ preserves the order of the uniform approximation error. It can be implemented by a ReLU network with a fixed increase in depth and width. If $\|S_n^* - S_0\|_\infty \leq C_1 \xi_n$, then $S_0/S_n^* \leq 1 + C_1$, so (4) holds. Moreover, (7) gives

$$\tilde{H}^2(S_n^*, S_0) \lesssim \xi_n^a \{1 + \log(1/\xi_n)\}. \quad (9)$$

Here and below, $\lesssim$ denotes an inequality up to a positive constant independent of n . Thus, (9) converts a uniform approximation bound into the Hellinger approximation bound required in Theorem 1.

4.2 Intensities with a compositional structure

Suppose that

$$S_0 = g_q \circ \cdots \circ g_0, \quad g_i : [a_i, b_i]^{d^{(i)}} \longrightarrow [a_{i+1}, b_{i+1}]^{d^{(i+1)}}, \quad (10)$$

where $d^{(0)} = d_x + d_z$ and $d^{(q+1)} = 1$. Each component of g_i depends on at most $t_i \leq d^{(i)}$ coordinates and belongs to a Hölder ball of smoothness $\beta_i > 0$ and radius K . The interval endpoints are bounded in absolute value by K . Denote the resulting class of nonnegative functions by $\mathcal{G}(q, \mathbf{t}, \boldsymbol{\beta}, K)$ and its subclass satisfying (7) by $\mathcal{G}_\alpha$. All dimensions, smoothness parameters, and constants defining these classes are fixed.

Define

$$\beta_i^* = \beta_i \prod_{\ell=i+1}^q (\beta_\ell \wedge 1), \quad \phi_n = \max_{0 \leq i \leq q} n^{-a\beta_i^*/(a\beta_i^*+t_i)}.$$

The effective smoothness indices β_i^* account for the propagation of approximation error through the composition (Schmidt-Hieber, 2020).

Theorem 2 (Convergence rate under a composition assumption). *Let $S_0 \in \mathcal{G}_\alpha$ and $F \geq K$. Choose*

$$L \asymp \log n, \quad r \asymp \sqrt{n\phi_n}, \quad B \asymp \max_{0 \leq i \leq q} n^{2(\beta_i+1)/(a\beta_i^*+t_i)}.$$

The maximum likelihood estimator over $\mathcal{F}_\sigma(L, r, B, F)$ satisfies

$$\sup_{S_0 \in \mathcal{G}_\alpha} \mathbb{E}[\tilde{H}^2(\hat{S}_n, S_0)] \leq C\phi_n \log^3 n.$$

The constant may depend on its fixed structural parameters, the observation domain, the reference measure, and the intensity bound.

The exponent in this rate depends on the component dimensions t_i , rather than directly on the total input dimension. The proof combines Hölder approximation by networks with bounded parameters (Kohler and Langer, 2021; Egels and Castillo, 2025), approximation of composite functions (Schmidt-Hieber, 2020), and covering number bounds for neural networks (Suzuki, 2019), together with the construction in (8).

Theorem 3 (Minimax lower bound). *Let $A = [0, 1]^{d_x}$ with Lebesgue reference measure, and suppose that the covariate density is bounded above and below by positive constants. If $t_i \leq \min\{d^{(0)}, \dots, d^{(i-1)}\}$ for $i = 1, \dots, q$ and K is sufficiently large, then*

$$\inf_{\hat{S}_n} \sup_{S_0 \in \mathcal{G}(q, \mathbf{t}, \boldsymbol{\beta}, K)} \mathbb{E}[\tilde{H}^2(\hat{S}_n, S_0)] \geq c \max_{0 \leq i \leq q} n^{-\beta_i^*/(\beta_i^*+t_i)}.$$

The infimum is over all estimators.

For $\alpha = 0$, the upper and lower bounds match up to logarithmic factors. In particular, for a single β -smooth component with t inputs, the exponent is $\beta/(\beta + t)$. Under the stronger condition $\alpha \geq 1$, the upper-bound exponent becomes $2\beta/(2\beta + t)$. Theorem 3 establishes minimax optimality in the Euclidean setting when no additional condition is imposed near zero; it does not establish optimality for every $\alpha > 0$.

4.3 Intensities on a known smooth manifold

Let $\mathcal{M} \subset \mathbb{R}^{d_x}$ be a fixed compact C^∞ embedded manifold without boundary, of dimension $d_{\mathcal{M}}$, and let ν be its known Riemannian volume measure. Put $k = d_{\mathcal{M}} + d_z > 0$. Fix a finite atlas with coordinate maps $P_j : V_j \rightarrow \Omega_j$ obtained by restricting affine maps on the ambient space. The images Ω_j are open balls. Also fix a smooth partition of unity subordinate to these charts, with each support compactly contained in the corresponding chart domain.

For noninteger $\beta > 0$, define the Hölder class by the bound

$$\max_j \left\| S(P_j^{-1}(u) \mid z) \right\|_{C^{\lfloor \beta \rfloor, \beta - \lfloor \beta \rfloor}(\Omega_j \times [0, 1]^{d_z})} \leq K.$$

The norm is the maximum of the derivative suprema through order $\lfloor \beta \rfloor$, plus the maximum Hölder seminorm of the derivatives of that order, using Euclidean distance. Derivatives extend continuously to the boundary of the covariate cube. Let $\mathcal{H}_{\mathcal{M}, \alpha}^\beta(K, F, C_{\text{svb}})$ denote the subclass also satisfying $0 \leq S \leq F$ and (7).

Theorem 4 (Convergence rate on a manifold). *Under the preceding assumptions, set*

$$\phi_{n,\mathcal{M}} = n^{-a\beta/(a\beta+k)}, \quad L \asymp \log n, \quad r \asymp \sqrt{n\phi_{n,\mathcal{M}}}, \quad B \asymp n^{2(\beta+1)/(a\beta+k)}.$$

The maximum likelihood estimator over $\mathcal{F}_\sigma(L, r, B, F)$ satisfies, for all sufficiently large n ,

$$\sup_{S_0 \in \mathcal{H}_{\mathcal{M},\alpha}^\beta(K, F, C_{\text{svb}})} \mathbb{E}[\tilde{H}^2(\hat{S}_n, S_0)] \leq C\phi_{n,\mathcal{M}} \log^3 n.$$

The approximation is constructed by extending local Hölder functions, approximating them by neural networks, and combining the local approximations with a partition of unity and multiplication networks (Yarotsky, 2017). The network inputs are ambient coordinates, whereas the exponent in the convergence rate involves $d_{\mathcal{M}} + d_z$, not $d_x + d_z$. The likelihood uses the exact volume measure, and the architecture is chosen using k , β , and α .

5 Numerical examples

5.1 Simulation settings and estimation methods

We compare two neural network estimators, denoted by DNN–Softplus and DNN–ReLU, with kernel estimators. The examples include intensities with a compositional structure, intensities with regions of small values that depend on the covariates, and intensities on the unit circle and sphere. The Euclidean domains are $[0, 1]$ and $[0, 1]^2$. Covariates are uniform on $[0, 1]^{d_z}$, with $d_z \in \{0, 1, 5\}$, and the sample sizes are $n \in \{100, 316, 1000, 3162, 10000\}$. Each setting is repeated 20 times. The intensities are numerically calibrated to yield approximately 30 expected events per point pattern. At most two covariates affect the generating intensity, so the settings with $d_z = 5$ include irrelevant covariates.

Both neural network estimators use ReLU hidden layers and the indicated output activation, with 10^{-8} added to ensure positive fitted intensities. Training uses AdamW, stochastic quadrature with 128 points per point pattern at each update, and batches of 64 point patterns. We use 80% of the data for fitting and 20% for validation. The maximum number of epochs is 150 for the Euclidean examples and 300 for the manifold examples; network sizes increase with n .

For the Euclidean examples, the comparison methods are a conditional Gaussian kernel estimator and a pooled kernel estimator that ignores Z . Bandwidths are selected by five-fold cross-validation based on the Poisson likelihood, with whole point patterns assigned to folds. For bandwidth selection, the data are restricted to at most 512 point patterns and 128 events per pattern; the final kernel estimate uses all observations. On the manifolds, the kernel estimator uses the true geodesic distances and geometry. The neural network estimators use ambient coordinates and an integration mesh reconstructed from independent samples on the manifold and the observed events. Thus, the methods use different sources of geometric information, and the neural network estimators receive information about the manifold in addition to the observed events.

5.2 Results

We evaluate the loss in (3) using independent covariates and numerical quadrature. Table 1 reports the mean loss and its Monte Carlo standard error, computed as the sample standard deviation divided by $\sqrt{20}$. Within each repetition, all methods use the same simulated data and evaluation points. Figure 1 shows the results for intensities with a compositional structure.

The relative performance of the methods varies with sample size. For a one-dimensional compositional intensity with $d_z = 1$, the mean loss at $n = 100$ is 0.120 for the conditional kernel estimator and 0.728 for DNN–Softplus. At $n = 10000$, the corresponding values are 0.0329 and 0.00926. At this largest sample size, a neural network estimator has the smallest mean loss in each of the four compositional settings with covariates. The differences from the conditional kernel estimator are larger in the settings with irrelevant covariates.

Table 1. Mean squared Hellinger loss at $n = 10000$, with Monte Carlo standard errors in parentheses, based on 20 repetitions. The dimension of the observation domain is d_s , and the covariate dimension is d_z . For the circle and sphere, d_s is the intrinsic dimension. A dash indicates that the pooled estimator is not included in the comparison.

Intensity/domain	d_s	d_z	Softplus	ReLU	Kernel	Pooled
Composition	1	1	0.00926 (0.00073)	0.0104 (0.00096)	0.0329 (0.0048)	0.474 (0.007)
Composition	1	5	0.0194 (0.0011)	0.0203 (0.0011)	0.144 (0.01)	0.544 (0.011)
Composition	2	1	0.0205 (0.00071)	0.0213 (0.00099)	0.0441 (0.0055)	0.398 (0.0052)
Composition	2	5	0.0346 (0.0013)	0.0329 (0.0012)	0.182 (0.0062)	0.466 (0.0067)
Low intensity	1	1	0.0092 (0.0012)	0.0229 (0.0023)	0.0442 (0.00036)	2.61 (0.044)
Low intensity	1	5	0.0179 (0.001)	0.0359 (0.0032)	0.597 (0.0051)	2.67 (0.031)
Low intensity	2	1	0.00788 (0.00056)	0.0108 (0.00085)	0.0455 (0.0035)	0.516 (0.012)
Low intensity	2	5	0.0296 (0.0011)	0.0408 (0.0022)	0.408 (0.0082)	0.951 (0.014)
Circle	1	1	0.0057 (0.0005)	0.00643 (0.00051)	0.0052 (0.00028)	—
Circle	1	5	0.0105 (0.00066)	0.0108 (0.00057)	0.0244 (0.00093)	—
Sphere	2	1	0.0102 (0.00058)	0.00997 (0.00057)	0.0205 (0.00084)	—
Sphere	2	5	0.0184 (0.00077)	0.0182 (0.00075)	0.0392 (0.0014)	—

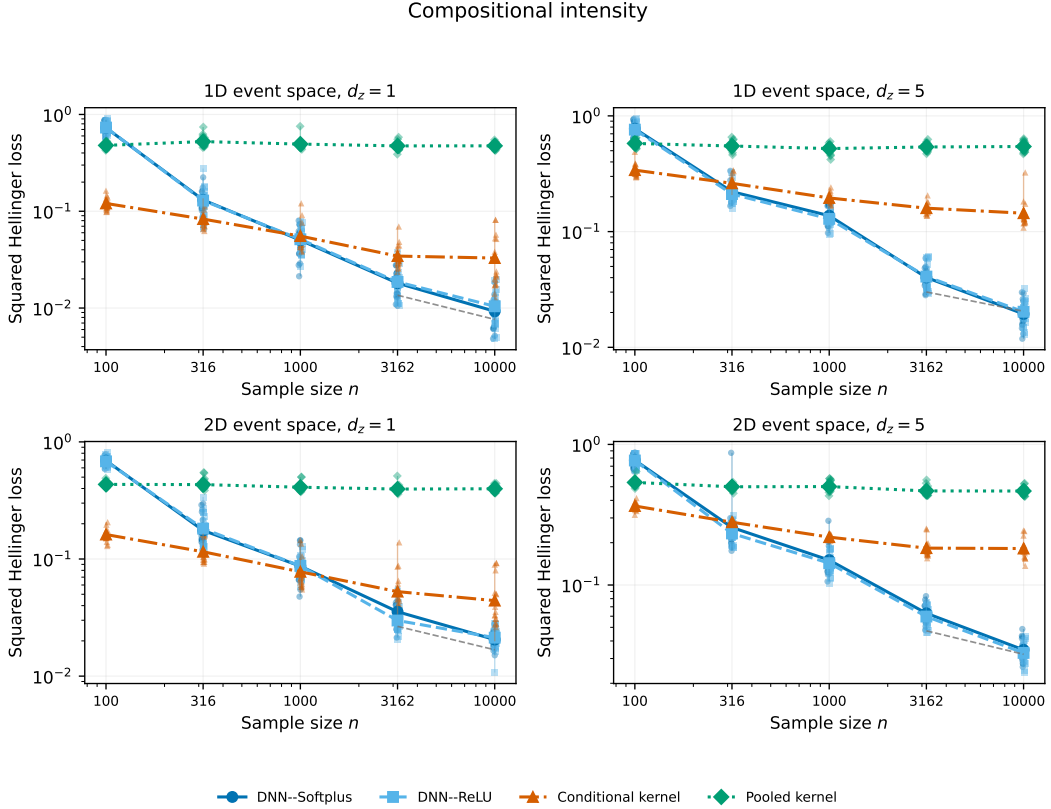


Figure 1. Squared Hellinger loss for intensities with a compositional structure. Small points show individual repetitions, connected markers show means, and vertical lines show the minimum and maximum over 20 repetitions.

The kernel estimator remains competitive in some manifold settings. On the circle with $d_z = 1$ and $n = 10000$, the mean loss is 0.00520 for the kernel estimator using the true geodesic distance, compared with 0.00570 for DNN-Softplus. For $d_z = 5$, the corresponding values are 0.0244 and 0.0105. Neither output activation gives a smaller mean loss in every setting, although Softplus performs better in many of the examples with low-intensity regions.

6 Application to stimulus-evoked spike trains

6.1 Data and estimation methods

We analyze Gabor receptive-field mapping data from Allen Brain Observatory Visual Coding Neuropixels session 715093703 (Siegle et al., 2021). The session was selected using a fixed rule based on metadata, before comparing the estimators. After excluding presentations that overlap invalid intervals and one presentation whose duration differs from the modal duration, 3,024 presentations remain. Each presentation provides a $T = 0.25$ -second observation window. The 243 stimulus conditions consist of 81 positions and three orientations. The covariates are horizontal and vertical position, normalized using the development data, and $\sin(2\theta)$, $\cos(2\theta)$ for orientation θ .

A fixed split, stratified by stimulus condition, assigns 2,417 presentations to a development set and 607 to a test set. All 243 conditions occur in both sets, so the evaluation concerns new presentations of conditions observed during model development, within the same session. We fit a separate intensity function for each of 12 primary visual cortex (VISp) units, selected from 62 eligible units by their development-set spike counts. The selected units are those with the largest counts. Test results for one unit (950931315) had been inspected in an earlier pilot analysis. We therefore also report a sensitivity analysis excluding this unit.

Both neural networks have three ReLU hidden layers of width 64 and use the output activations described in Section 5. The number of training epochs is selected using a fixed split of 1,935 presentations for fitting and 482 for validation, followed by refitting on all 2,417 development presentations. Kernel bandwidths are selected by five-fold cross-validation within the development set, with whole presentations assigned to folds. All methods are evaluated on the same test presentations.

6.2 Evaluation criterion and bootstrap intervals

For unit v and test presentation i , define the negative log-likelihood (NLL)

$$L_{vi}(\hat{S}) = \int_0^T \hat{S}_{v,t}(t | z_i) dt - \sum_{j=1}^{m_{vi}} \log \hat{S}_{v,t}(t_{vij} | z_i),$$

where t_{vij} are spike times in seconds and m_{vi} is the spike count. The models are fitted using normalized time $u = t/T$, with $\hat{S}_{v,t}(t | z) = \hat{S}_{v,u}(t/T | z)/T$. The integral is approximated by the trapezoidal rule on a deterministic grid of 1,024 points. Differences in NLL between methods are unchanged by this conversion of time units.

For each unit, we average the difference between the DNN and conditional-kernel NLL over the test presentations. We then average these differences across units with equal weights. Negative values indicate a lower NLL for the DNN. We compute 95% percentile bootstrap intervals using 2,000 resamples of the 243 stimulus conditions. Each resample retains all test presentations within a selected condition, and the same resampled conditions are used for all units and methods. The intervals are pointwise, with no adjustment for multiple comparisons.

6.3 Results

Both neural network estimators have a lower average test NLL than the conditional kernel estimator (Table 2). For the 12-unit analysis, the mean differences are -0.02901 and -0.02080 nats per presentation for Softplus and ReLU, respectively. The differences remain negative in the 11-unit sensitivity analysis. For comparison, the conditional kernel estimator reduces NLL by 0.38926 nats per presentation relative to the pooled estimator. Thus, the additional reduction obtained by the neural networks is smaller than the reduction associated with including stimulus covariates in the kernel estimator.

The results vary across units. The estimated NLL difference is negative for 10 of 12 units with Softplus and nine with ReLU, but many of the intervals for individual units include zero. The average difference between Softplus and ReLU is -0.00821 , with interval $[-0.01728, 0.00092]$.

Table 2. Difference in average test NLL between each DNN and the conditional kernel estimator, in nats per presentation, with equal weights across units. Negative values indicate lower NLL for the DNN. The intervals are 95% percentile bootstrap intervals obtained by resampling stimulus conditions. The 11-unit analysis excludes the unit examined in the earlier pilot.

Units	Output	Mean difference	95% interval	Units with lower NLL
12	Softplus	−0.02901	[−0.04042, −0.01729]	10/12
12	ReLU	−0.02080	[−0.03092, −0.01026]	9/12
11	Softplus	−0.02610	[−0.03791, −0.01378]	9/11
11	ReLU	−0.01954	[−0.02991, −0.00933]	8/11

The Poisson likelihood is used as a criterion for predictive comparison. These results do not establish that the spike trains follow the conditional Poisson model, and the analysis does not fully account for dependence on previous spikes or dependence between presentations.

7 Discussion

We have derived an oracle inequality for nonparametric maximum likelihood estimation of Poisson intensities under squared Hellinger loss. The inequality requires a reference intensity for which S_0/S_n^* is uniformly bounded, rather than a common positive lower bound on the candidate class. For deep ReLU networks, it gives convergence rates determined by compositional structure or intrinsic dimension and by the small value bound condition. In the Euclidean setting, the rate is minimax optimal up to logarithmic factors when no additional condition is imposed near zero.

The numerical examples and the spike-train analysis illustrate settings in which neural network estimators achieve lower loss than kernel estimators, as well as settings in which kernel estimators remain competitive. The theoretical bounds concern exact ReLU maximum likelihood estimators, whereas the numerical procedures use iterative optimization and numerical integration. Further work is needed to quantify optimization and integration errors, to treat unknown reference measures.

References

- Bos, T. and Schmidt-Hieber, J. (2022). Convergence rates of deep ReLU networks for multiclass classification. *Electronic Journal of Statistics* **16**, 2724–2773.
- Egels, P. and Castillo, I. (2025). Posterior and variational inference for deep neural networks with heavy-tailed weights. *Journal of Machine Learning Research* **26**(122), 1–58.
- Kohler, M. and Langer, S. (2021). On the rate of convergence of fully connected deep neural network regression estimates. *Annals of Statistics* **49**, 2231–2249.
- Last, G. and Penrose, M. (2017). *Lectures on the Poisson Process*. Cambridge University Press.
- Leskelä, L. (2024). Information divergences and likelihood ratios of Poisson processes and point patterns. *IEEE Transactions on Information Theory* **70**, 9084–9101.
- Sart, M. (2015). Model selection for Poisson processes with covariates. *ESAIM: Probability and Statistics* **19**, 204–235.
- Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *Annals of Statistics* **48**, 1875–1897.
- Siegle, J. H., Jia, X., Durand, S., Gale, S., Bennett, C., Graddis, N., et al. (2021). Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature* **592**, 86–92.
- Suzuki, T. (2019). Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. *International Conference on Learning Representations*.
- van de Geer, S. A. (2000). *Empirical Processes in M-estimation*. Cambridge University Press.
- Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Networks* **94**, 103–114.