NONPARAMETRIC POISSON REGRESSION
FROM INDEPENDENT AND WEAKLY DEPENDENT OBSERVATIONS
BY MODEL SELECTION

MARTIN KROLL

Universität Mannheim

SUMMARY. We consider the non-parametric Poisson regression problem where the integer valued response $Y$ is the realization of a Poisson random variable with parameter $\lambda(X)$. The aim is to estimate the functional parameter $\lambda$ from independent or weakly dependent observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ in a random design framework.

First we determine upper risk bounds for projection estimators on finite dimensional subspaces under mild conditions. In the case of Sobolev ellipsoids the obtained rates of convergence turn out to be optimal.

The main part of the paper is devoted to the construction of adaptive projection estimators of $\lambda$ via model selection. We proceed in two steps: first, we assume that an upper bound for $\|\lambda\|_\infty$ is known. Under this assumption, we construct an adaptive estimator whose dimension parameter is defined as the minimizer of a penalized contrast criterion. Second, we replace the known upper bound on $\|\lambda\|_\infty$ by an appropriate plug-in estimator of $\|\lambda\|_\infty$. The resulting adaptive estimator is shown to attain the minimax optimal rate up to an additional logarithmic factor both in the independent and the weakly dependent setup. Appropriate concentration inequalities for Poisson point processes turn out to be an important ingredient of the proofs.

We illustrate our theoretical findings by a short simulation study and conclude by indicating directions of future research.

1. INTRODUCTION

We consider the non-parametric estimation of a regression function $\lambda: X \rightarrow [0, \infty)$ defined on some Polish space $X$ from observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ where, conditional on $X_1, \ldots, X_n$, the $Y_i$ are independent and Poisson distributed with parameter $\lambda(X_i)$. The covariates $X_1, \ldots, X_n$ are drawn from some strictly stationary process $(X_i)_{i \in \mathbb{Z}}$, and we will consider the two cases where either (i) the $X_1, \ldots, X_n$ are independent, or (ii) some adequate condition on the dependence of the underlying process $(X_i)_{i \in \mathbb{Z}}$ is satisfied. Although we will also provide minimax theoretical results for the non-parametric estimation problem at hand, our focus will be on the adaptive estimation of $\lambda$, that is, the construction of estimators that depend only on the observations but not on any structural presumptions concerning the regression function.

E-mail address: martin.kroll@math.uni-mannheim.de.

Date: May 14, 2018.

2010 Mathematics Subject Classification. 62G05, 62G08.

Key words and phrases. Poisson regression. Nonparametric estimation. Projection estimator. Adaptive estimation. Model selection. Minimax lower bound.

This paper is an extension of results from the author’s PhD thesis [Kro17]. I am grateful to my supervisors Jan Johannes and Martin Schlather for constructive discussions and support. Financial support by the Deutsche Forschungsgemeinschaft (DFG) through the Research Training Group RTG 1953 is gratefully acknowledged.

Present address: École Nationale de la Statistique et de l’Administration Économique (ENSAE), 5 avenue Henry Le Chatelier, F-91120 Palaiseau. martin.kroll@ensae.fr.

arXiv:1708.02405v2 [math.ST] 11 May 2018
Regression models with count data, i.e., non-negative and integer-valued response, are of interest in a wide range of applications, for instance in economics [Win08], quantitative criminology [BM08], and ecology [VHB07]. The Poisson regression model introduced above is the most natural example of such a count data regression model. Other models with count data response include models based on the negative binomial distribution which can also deal with overdispersion. Such more advanced models will not be considered in this paper. Most of the work in the area of count data regression has been devoted to parametric models, see for instance the monograph [CT98] for a comprehensive overview of methods. Let us just mention some examples: the paper [DTM98] gives an application of a Poisson regression model in a geostatistical context. It provides a fully parametric approach and suggests MCMC techniques for fitting a model to the given data. The paper [CP02] introduces a semi-parametric Bayesian model for count data regression and applies it as a prognostic model for early breast cancer data. The article [Nak+05] considers geographically weighted Poisson regression for disease association mapping.

Despite its potential utility in many applications, non-parametric Poisson regression has hardly been studied from a theoretical point so far. One possible approach is to apply the so-called Anscombe transform [Ans48] to the data and treat the data as if they were Gaussian. Another approach would be to consider the generalized linear model representation of Poisson regression and allow for varying coefficients [HT93, FZ99]. Recent work has also considered the Poisson regression model in a high-dimensional framework using the LASSO and the group LASSO [IPR16]. Another interesting reference is [Fry08]: in this paper the author considers a very general model with Poisson regression as a special case. In contrast to our model, only regression with deterministic design is considered (note that the distinction between independent and weakly dependent covariates considered by us is not possible in the model with deterministic design). Moreover, the automatic choice of the smoothing parameter is not addressed from a theoretical point of view in [Fry08] whereas this is the major topic of our contribution. Finally, let us mention the work [BFS04] where an extensive simulation study for count data regression using wavelet methods was performed. That paper contains also further references to Bayesian methods in the context of count data regression. In this paper, we study adaptive non-parametric Poisson regression via the model selection approach. To the best of our knowledge, this approach has not been used for non-parametric Poisson regression so far (in a parametric framework, however, the recent paper [KYS13] considers a model selection approach via a bias-corrected AIC criterion).

Note that a characteristic feature of the non-parametric Poisson regression model is the fact that it naturally incorporates heteroscedastic noise. Besides work on regression in presence of homoscedastic errors (see for instance [Bar00]), there already exists research that considers model selection techniques in regression frameworks containing heteroscedasticity [Sau13]. However, in [Sau13] the observations are of the form

\[
Y = r(X) + \sigma(X)\varepsilon
\]

where \( r \) is the unknown regression function to be estimated, the residuals \( \varepsilon \) have zero mean and variance one, and the function \( \sigma \) models the unknown heteroscedastic noise level. Note that this model does not contain the Poisson regression model to be considered in this paper as a special case.

Our paper is also more general with regard to another aspect: we do not exclusively stick to the case that the covariates \( X_i \) are independent but also consider the more general case that the covariates are weakly dependent which seems to be more realistic at least in some real world scenarios. For instance, when studying clutch sizes of bird eggs that are modeled via count data models (that usually go beyond the Poisson model studied here due to over-dispersion of the data) in ornithology [RB04], the covariates (e.g.,
Nonparametric Adaptive Poisson Regression

2. Methodology

2.1. Notation. Throughout the paper, let \((\mathcal{X}, \mathcal{X}, \mathbb{P})\) be a fixed probability space and denote by \(L^2 = L^2(\mathcal{X}, \mathcal{X}, \mathbb{P})\) the space of square-integrable random variables. The space \(\mathcal{X}\) is assumed to be Polish with \(\mathcal{X}\) being the \(\sigma\)-field generated by the topology of \(\mathcal{X}\). The regression function \(\lambda\) is always assumed to belong to \(L^2\). For \(p = 1, 2\), let \(\|\cdot\|_p\) denote the usual \(L^p\) norm, i.e., \(\|g\|_p = (\int \|g\|^p d\mathbb{P})^{1/p}\) (in the case \(p = 2\) we usually suppress the index \(p\)). In the special case \(p = 2\), we denote the scalar product corresponding to the norm \(\|\cdot\|_2\) by \(\langle \cdot, \cdot \rangle\). \(\|\cdot\|_\infty\) denotes the sup norm on the space \(\mathcal{X}\). We write \(a_n \lesssim b_n\) if \(a_n \leq Cb_n\) holds for all \(n \in \mathbb{N}\) with some constant independent of \(n\).

2.2. Projection estimators. Let \(\mathcal{S}_m\) be a finite-dimensional subspace of \(L^2\). For a subspace \(\mathcal{S}_m \subseteq \mathcal{S}_n\) with orthonormal basis \(\{\varphi_\eta\}_{\eta \in \mathcal{I}_m}\) (\(\mathcal{I}_m\) being an appropriate index set of cardinality \(\mathcal{D}_m\) equal to the dimension of the model) we denote by \(\hat{\lambda}_m\) the projection estimator given through

\[ \hat{\lambda}_m = \sum_{\eta \in \mathcal{I}_m} \hat{\theta}_\eta \varphi_\eta \]

where \(\hat{\theta}_\eta = \frac{1}{n} \sum_{i=1}^n Y_i \varphi_\eta(X_i)\) is an unbiased estimator of the true generalized Fourier coefficient \(\theta_\eta = \int \lambda(x) \varphi_\eta(x) d\mathbb{P}(dx)\). As often in non-parametric statistics, the theoretical...
investigation of the estimator $\hat{\lambda}_m$ will be based on the bias-variance decomposition of the mean integrated squared error
\[
\mathbb{E}\|\hat{\lambda}_m - \lambda\|^2 = \|\lambda - \lambda_m\|^2 + \mathbb{E}\|\hat{\lambda}_m - \lambda_m\|^2 = \|\lambda - \lambda_m\|^2 + \sum_{\eta \in \mathcal{I}_m} \mathbb{E}[(\hat{\theta}_\eta - \theta_\eta)^2]
\]
where $\lambda_m = \sum_{\eta \in \mathcal{I}_m} \theta_\eta \varphi_\eta$ denotes the orthogonal projection of $\lambda$ on the space $\mathcal{S}_m$. For our theoretical treatment we impose the following condition on the models $\mathcal{S}_m = \text{span}\{\varphi_\eta\}_{\eta \in \mathcal{I}_m}$.

**Assumption 2.1.** There exists a positive constant $\Phi$ such that for any $f \in \mathcal{S}_m$ it holds $\|f\|_\infty \leq \Phi \sqrt{D_m}\|f\|$.

**Remark 2.2.** As remarked by [Vie97], Assumption 2.1 is equivalent to the assumption that for any orthonormal basis $\{\varphi_\eta\}_{\eta \in \mathcal{I}_m}$ of $\mathcal{S}_m$ it holds $\|\sum_{\eta \in \mathcal{I}_m} \varphi_\eta^2\|_\infty \leq \Phi^2 D_m$. In our proofs, we will exploit this characterization of Assumption 2.1. The class of models satisfying Assumption 2.1 incorporates, for instance, all bounded bases (such as the trigonometrical basis) as well as piecewise polynomials, splines, and wavelets (see [Vie97], p. 475 for further details).

When studying adaptive estimators we have to impose further conditions on the set of potential models (see Assumption 3.4).

### 2.3. Dependency assumptions.

In this paper, we aim at developing the theory both for independent and weakly dependent observations. In order to describe dependency between subsequent observations of the covariates $X_1, \ldots, X_n$, several concepts of mixing coefficients have been introduced (see [Bos98] for a comprehensive introduction): in this paper we consider the $\beta$-mixing (or absolutely regular-mixing) coefficients that were originally introduced in [KR60]. For a probability space $(\Omega, \mathcal{F}, \mathbb{Q})$ and two sub-$\sigma$-fields $\mathcal{U}$ and $\mathcal{V}$ of $\mathcal{F}$ the $\beta$-mixing coefficient is defined by
\[
\beta(\mathcal{U}, \mathcal{V}) = \frac{1}{2} \sup \left\{ \sum_i \sum_j |\mathbb{Q}(U_i)\mathbb{Q}(V_j) - \mathbb{Q}(U_i \cap V_j)| \right\}
\]
where the supremum is taken over all finite partitions $(U_i)_{i \in I}$ and $(V_j)_{j \in J}$ of $\Omega$ which are measurable with respect to $\mathcal{U}$ and $\mathcal{V}$, respectively. For random variables $X_1, X_2$ we define $\beta(X_1, X_2)$ as the $\beta$-mixing coefficient between the $\sigma$-fields generated by $X_1$ and $X_2$, respectively, i.e., $\beta(X_1, X_2) = \beta(\sigma(X_1), \sigma(X_2))$. For a strictly stationary process $(X_i)_{i \in \mathbb{Z}}$ of random elements in a Borel space, let us denote $\mathcal{F}_0 = \sigma(\{X_i\}_{i \leq 0})$ and $\mathcal{F}_k = \sigma(\{X_i\}_{i \geq k})$ for any positive integer $k$. The sequence of $\beta$-mixing coefficients $(\beta_k)_{k \geq 0}$ of the process $(X_i)_{i \in \mathbb{Z}}$ is defined by $\beta_k = \beta(\mathcal{F}_0, \mathcal{F}_k)$ and the sequence $(X_i)_{i \in \mathbb{Z}}$ is called $\beta$-mixing (or absolutely regular) if $\beta_k \rightarrow 0$ as $k \rightarrow \infty$.

Examples of $\beta$-mixing processes are given in [Lac08] and include, for instance, autoregressive processes of order 1. Concerning the minimax theory, the main difficulty in the weakly dependent case is to find suitable bounds for the variance of the estimated coefficients $\hat{\theta}_\eta$ in the bias-variance decomposition. The main tool to deal with this problem will be Lemma B.1 below.

### 2.4. Adaptive estimation via model selection.

Given a finite collection of models $\mathcal{M}_n$, the model corresponding to the optimal estimator from the set $\{\hat{\lambda}_m\}_{m \in \mathcal{M}_n}$ depends on the unknown $\lambda$ and is thus not accessible. Often it is possible to choose an optimal model when imposing smoothness restrictions on the unknown $\lambda$ (see, for instance, Example 3.2 below) but even this assumption usually seems to be too hard in practise where one wants to construct an estimator of $\lambda$ in a fully-data driven way. In applications, cross-validation techniques [AC10] are quite popular. There are several other methods that aim at the construction of one single estimator from a given set of estimators, among them Lepski’s method [Lep91], aggregation [BTW07, LM09, RT12] or model selection [BBM99]. In this
The principal idea of the model selection approach is to choose a model \( \hat{m} \) from the collection \( \mathcal{M}_n \) by means of a so-called penalized contrast criterion

\[
\hat{m} = \arg\min_{m \in \mathcal{M}_n} \{ \Upsilon(\hat{\lambda}_m) + \text{pen}(m) \}
\]

where \( \Upsilon \) is the contrast function and pen the penalty (in case of non-uniqueness of the minimizer, one chooses an arbitrary one). Usually, one can prove for the adaptive estimator \( \hat{\lambda}_m \) oracle inequalities of the form

\[
\mathbb{E} \| \hat{\lambda}_m - \lambda \|^2 \lesssim \inf_{m \in \mathcal{M}_n} \{ \| \lambda - \lambda_m \|^2 + \text{pen}(m) \} + \text{‘terms of lower order’}.
\]

The general form of the result already shows that if one is able to choose the penalty term of the same order as the variance under the model \( m \), the term over which the infimum is taken mimicks the bias-variance trade-off and one obtains an estimator that attains the optimal rate of convergence. In our case, we will have to introduce an additional logarithmic factor in the penalty leading to a deterioration of the optimal rate by this logarithmic factor in the adaptive case. A crucial tool in order to prove oracle inequalities of the above form are suitable concentration inequalities, and in our Poisson regression setup we will use concentration results for Poisson processes. More precisely, our theoretical analysis is based on a special consequence of a Talagrand type inequality (our Lemma E.2) that has turned out to be fruitful in non-parametric estimation (see for instance [Cha13], Proposition 2.2) but has only been transferred to the Poisson setup in [Kro17].

Concerning the case of dependent covariates, an additional difficulty appears by the fact that concentration inequalities which are crucial for the proof in the independent case are not available in the dependent one. In order to deal with this case, we will exploit a construction due to [Vie97] where the sample \( X_1, \ldots, X_n \) is substituted by another sample \( X^*_1, \ldots, X^*_n \) such that non-neighbouring blocks of a certain size of the \( X^*_i \) are independent. Simultaneously, the \( X^*_i \) are constructed such that they coincide with the original \( X_i \) with high probability. Concentration inequalities will then be applied to the independent blocks instead of to the original \( X_i \). The adaptive estimation will be discussed in detail in Subsections 3.3 (independent case) and 4.2 (weakly dependent case), respectively.

### 3. Independent observations

We first focus on the case of independent observations. In this case we will, besides an upper bound on the risk, also derive a lower bound which shows the optimality of the estimator over certain classes of \( L^2 \)-ellipsoids. Subsection 3.3 is then dedicated to the adaptive estimation by model selection.

#### 3.1. Upper bound

The following proposition states an upper bound for a general model.

**Proposition 3.1.** For a model \( m \) satisfying Assumption 2.1, let \( \hat{\lambda}_m \) be the corresponding projection estimator defined in (1). Then

\[
\mathbb{E} \| \hat{\lambda}_m - \lambda \|^2 \leq \| \lambda_m - \lambda \|^2 + \Phi^2 \frac{\| \lambda \| + \| \lambda \|_1}{n}.
\]

Unfortunately, given a collection of models, a model optimizing the upper bound cannot be specified in advance since the bias term depends on the unknown regression function. The aim of the following example is to illustrate the general result of Proposition 3.1 in the special case when the models under consideration are given by nested spaces generated by the trigonometric basis. As usual in non-parametric statistics, one can determine a rate optimal model by imposing some *a priori* smoothness conditions on the regression function.
Example 3.2 ($L^2$-ellipsoids). In order to illustrate the result of Proposition 3.1, let us consider the special case of the trigonometric basis on the space $(\mathcal{X}, \mathcal{F}, \mathbb{P}) = ([0, 1], \mathcal{B}([0, 1]), dx)$: for $m \in \mathbb{N}_0$, let $\mathcal{S}_m = \{-m, \ldots, m\}$ and $\mathcal{S}_m = \text{span}(\varphi_j)_{j \in \mathcal{S}_m}$ where

$\varphi_0 = 1, \quad \varphi_j(x) = \sqrt{2}\cos(2\pi jx), \quad \text{and} \quad \varphi_{-j}(x) = \sqrt{2}\sin(2\pi jx)$ \quad \text{for} \quad j \in \mathbb{N}.

Smoothness of the regression function $\lambda$ may be expressed by assuming its membership to a suitable ellipsoid

$$\Theta^R_\gamma = \left\{ \lambda = \sum_{j \in \mathbb{Z}} \theta_j \varphi_j \in L^2 : \lambda \geq 0 \text{ and } \sum_{j \in \mathbb{Z}} \theta_j^2 \gamma_j^2 \leq R \right\}$$

where $R > 0$ and $\gamma = (\gamma_j)_{j \in \mathbb{Z}}$ is a strictly positive symmetric sequence such that $\gamma_0 = 1$ and the sequence $(\gamma_n)_{n \in \mathbb{N}_0}$ is non-decreasing. Typical examples of $\gamma$ are $\gamma_j = |j|^p$ (for $j \geq 1$) and $\gamma_j = \exp(p|j|)$ for $p > 0$. Under the stated assumption on the sequence $\gamma$, the bias term in the proof of Proposition 3.1 may be bounded as follows (as introduced above, $\lambda_m$ denotes the projection of $\lambda$ on $\mathcal{S}_m$):

$$\|\lambda_m - \lambda\|^2 = \sum_{|j| > m} \theta_j^2 \leq \gamma_m^{-2} \sum_{|j| > m} \theta_j^2 \gamma_j^2 \leq R \gamma_m^{-2}.$$

The trade-off between squared bias and variance is thus equivalent to the best compromise between $\gamma_m^{-2}$ and $m n^{-1}$. In the polynomial case $\gamma_j = |j|^p$, the best compromise is realized by $m^* \approx n^{1/(2p+1)}$, and we get the classical non-parametric rate $n^{-2p/(2p+1)}$. In the exponential case $\gamma_j = \exp(p|j|)$, we have $m^* \propto \log n$ and the rate is $\log n/n$.

3.2. Lower bound for Sobolev ellipsoids. The following theorem provides a lower bound on the minimax risk in the framework of Example 3.2.

Theorem 3.3. Consider $\Theta^R_\gamma$ defined as in Example 3.2. Let $\gamma = (\gamma_j)_{j \in \mathbb{Z}}$ be a strictly positive symmetric sequence such that $\gamma_0 = 1$ and the sequence $(\gamma_n)_{n \in \mathbb{N}_0}$ is non-decreasing. Set $m^* = \arg\min_{k \in \mathbb{N}_0} \max\{\gamma_k^{-2}, 2k+1\}$ and $\Psi_n = \max\{\gamma_m^{-2}, \frac{2m^*+1}{n}\}$. Assume that

1. $\Gamma := \sum_{j \in \mathbb{Z}} \gamma_j^{-2} < \infty$, and
2. $0 < \eta^{-1} := \inf_{n \in \mathbb{N}} \Psi_n^{-1} \min\{\gamma_m^{-2}, \frac{2m^*+1}{n}\}$ for some $1 \leq \eta < \infty$.

Then, for any $n \in \mathbb{N}$,

$$\inf_{\lambda} \sup_{\lambda \in \Theta^R_\gamma} \mathbb{E}[\|\lambda - \lambda\|^2] \geq \Psi_n,$$

where the infimum is taken over all estimators $\hat{\lambda}$ of $\lambda$ based on the observation of the tuples $(X_1, Y_1), \ldots, (X_n, Y_n)$.

The heuristic behind condition (2) is that for the optimal model $m^*$ the corresponding squared bias $\gamma_{m^*}^{-2}$ and the variance $\frac{2m^*+1}{n}$ should be of the same order. It is satisfied for both the case that $\gamma_j = |j|^p$ and the case that $\gamma_j = \exp(p|j|)$.

The lower bound of Theorem 3.3 shows together with Example 3.2 that under the given assumptions the rate $\Psi_n$ is optimal.

3.3. Adaptive estimation. In order to construct an adaptive estimator of the regression function we stick to the model selection method sketched in the introduction. For this approach we define the contrast function

$$\Upsilon_n(f) = \|f\|^2 - 2\langle \hat{\lambda}_n, f \rangle$$

for $f \in L^2$ and $\hat{\lambda}_n$ is the projection estimator associated to the subspace $\mathcal{S}_n$.

The aim of the model selection approach is to select from a given collection $\mathcal{M}_n$ of submodels of $\mathcal{S}_n$ in a completely data-driven way a candidate that behaves as well as possible as the best model in the collection in the sense of an oracle inequality like (2). In order to establish our theoretical results, we have to introduce a further assumption on the collection of models.
Assumption 3.4. The models \( m \in \mathcal{M}_n \) are nested in the sense that for \( m, m' \in \mathcal{M}_n \) the inequality \( D_m \leq D_{m'} \) implies that \( S_m \subseteq S_{m'} \) (in particular, this implies that there is at most one model with \( D_m = d \) for a given dimension \( d \in \mathbb{N} \)). In addition, there exist universal constants \( c_m, c_M > 0 \) such that

- \( D_m \leq c_m n \) for all \( m \in \mathcal{M}_n \) and \( n \in \mathbb{N} \),
- \( |\mathcal{M}_n| \leq c_M n \) for all \( n \in \mathbb{N} \).

The subspace associated with the maximal model in the collection \( \mathcal{M}_n \) will be denoted with \( S_n \) and the corresponding basis with \( \{ \varphi_{\eta} \}_{\eta \in I_n} \).

The nestedness assumption together with Assumption 2.1 is quite standard (see [BM97], p. 58) and satisfied, e.g., by the trigonometric basis, piecewise polynomials and wavelets (see [BM97], p. 71).

3.3.1. Known upper bound of the regression function. Before we derive a fully adaptive estimator we first stick to the following assumption.

Assumption 3.5. We have access to some \( \xi > 0 \) such that \( \| \lambda \|_\infty \leq \xi \).

Based on the knowledge of \( \xi \), we define for a model \( m \in \mathcal{M}_n \) the penalty:

\[
pen(m) = 24 \mu \cdot \frac{\Phi^2 D_m}{n} + 400 \mu \cdot \Phi^2 D_m \cdot \frac{\log(n+2)}{n}
\]

(4)

where \( \mu = 1 \lor \xi^2 \). Then a model \( \tilde{m} \) is chosen as follows:

\[
\tilde{m} = \arg \min_{m \in \mathcal{M}_n} \{ \Upsilon(\hat{\lambda}_m) + pen(m) \}
\]

(in case that the minimizer is not unique one chooses an arbitrary one).

Remark 3.6. Note that the definition of the penalty in (4) (and in all subsequent definitions within this work) contains an additional logarithmic term which is in contrast to the standard Gaussian regression setup where such a factor is not necessary (see [Mas07], Section 4.3.3). This additional term will cause an extra logarithmic factor in the adaptive upper bounds in contrast to the minimax upper bounds. Unfortunately, it is not clear to us if this factor is indeed inevitable or just an artefact of our method. A heuristic explanation for the necessity of this extra logarithmic factor in our case would be the fact that the concentration inequalities used for our theoretical analysis are not derived in a sub-gaussian setup but in a Poisson setup with sub-gamma tails.

Theorem 3.7. For every \( n \in \mathbb{N} \), let \( \mathcal{M}_n \) be a collection of models such that Assumption 2.1 is satisfied for all \( m \in \mathcal{M}_n \). Further assume that the collection \( \mathcal{M}_n \) satisfies Assumption 3.4 and that Assumption 3.5 holds. Then

\[
\mathbb{E} \| \hat{\lambda}_{\tilde{m}} - \lambda \|^2 \lesssim \min_{m \in \mathcal{M}_n} \max \{ \| \lambda - \lambda_m \|^2, \text{pen}(m) \} + \frac{1}{n}.
\]

Example 3.8 (Continuation of Example 3.2). Setting \( \mathcal{M}_n = \{0, \ldots, n\} \) and defining the spaces \( S_m \) for \( m \in \mathcal{M}_n \) exactly as in Example 3.2, the penalty function reads

\[
\text{pen}(m) = 24 \mu \cdot \frac{2m + 1}{n} + 400 \mu \cdot (2m + 1) \cdot \frac{\log(n+2)}{n}.
\]

Up to the additional logarithmic factor in the second summand, the penalty term behaves exactly as the variance. Hence, the adaptive estimate attains the optimal rate up to this extra logarithmic factor.

Remark 3.9. A careful inspection of the proof of Theorem 3.7 shows that the established bound holds uniformly over sets \( \Theta_\gamma^F \cap \{ \| \lambda \|_\infty \leq \xi \} \) for \( \xi > 0 \). In the framework of Example 3.2 the rate is then deteriorated by an additional logarithmic factor in the case of polynomially increasing \( \gamma \) and optimal in the case of exponentially increasing \( \gamma \).
3.3.2. Unknown upper bound of the regression function. We now propose an adaptive estimator of the regression function \( \lambda \) that does not depend on a priori knowledge of an upper bound for \( \| \lambda \|_\infty \), and is thus fully data-driven. Of course, the key idea is to replace the quantity \( \xi \) from Assumption 3.5 appearing in the definition of the penalty (4) by an appropriate estimator of \( \| \lambda \|_\infty \). For the construction of the estimator of \( \| \lambda \|_\infty \), we take inspiration from an approach that was used in [Com01] in the context of adaptive estimation of the spectral density from a stationary Gaussian sequence. More precisely, the estimator of \( \| \lambda \|_\infty \) is obtained as the plug-in estimator \( \| \hat{\lambda}_\Pi \|_\infty \) where \( \hat{\lambda}_\Pi \) is a suitable histogram estimator of \( \lambda \) based on some partition \( \Pi = \{ X_1, \ldots, X_M \} \) of the space \( X \) in mutually disjoint measurable sets \( \lambda_j \). More precisely, \( \hat{\lambda}_\Pi \) is defined as \( \hat{\lambda}_\Pi = \sum_{j=1}^M \pi_j \mathbf{1}_{X_j} \) where \( \pi_j = \frac{1}{n \sqrt{P(X_j)}} \sum_{i=1}^n Y_i \mathbf{1}_{(X_i \in X_j)} \). Obviously, \( \hat{\lambda}_\Pi \) is the projection estimator on the space \( S_\Pi \) generated by the orthonormal basis functions \( \frac{1}{\sqrt{p(X_j)}} \mathbf{1}_{X_j}, j = 1, \ldots, m \) which has dimension \( D_\Pi = |\Pi| = M \).

We substitute the quantity \( \xi \) in the definition of the penalty term defined in (4) with \( \| \hat{\lambda}_\Pi \|_\infty \). More precise assumptions on the partition \( \Pi \) will be stated in Theorem 3.10 below. Further, by adapting the numerical constants in the definition of the penalty (which is necessary for our proof), we replace the deterministic penalty term used under the validity of Assumption 3.5 by the random penalty

\[
\hat{\text{pen}}(m) = 384 \hat{\mu} \cdot \frac{\Phi^2 D_m}{n} + 6400 \hat{\mu} \cdot \Phi^2 D_m \cdot \frac{\log(n + 2)}{n}
\]

where \( \hat{\mu} = 1 \vee \| \hat{\lambda}_\Pi \|_\infty^2 \). Keeping the contrast function \( Y_n \) as defined in (3) we finally put

\[
\hat{m} = \arg\min_{m \in M_n} \{ Y_n(\hat{\lambda}_m) + \hat{\text{pen}}(m) \}.
\]

**Theorem 3.10.** For every \( n \in \mathbb{N} \), let \( M_n \) be a collection of models such that Assumption 2.1 is satisfied for all \( m \in M_n \). Further assume that the collection of models satisfies Assumption 3.4 and that the following conditions hold:

1. (\( \Pi_1 \)) \( \| \lambda - \lambda_\Pi \|_\infty \leq \frac{1}{2} \| \lambda \|_\infty \) where \( \lambda_\Pi \) denotes the projection of \( \lambda \) on \( S_\Pi \), and
2. (\( \Pi_2 \)) the partition \( \Pi = \{ X_1, \ldots, X_M \} \) in the definition of the auxiliary estimator \( \hat{\lambda}_\Pi \) satisfies \( P(X_j) \geq c_\Pi/M \) for some constant \( c_\Pi > 0 \) and

\[
M \leq \frac{c_\Pi n}{320 \log n}.
\]

Then

\[
\mathbb{E}\| \hat{\lambda}_m - \lambda \|^2 \leq \min_{m \in M_n} \max\{ \| \lambda - \lambda_m \|^2, \text{pen}(m) \} + \frac{1}{n}
\]

where \( \text{pen}(m) = 24 \mu \cdot \frac{\Phi^2 D_m}{n} + 400 \mu \cdot \Phi^2 D_m \cdot \frac{\log(n + 2)}{n} \) and \( \mu = 1 \vee \| \lambda \|_\infty^2 \).

The additional Assumptions (\( \Pi_1 \)) and (\( \Pi_2 \)) are inspired by similar assumptions made in Theorem 2 of [Com01]. Note that Assumption (\( \Pi_2 \)) is especially satisfied in the case that \( X \) is some compact subset of some \( \mathbb{R}^d \) and \( P \) admits a density with respect to the Lebesgue measure that is bounded from below by some strictly positive constant. In this case, one can take an arbitrary partition of \( X \) into \( M \) sets \( X_1, \ldots, X_M \) of equal Lebesgue measure.

4. Dependent observations

4.1. Upper bound. The following proposition provides an upper bound on the risk in the weakly dependent case.
Proposition 4.1. For a model $m$ satisfying Assumption 2.1 let $\hat{\lambda}_m$ be the corresponding projection estimator. Then

$$E[|\hat{\lambda}_m - \lambda|^2 \leq \|\lambda_m - \lambda\|^2 + \frac{\Phi^2 D_m}{n} \left\|\lambda\|_1 + 4\|\lambda\|_\infty \left(\sum_{k=0}^{n} \beta_k\right)\right\|_2.$$ 

Note that the numerical constant is uniformly bounded in $n$ under the assumption that $\sum_{k=0}^{\infty} \hat{\beta}_k < \infty$.

Remark 4.2. Under the assumption that $\sum_{k=0}^{\infty} \beta_k < \infty$ the bound in the weakly dependent case coincides (apart from the numerical constants appearing) with the one in the independent case. Moreover, since the weakly dependent case incorporates the independent one, the lower bound given in Theorem 3.3 provides also the benchmark for the weakly dependent case.

4.2. Adaptive estimation. We now consider adaptive estimation in case of weak dependence. We stick directly to the case that Assumption 3.5 is not satisfied. We keep the contrast function from Section 3 but define the penalty as

$$\hat{\text{pen}}(m) = D_m \cdot \frac{\log(n+2)}{n} + 6400 \cdot \Phi^2 D_m \cdot \frac{\log(n+2)}{n}.$$ 

In addition we have to impose the following assumption on the $\beta$-mixing coefficients which is similar to Assumption A4 in [Lac08].

Assumption 4.3. The process $(X_i)_{i \in \mathbb{Z}}$ is geometrically $\beta$-mixing ($\beta_q \lesssim e^{-\theta q}$) or arithmetically $\beta$-mixing ($\beta_q \lesssim M q^{-\theta}$) with $\theta \geq 9$ in the latter case.

Note that under Assumption 4.3 the condition $\sum_{k=0}^{\infty} \beta_k < \infty$ is satisfied. Examples of processes satisfying Assumption 4.3 are given in [Lac08] and include autoregressive processes of order 1. We have the following theorem.

Theorem 4.4. Let the assumptions of Theorem 3.10 be satisfied with the condition on $M$ replaced with

$$M \leq \frac{c \Pi n^{1/3}}{320 \log n}.$$ 

Further assume that Assumption 4.3 holds. Then, for every $n \in \mathbb{N}$,

$$E[\|\hat{\lambda}_m - \lambda\|^2 \lesssim \inf_{m \in M_n} \max\{\|\lambda - \lambda_m\|^2, \text{pen}(m)\} + \frac{1}{n},$$

where $\text{pen}(m) = D_m \cdot \frac{\log(n+2)}{n}.$

5. Numerical results

Although the main focus of the present paper is the derivation of theoretical results we provide a short simulation study which is particularly intended to motivate and stimulate further research on the topic of the paper. More precisely, we test our approach for the test function $\lambda : [0, 1] \to (0, \infty)$ where

$$\lambda(x) = (5 + 5 \cos(2\pi x)) \cdot \mathbb{I}_{[0,0.5]}(x) + 10x \cdot \mathbb{I}_{(0.5,1]}(x).$$

Concerning the covariates $X_i$, we assume $X_i \sim U([0, 1])$ where $U([0, 1])$ denotes the uniform distribution on the interval $[0, 1]$. In a first experiment, we assume the covariates to be independent, in a second one we introduce dependencies by the means of the following model: $X_1 \sim U([0, 1])$, and then in order to generate $X_2, \ldots, X_n$ one recursively sets for $i = 2, \ldots, n$

$$X_i = [0.5X_i-1 + \varepsilon_i]$$

where the $\varepsilon_i$ are i.i.d. $\sim \mathcal{N}(0, 1)$. 

Concerning the models, we consider approximating spaces in terms of piecewise constant functions (histograms). More precisely, we put $M_n = \{0, \ldots, \lfloor \log_2 n \rfloor \}$ and for $m \in M_n$ define the space $S_m$ as the linear span of the functions $\varphi_j(x)$, $j = 1, \ldots, 2^m$ where

$$
\varphi_j(x) = \sqrt{2^m} 1_{\left[\frac{j-1}{2^m}, \frac{j}{2^m}\right)}(x).
$$

Note that the assumptions on the models made by us are satisfied: Assumption 2.1 holds true with $\Phi = 1$, and Assumption 3.4 is satisfied with $c_m = c_M = 1$. For the sake of convenience, we assume that we a priori know that $\|\lambda\|_{\infty} \leq 10$ and consider a penalty of the form $\text{pen}(m) = \kappa \cdot D_m \cdot \xi^2 \cdot \log_n n$ with $\xi^2 = 100$ and $\kappa$ a numerical constant. Concerning the latter, we test various choices. It is a phenomenon often recognized in non-parametric model selection that the numerical constant in the definition of the penalty term which is convenient to derive theoretical results is by much too large to obtain reasonable results for samples of small size. This phenomenon is also encountered in our simulation study. In contrast to this unpleasant behaviour, the overall method is not very demanding from a computational point of view because we essentially minimize the penalized contrast criterion over the set of admissible models which is at most of order $n$ by Assumption 3.4.

The results of our simulation experiments are summarized in Tables 1 and 2 whereas illustrations are given in Figures 1 and 2.

| $n$  | $\kappa = 0.08$ error | $\kappa = 0.09$ error | $\kappa = 0.10$ error |
|-----|-----------------------|-----------------------|-----------------------|
|     | sd                    | sd                    | sd                    |
| 1024| 1.2171 0.8069         | 1.1901 0.7209         | 1.2496 0.6790         |
| 2048| 0.7326 0.3354         | 0.7097 0.2510         | 0.7428 0.2786         |

Table 1. Performance of the histogram estimator in the case of independent covariates for different values of $\kappa$. Mean error and standard deviation were computed over 500 independent replicates of the experiment. Minimal errors were obtained for $\kappa = 0.09$ for both sample sizes considered.

Figure 1. Illustration of the simulation experiment in the case of independent covariates. The true function is plotted in green, whereas the median of the estimator over $n = 100$ experiments is in blue. The dashed red lines indicate the empirical pointwise 0.25 resp. 0.75 quantile. The dotted red lines indicate the empirical pointwise 0.99 resp. 0.01 quantile. The grey lines show some exemplary outcomes of single experiments.

6. Conclusion and outlook to future research

In this paper, we have considered adaptive non-parametric Poisson regression via model selection in the case of independent and $\beta$-mixing covariates. More precisely, the main
Table 2. Performance of the histogram estimator in the case of dependent covariates for different values of $\kappa$ and different sample sizes $n$. Mean error and standard deviation were computed over 500 independent replicates of the experiment. Minimal errors were obtained for $\kappa = 0.09$ for both sample sizes considered.

| $n$  | $\kappa = 0.08$ error | sd     | $\kappa = 0.09$ error | sd     | $\kappa = 0.10$ error | sd     |
|------|-----------------------|--------|------------------------|--------|------------------------|--------|
| 1024 | 1.2785                | 1.1981 | 1.1505                 | 0.6012 | 1.2588                 | 0.6931 |
| 2048 | 0.7112                | 0.2899 | 0.6967                 | 0.2370 | 0.7375                 | 0.2733 |

Figure 2. Illustration of the simulation experiment in the case of dependent covariates. The linetypes and colours are chosen as in Figure 1.

The most important open question concerns either the development of an adaptive estimation technique that is capable to obtain the optimal rate of convergence without the additional logarithmic factor or the derivation of a theoretical result showing that this is not possible. This aspect will be considered in future research as well as the investigation of kernel type estimators instead of projection estimators.
Appendix A. Proofs of Section 3

A.1. Proof of Proposition 3.1. We have the bias-variance decomposition

$$\mathbb{E}\|\bar{\lambda}_m - \lambda\|^2 = \|\lambda_m - \lambda\|^2 + \mathbb{E}\|\tilde{\lambda}_m - \lambda_m\|^2.$$  

Exploiting Assumption 2.1 and the independence assumption on the $X_i$ yields for the variance term the estimate

$$\mathbb{E}\|\tilde{\lambda}_m - \lambda_m\|^2 = \sum_{\eta \in \mathcal{I}_m} \text{Var}(\hat{\theta}_\eta) = \frac{1}{n} \sum_{\eta \in \mathcal{I}_m} \text{Var}(Y_1 \varphi_\eta(X_1))$$

$$\leq \frac{1}{n} \mathbb{E} \left[ \sum_{\eta \in \mathcal{I}_m} Y_1^2 \varphi_\eta^2(X_1) \right] \leq \frac{\Phi^2 \mathbb{D}_m}{n} \cdot \mathbb{E} Y_1^2$$

$$\leq \frac{\Phi^2 \mathbb{D}_m}{n} \cdot (\|\lambda\|^2 + \|\lambda\|_1),$$

and hence the result follows.

A.2. Proof of Theorem 3.3. For each $\tau = (\tau_j)_{0 \leq j \leq m^*} \in \{\pm 1\}^{2m^*+1}$ we define the function $\lambda_\tau$ as

$$\lambda_\tau = \left( \frac{R}{4} \right)^{1/2} + \tau_0 \left( \frac{R \zeta}{16n} \right)^{1/2} + \left( \frac{R \zeta}{16n} \right)^{1/2} \sum_{0 \leq j \leq m^*} \tau_j \varphi_j$$

$$= \left( \frac{R}{4} \right)^{1/2} + \left( \frac{R \zeta}{16n} \right)^{1/2} \sum_{0 \leq j \leq m^*} \tau_j \varphi_j$$

where $\zeta = \min\{1/(\Gamma \eta), 2/R\}$ and the $\varphi_j$ are defined as in Example 3.2. We have

$$\left\| \left( \frac{R \zeta}{16n} \right)^{1/2} \sum_{0 \leq j \leq m^*} \tau_j \varphi_j \right\|_\infty \leq \left( \frac{R \zeta}{16n} \right)^{1/2} \sum_{0 \leq j \leq m^*} \sqrt{2}$$

$$\leq \left( \frac{R \zeta}{8} \right)^{1/2} \left( \sum_{0 \leq j \leq m^*} \gamma_j^2 \right)^{1/2} \left( \sum_{0 \leq j \leq m^*} \gamma_j^2 n \right)^{1/2}$$

$$\leq \left( \frac{\Gamma R \zeta}{8} \right)^{1/2} \left( \gamma_{2m^*}^2 \frac{2m^*+1}{n} \right)^{1/2} \leq \left( \frac{\Gamma R \zeta}{n} \right)^{1/2} \leq \left( \frac{R}{8} \right)^{1/2},$$

and hence $\lambda_\tau \geq \sqrt{R} \cdot (1/2 - 1/(2\sqrt{2}))$ (in particular, $\lambda_\tau$ is non-negative). Together with the calculation

$$\left[ \left( \frac{R}{4} \right)^{1/2} + \tau_0 \left( \frac{R \zeta}{16n} \right)^{1/2} \right]^2 + \left( \frac{R \zeta}{16n} \right) \sum_{1 \leq j \leq m^*} \gamma_j^2 n \leq \frac{R}{2} + \frac{R \zeta}{16n} \gamma_{2m^*} \frac{2m^*+1}{n} \leq R$$

this shows that $\lambda_\tau \in \Theta^R_\gamma$ for every $\tau \in \{\pm 1\}^{2m^*+1}$.

We now derive a reduction scheme which holds for an arbitrary estimator $\hat{\lambda}$ of $\lambda$. For this purpose, denote $X = (X_1, \ldots, X_n)$, $Y = (Y_1, \ldots, Y_n)$, and by $\mathbb{E} Y, \mathbb{E} X$ the conditional distribution of $Y$ given $X$ when the true regression function is $\lambda_\tau$. By $\mathbb{E}_\tau[\cdot|X]$ we denote the corresponding conditional expectation operator. Then the following reduction scheme holds

$$\sup_{\lambda \in \Theta^R_\gamma} \mathbb{E}\|\tilde{\lambda} - \lambda\|^2 \geq \frac{1}{2^{2m^*+1}} \sum_{\tau \in \{\pm 1\}^{2m^*+1}} \sum \mathbb{E} \mathbb{E}_\tau[|\tilde{\theta}_j - \theta_{\tau j}|^2|X]$$

$$= \frac{1}{2^{2m^*+1}} \sum_{0 \leq j \leq m^*} \sum_{\tau \in \{\pm 1\}^{2m^*+1}} \frac{1}{2} \left\{ \mathbb{E} \mathbb{E}_\tau[|\tilde{\theta}_j - \theta_{\tau j}|^2|X] + \mathbb{E} \mathbb{E}_\tau[|\tilde{\theta}_j - \theta_{\tau j}|^2|X] \right\}$$

(5)
where \( \tilde{\theta}_j \), \( \theta_{\tau j} \) are the coefficients of \( \tilde{\lambda} \) and \( \lambda_\tau \) corresponding to the basis function \( \varphi_j \), respectively, and for \( \tau \in \{ \pm 1 \}^{2m^*+1} \) the element \( \tau_j \in \{ \pm 1 \}^{2m^*+1} \) is defined by \( \tau_k = \tau_k \) for \( k \neq j \) and \( \tau_j = -\tau_j \). Consider the Hellinger affinity defined through

\[
\rho(P^Y \mid X, P^Y \mid X_{\tau j}) = \int \sqrt{dP^Y \mid X \cdot dP^Y \mid X_{\tau j}}.
\]

We have

\[
\rho(P^Y \mid X, P^Y \mid X_{\tau j}) \leq \int \frac{\tilde{\theta}_j - \theta_{\tau j}}{|\tilde{\theta}_j - \theta_{\tau j}|} \sqrt{dP^Y \mid X \cdot dP^Y \mid X_{\tau j}} + \int \frac{\tilde{\theta}_j - \theta_{\tau j}}{|\tilde{\theta}_j - \theta_{\tau j}|} \sqrt{dP^Y \mid X \cdot dP^Y \mid X_{\tau j}}
\]

\[
\leq \left( \int \frac{\tilde{\theta}_j - \theta_{\tau j}}{|\tilde{\theta}_j - \theta_{\tau j}|}^2 \sqrt{dP^Y \mid X} \right)^{1/2} + \left( \int \frac{\tilde{\theta}_j - \theta_{\tau j}}{|\tilde{\theta}_j - \theta_{\tau j}|}^2 \sqrt{dP^Y \mid X_{\tau j}} \right)^{1/2}.
\]

By means of the estimate \((a + b)^2 \leq 2a^2 + 2b^2\) we obtain

\[
\frac{1}{2} |\theta_{\tau j} - \theta_{\tau j}|^2 (P^Y \mid X_{\tau j}) \leq \mathbb{E} \mathbb{E} \mathbb{E} \mathbb{E} \theta_j - \theta_{\tau j}|^2 \mathbb{E} \theta_j - \theta_{\tau j}|^2 \mathbb{X}.
\]

For any \( \tau \in \{ \pm 1 \}^{2m^*+1} \), let us denote by \( \mathbb{P}^{Y \mid X}_{\tau} \) the marginal distribution of \( Y_i \) given \( X \) (since the distribution of \( Y_i \) given \( X \) depends on \( X_i \) only, we could equally write \( \mathbb{P}^{Y \mid X}_{\tau(k)} \)). Formula (5) from [Roo03] allows us to bound the total variation distance \( TV(P^Y \mid X, \mathbb{P}^{Y \mid X}_{\tau j}) \) between \( \mathbb{P}^{Y \mid X}_{\tau} \) and \( \mathbb{P}^{Y \mid X}_{\tau j} \) as

\[
TV(P^Y \mid X, \mathbb{P}^{Y \mid X}_{\tau j}) \leq |\lambda_\tau(X_i) - \lambda_{\tau j}(X_i)| \leq |2 \sqrt{2} (R \zeta/(16n))^{1/2}| \leq \left( \frac{R \zeta}{2n} \right)^{1/2} \leq \frac{1}{\sqrt{n}}
\]

due to the definition of \( \zeta \). Recall the definition \( H^2(P, Q) = \int [\sqrt{dP} - \sqrt{dQ}]^2 \) of the squared Hellinger distance between two probability measures \( P \) and \( Q \). By formula (2.20) from [Tsy09] it follows that

\[
H^2(P^Y \mid X, \mathbb{P}^{Y \mid X}_{\tau j}) \leq \frac{1}{n}.
\]

Since \( Y_1, \ldots, Y_n \) are independent conditionally on \( X_1, \ldots, X_n \) we obtain by Lemma 3.3.10 (i) from [Rei93] that

\[
H^2(P^Y \mid X, \mathbb{P}^{Y \mid X}_{\tau j}) \leq \sum_{i=1}^n H^2(P^Y \mid X, \mathbb{P}^{Y \mid X}_{\tau j}) \leq 1.
\]

Hence the relation \( \rho(P^Y \mid X, \mathbb{P}^{Y \mid X}_{\tau j}) = 1 - H^2(P^Y \mid X, \mathbb{P}^{Y \mid X}_{\tau j})/2 \) (see [Tsy09], p. 87) implies \( \rho(P^Y \mid X, \mathbb{P}^{Y \mid X}_{\tau j}) \geq 1/2 \). Putting this estimate into the reduction scheme (5) finally yields using (6) that

\[
\sup_{\lambda \in \Theta^R_\gamma} \mathbb{E} ||\tilde{\lambda} - \lambda||^2 \geq \frac{1}{2} \sum_{\tau \in \{ \pm 1 \}^{2m^*+1}} \sum_{0 \leq |j| \leq m^*} \frac{1}{2} \mathbb{E} \mathbb{E} \mathbb{E} \mathbb{E} \theta_j - \theta_{\tau j}|^2 \mathbb{E} \theta_j - \theta_{\tau j}|^2 \mathbb{X} \geq \frac{1}{16} \sum_{0 \leq |j| \leq m^*} |\theta_j - \theta_{\tau j}|^2 = \frac{R \zeta}{64} \sum_{0 \leq |j| \leq m^*} \frac{1}{n} = \frac{R \zeta}{64} \frac{2m^* + 1}{n}.
\]

Since the last estimate holds for arbitrary \( \tilde{\lambda} \), we obtain the claim assertion by means of Assumption 2.
A.3. Proof of Theorem 3.7. Note that the identity \( \Upsilon_n(f) = \| \hat{\lambda}_n - f \|^2 - \| \hat{\lambda}_n \|^2 \) holds for all \( f \in L^2 \). Hence \( \hat{\lambda}_n = \argmin_{f \in S_m} \Upsilon_n(f) \) for all \( m \in M_n \), and exploiting the definition of \( \hat{m} \) yields for all \( m \in M_n \) that

\[
\Upsilon_n(\hat{\lambda}_{\hat{m}}) + \text{pen}(\hat{m}) \leq \Upsilon_n(\hat{\lambda}_m) + \text{pen}(m) \leq \Upsilon_n(\lambda_m) + \text{pen}(m)
\]

where \( \lambda_m = \sum_{\eta \in I_m} \theta_\eta \varphi_\eta \) is the projection of \( \lambda \) on the finite-dimensional space \( S_m \). In a similar manner, let us denote by \( \hat{\lambda}_n \) the projection of \( \lambda \) on the space \( S_n \), i.e., \( \lambda_n = \sum_{\eta \in I_n} \theta_\eta \varphi_\eta \). By definition of the contrast and some algebra we obtain

\[
\| \hat{\lambda}_{\hat{m}} - \lambda \|^2 \leq \| \lambda_m - \lambda \|^2 + 2(\hat{\lambda}_m - \lambda_m, \hat{\lambda}_{\hat{m}} - \lambda_m) + \text{pen}(m) - \text{pen}(\hat{m}).
\]

Setting \( \tilde{\theta}_\eta = \frac{1}{n} \sum_{i=1}^n \lambda(X_i) \varphi_\eta(X_i) \) and \( \tilde{\lambda}_m = \sum_{\eta \in I_m} \tilde{\theta}_\eta \varphi_\eta \) we obtain

\[
\| \tilde{\lambda}_m - \lambda \|^2 \leq \| \lambda_m - \lambda \|^2 + 2(\thetahat_n, \tilde{\lambda}_m - \lambda_m) + 2(\thetahat_n, \hat{\lambda}_m - \lambda_m) + \text{pen}(m) - \text{pen}(\hat{m})
\]

where \( \tilde{\theta}_\eta = \tilde{\lambda}_n - \lambda \) and \( \tilde{\lambda}_m = \tilde{\lambda}_n - \lambda \). Set \( B_m = \{ \lambda \in S_m : \| \lambda \| \leq 1 \} \). Using the estimate \( 2xy \leq x^2 + y^2 \) for positive \( \tau \) (below we specialize with \( \tau = 1/8 \)) we conclude

\[
\| \hat{\lambda}_m - \lambda \|^2 \leq 3\| \lambda - \lambda_m \|^2 + 16 \sup_{t \in B_{m \setminus \hat{m}}} |(\hat{\theta}_n, t)|^2 + 16 \sup_{t \in B_{m \setminus \hat{m}}} |(\tilde{\theta}_n, t)|^2 + 2\text{pen}(m) - 2\text{pen}(\hat{m})
\]

\[
\leq 3\| \lambda - \lambda_m \|^2 + 16 \left( \sup_{t \in B_{m \setminus \hat{m}}} |(\hat{\theta}_n, t)|^2 - 50\mu : \frac{\Phi^2 D_{m \setminus \hat{m}} \log(n + 2)}{n} \right) + 16 \left( \sup_{t \in B_{m \setminus \hat{m}}} |(\tilde{\theta}_n, t)|^2 - 3\mu : \frac{\Phi^2 D_{m \setminus \hat{m}} \log(n + 2)}{n} \right) + 800\mu : \frac{\Phi^2 D_{m \setminus \hat{m}} \log(n + 2)}{n} + 48\mu : \frac{\Phi^2 D_{m \setminus \hat{m}}}{n} + 2\text{pen}(m) - 2\text{pen}(\hat{m}).
\]

By definition of the penalty, the estimate \( D_{m \setminus \hat{m}} \leq D_m + D_{\hat{m}} \) and roughly bounding the brackets \( (\ldots)_+ \) by summing over all potential models \( m \in M_n \), we have

\[
\| \hat{\lambda}_{\hat{m}} - \lambda \|^2 \leq 3\| \lambda - \lambda_m \|^2 + 16 \sum_{m' \in M_n} \left( \sup_{t \in B_{m'}} |(\hat{\theta}_n, t)|^2 - 50\mu : \frac{\Phi^2 D_{m'} \log(n + 2)}{n} \right) + 16 \sum_{m' \in M_n} \left( \sup_{t \in B_{m'}} |(\tilde{\theta}_n, t)|^2 - 3\mu : \frac{\Phi^2 D_{m'}}{n} \right) + 4\text{pen}(m).
\]

Taking expectations and into account that the last estimate holds for each \( m \in M_n \), we obtain

\[
\mathbb{E}\| \hat{\lambda}_{\hat{m}} - \lambda \|^2 \leq \min_{m \in M_n} \{ 3\| \lambda - \lambda_m \|^2 + 4\text{pen}(m) \}
\]

\[
+ 16 \sum_{m' \in M_n} \mathbb{E} \left( \sup_{t \in B_{m'}} |(\hat{\theta}_n, t)|^2 - 50\mu : \frac{\Phi^2 D_{m'} \log(n + 2)}{n} \right) + 16 \sum_{m' \in M_n} \mathbb{E} \left( \sup_{t \in B_{m'}} |(\tilde{\theta}_n, t)|^2 - 3\mu : \frac{\Phi^2 D_{m'}}{n} \right).
\]
all the appearing quantities are defined as in the proof of Theorem 3.7.

Bounding the terms in the brackets (7), we obtain

\[ E\|\hat{\lambda}_m - \lambda\|^2 \lesssim \min_{m \in \mathcal{M}_n} \max\{\|\lambda_m - \lambda\|^2, \text{pen}(m)\} + \frac{1}{n}. \]

A.4. **Proof of Theorem 3.10.** Let us introduce the event \( \Xi = \left\{ \frac{\|\hat{\lambda}_m\|_{\infty} \vee 1}{\|\lambda\|_{\infty} \vee 1} - 1 \right\} \). It is readily verified that on \( \Xi \) it holds that

\[ \|\lambda\|_{\infty} \vee 1 \leq 4(|\hat{\lambda}_m|_{\infty} \vee 1) \quad \text{and} \quad \|\hat{\lambda}_m\|_{\infty} \vee 1 \leq \frac{7}{4}(\|\lambda\|_{\infty} \vee 1). \]

These estimates will be used below without further reference. We consider the decomposition

\[ E\|\hat{\lambda}_m - \lambda\|^2 = E\|\hat{\lambda}_m - \lambda\|^2_{\Xi} + E\|\hat{\lambda}_m - \lambda\|^2_{\Xi^c} =: T_1 + T_2. \]

*Upper bound for \( T_1 \):* In analogy to the proof of Theorem 3.7 one can derive (in the following, all the appearing quantities are defined as in the proof of Theorem 3.7)

\[ \|\hat{\lambda}_m - \lambda\|^2 \leq 3\|\lambda_m - \lambda\|^2 + 16 \sup_{t \in B_m} |\langle \hat{\Theta}_n, t \rangle|^2 + 16 \sup_{t \in B_m} |\langle \hat{\Theta}_n, t \rangle|^2 + 2\text{pen}(m) - 2\text{pen}(\hat{m}) \]

\[ \leq 3\|\lambda_m - \lambda\|^2 + 16 \left( \sup_{t \in B_m} |\langle \hat{\Theta}_n, t \rangle|^2 - 50\mu \cdot \frac{\Phi^2_{D_m \vee \hat{m}} \log(n + 2)}{n} \right) + \]

\[ + 16 \left( \sup_{t \in B_m} |\langle \hat{\Theta}_n, t \rangle|^2 - 3\mu \cdot \frac{\Phi^2_{D_m \vee \hat{m}}}{n} \right) + \]

\[ + 800\mu \cdot \frac{\Phi^2_{D_m \vee \hat{m}} \log(n + 2)}{n} + 48\mu \cdot \frac{\Phi^2_{D_m \vee \hat{m}}}{n} + 2\text{pen}(m) - 2\text{pen}(\hat{m}). \]

By definition of \( \Xi \) and the random penalty function, we obtain (note that \( D_m \vee \hat{m} \leq D_m + D_{\hat{m}} \))

\[ \|\hat{\lambda}_m - \lambda\|^2_{\Xi} \leq 3\|\lambda_m - \lambda\|^2 + 16 \left( \sup_{t \in B_m} |\langle \hat{\Theta}_n, t \rangle|^2 - 50\mu \cdot \frac{\Phi^2_{D_m \vee \hat{m}} \log(n + 2)}{n} \right) + \]

\[ + 16 \left( \sup_{t \in B_m} |\langle \hat{\Theta}_n, t \rangle|^2 - 3\mu \cdot \frac{\Phi^2_{D_m \vee \hat{m}}}{n} \right) + 100\text{pen}(m). \]

Bounding the terms in the brackets \((\ldots)_+\) by summing over all admissible models \( m \in \mathcal{M}_n \) and taking expectations on both sides yields

\[ E\|\hat{\lambda}_m - \lambda\|^2_{\Xi} \leq \sum_{m \in \mathcal{M}_n} \left[ \left( \sup_{t \in B_m} |\langle \hat{\Theta}_n, t \rangle|^2 - 50\mu \cdot \frac{\Phi^2_{D_m \vee \hat{m}} \log(n + 2)}{n} \right)_+ \right] + \]

\[ + \sum_{m \in \mathcal{M}_n} \left[ \left( \sup_{t \in B_m} |\langle \hat{\Theta}_n, t \rangle|^2 - 3\mu \cdot \frac{\Phi^2_{D_m \vee \hat{m}}}{n} \right)_+ \right] \]
Applying Lemmata C.1 and C.2 as in the proof of Theorem 3.7 then implies
\[ \mathbb{E}\|\hat{\lambda}_n - \lambda\|^2 \mathbb{I}_{\|\cdot\|} \lesssim \inf_{m \in M_n} \max\{\|\lambda_m - \lambda\|^2, \text{pen}(m)\} + \frac{1}{n}. \]

**Upper bound for \( T_2 \):** First, take note of the estimate
\[ \mathbb{E}\|\hat{\lambda}_n - \lambda\|^2 \mathbb{I}_{\|\cdot\|} \leq 2\mathbb{E}\|\lambda_n\|^2 \mathbb{I}_{\|\cdot\|} + 2\mathbb{E}\|\lambda\|^2 \mathbb{I}_{\|\cdot\|}. \]
We have
\[ \|\hat{\lambda}_n\|^2 = \sum_{\eta \in \mathcal{F}_n} \hat{\rho}_\eta^2 \leq \frac{1}{n} \sum_{i=1}^{n} Y_i^2 \sum_{\eta \in \mathcal{F}_n} \varphi_\eta^2(X_i) \leq \frac{\Phi^2 \mathcal{D}_m}{n} \sum_{i=1}^{n} Y_i^2. \]
Hence, by means of the Cauchy-Schwarz inequality and the fact that \( \mathcal{D}_m \leq c_m n \) due to Assumption 3.4,
\[ \mathbb{E}\|\hat{\lambda}_n - \lambda\|^2 \mathbb{I}_{\|\cdot\|} \leq 2\Phi^2 c_m \left( \mathbb{E}\left[ \left( \sum_{i=1}^{n} Y_i^2 \right)^2 \right] \right)^{1/2} \mathbb{P}(\Xi^6)^{1/2} + 2\|\lambda\|^2 \mathbb{P}(\Xi^6) \leq 2\Phi^2 c_m n T_4(\|\lambda\|)^{1/2} \mathbb{P}(\Xi^6)^{1/2} + 2\|\lambda\|^2 \mathbb{P}(\Xi^6) \]
where \( T_4 \) is the fourth Touchard polynomial \( T_4 \). By the above estimates it suffices to show that \( \mathbb{P}(\Xi^6) \lesssim n^{-d} \). Note that we have
\[ \|\hat{\lambda}_n - \lambda\|_\infty = \sup_{1 \leq j \leq M} \|\hat{\lambda}_n - \lambda\|_{X_j} \]
where \( \nu(\varphi_j) = (\hat{\lambda}_n - \mathbb{E}[\hat{\lambda}_n|X], \varphi_j) \) and \( \tilde{\nu}(\varphi_j) = (\mathbb{E}[\hat{\lambda}_n|X] - \lambda\|, \varphi_j) \). Using (8) and the estimate \( |a \vee 1 - b \vee 1| \leq |a - b| \), we obtain
\[ \mathbb{P}(\Xi^6) = \mathbb{P}(\|\hat{\lambda}_n\|_\infty \vee 1 - \|\lambda\|_\infty \vee 1) \geq 3/4 \cdot (\|\lambda\|_\infty \vee 1) \]
where \( \xi = (\|\lambda\|_\infty \vee 1)/4 \). Note that \( \|\varphi_j\| = 1 \) and \( \|\varphi_j\|_\infty = \mathbb{P}(X_j)^{-1/2} \). By application of Proposition E.3 we obtain putting \( p = \inf_{j=1,...,M} \mathbb{P}(X_j) \)
\[ \mathbb{P}(\pm\nu(\varphi_j) \geq \xi) \leq \exp \left( -\frac{n\xi^2}{2\|\varphi_j\|^2_\infty \|\lambda\|_\infty + 2/3\|\varphi_j\|_\infty} \right) \]
For the variance term, we obtain exploiting Assumption 2.1 and Lemma B.1 with use the bias-variance decomposition $b$ which finishes the proof by defining $\lambda$.

Thus thanks to Lemma 4.1 in [Vie97], there exist two functions $\beta$ such that for any $g \in L^2$ and any $n \in \mathbb{N}$

$$\text{Var} \left( \sum_{i=1}^{n} g(X_i) \right) \leq 4n \mathbb{E} \left[ \left( \sum_{k=0}^{n} b_k \right) ^2 g^2(X_0) \right].$$

Proof of the Lemma. Thanks to Lemma 4.1 in [Vie97], there exist two functions $b_k'$ and $b_k''$ such that $b_k = b_k' - b_k''$ and

$$\text{Cov}(g(X_0), g(X_k)) \leq 2 \mathbb{E} |b_k' g_1 |^{1/2} \mathbb{E} |b_k'' g_2 |^{1/2}.$$

Thus

$$\text{Var} \left( \sum_{i=1}^{n} g(X_i) \right) \leq 2 \sum_{k=0}^{n} (n - k) \text{Cov}(g(X_0), g(X_k)) |$$

$$\leq 4n \sum_{k=0}^{n} \mathbb{E} |b_k' g_1 |^{1/2} \mathbb{E} |b_k'' g_2 |^{1/2}$$

$$\leq 4n \sum_{k=0}^{n} \mathbb{E} (b_k' + b_k'') g^2 / 2$$

which finishes the proof by defining $b_k = (b_k' + b_k'') / 2$. \hfill \Box

We come now to the proof of Proposition 4.1. As in the proof of Proposition 3.1, we use the bias-variance decomposition

$$\mathbb{E} \Vert \hat{\lambda}_m - \lambda \Vert^2 = \Vert \lambda_m - \lambda \Vert^2 + \mathbb{E} \Vert \hat{\lambda}_m - \lambda_m \Vert^2.$$

For the variance term, we obtain exploiting Assumption 2.1 and Lemma B.1

$$\mathbb{E} \Vert \hat{\lambda}_m - \lambda_m \Vert^2 = \sum_{\eta \in I_m} \text{Var}(\theta_{\eta}) = \frac{1}{n^2} \sum_{\eta \in I_m} \sum_{i=1}^{n} \text{Var} \left( \sum_{i=1}^{n} Y_i \varphi_\eta(X_i) \right)$$

$$= \frac{1}{n^2} \sum_{\eta \in I_m} \text{Var} \left( \sum_{i=1}^{n} \mathbb{E}[Y_i \varphi_\eta(X_i) | X_i] \right) + \frac{1}{n^2} \sum_{\eta \in I_m} \mathbb{E} \left[ \text{Var} \left( \sum_{i=1}^{n} Y_i \varphi_\eta(X_i) | X_i \right) \right]$$

$$= \frac{1}{n^2} \sum_{\eta \in I_m} \text{Var} \left( \sum_{i=1}^{n} \lambda(X_i) \varphi_\eta(X_i) \right) + \frac{1}{n^2} \sum_{\eta \in I_m} \mathbb{E} \left[ \sum_{i=1}^{n} \varphi_\eta^2(X_i) \lambda(X_i) \right]$$
1, and consider the event $\Xi^*$. The term $A^*$ and $A^*_\ell$ are independent if $\ell \neq \ell'$, and $P(A_\ell \neq A^*_\ell) \leq \beta_{q_n}$.

In the same fashion, one can build $B^*_\ell$ for $\ell = 0, \ldots, p_n - 1$. We now define the sequence $X^*_i$ via $A^*_i = (X^*_{2q_n+1}, \ldots, X^*_{(2\ell+1)q_n})$ and $B^*_i = (X^*_{(2\ell+1)q_n+1}, \ldots, X^*_{(2\ell+2)q_n})$ for $\ell = 0, \ldots, p_n - 1$, and consider the event $\Xi^* = \{X_i = X^*_i \text{ for all } i = 1, \ldots, n\}$. In addition consider the event $\Xi$ defined as in the proof of Theorem 3.10. We consider the decomposition

$$E\|\hat{\lambda}_m - \lambda\|^2 = \underbrace{E\|\hat{\lambda}_m - \lambda\|^2 I_{\Xi \cap \Xi^*}}_{T_1} + \underbrace{E\|\hat{\lambda}_m - \lambda\|^2 I_{\Xi \cap \Xi^*}}_{T_2} + \underbrace{E\|\hat{\lambda}_m - \lambda\|^2 I_{\Xi \cap \Xi^*}}_{T_3},$$

and bound the three terms separately.

**Upper bound for $T_1$:** Following along the lines of the proofs of Theorems 3.7 and 3.10 one can show

$$E\|\hat{\lambda}_m - \lambda\|^2 I_{\Xi \cap \Xi^*} \leq C \min_{m \in \mathcal{M}_n} \{\|\lambda_m - \lambda\|^2 + \text{pen}(m)\} + 16 \sum_{m' \in \mathcal{M}_n} \underbrace{E\left(\sup_{t \in \mathcal{B}_{m'}} |(\hat{\Theta}, t)|^2 - 50\mu \cdot \frac{\phi^2 D_{m'} \log(n + 2)}{n}\right)}_{T_1} + 16 \sum_{m' \in \mathcal{M}_n} \underbrace{E\left(\sup_{t \in \mathcal{B}_{m'}} |(\hat{\Theta}, t)|^2 - 8\phi^2 D_{m'} \mu \cdot \frac{\sum_{k=0}^{\infty} \beta_k}{n}\right)}_{T_2} + 16 \sum_{m' \in \mathcal{M}_n} \left(8\phi^2 D_{m'} \mu \cdot \frac{\sum_{k=0}^{\infty} \beta_k}{n} - \frac{D_{m'} \log(n + 2)}{8n}\right)_{T_3}.$$

Under the given assumptions on the potential models we obtain for the last term the estimate

$$\sum_{m' \in \mathcal{M}_n} \left(8\phi^2 D_{m'} \mu \cdot \frac{\sum_{k=0}^{\infty} \beta_k}{n} - \frac{D_{m'} \log(n + 2)}{8n}\right) \leq \sum_{m' \in \mathcal{M}_n} \left(8\phi^2 D_{m'} \mu \cdot \frac{\sum_{k=0}^{\infty} \beta_k}{n}\right) \leq \sum_{m=1}^{m_0} \left(8\phi^2 m \mu \cdot \frac{\sum_{k=0}^{\infty} \beta_k}{n}\right) = 4\phi^2 \mu \left(\sum_{k=0}^{\infty} \beta_k\right) \frac{m_0(m_0 + 1)}{n},$$

where $m_0 = m_0(\Phi^2, \mu, (\beta_k)_{k \in \mathbb{N}_0})$ is some non-negative integer that depends on $\mu$, $\Phi$, and the sequence of $\beta$-mixing coefficients. The term $E_{m'1} = \mathbb{E}(\sup_{t \in \mathcal{B}_m} |(\hat{\Theta}, t)|^2 - 50\mu \cdot \frac{\phi^2 D_{m'} \log(n + 2)}{n})$.
\[ \Phi_2^2 D_m (\log (n + 2))_+ \] can be bounded exactly as in the proof of Theorem 3.7 via Lemma C.2 (since the investigation is based on conditioning on the covariates \( X_1, \ldots, X_n \)):

\[
\E \left[ \left( \sup_{t \in B_m} |(\tilde{\Theta}_n, t)|^2 - 50 \mu \cdot \Phi_2^2 D_m \log (n + 2) \right)_+ \right] \leq K_1 \left\{ \frac{D_m'}{n} \exp (-2 \log (n + 2)) + \frac{D_m'}{n^2} \exp (-K'_2 \sqrt{n}) \right\}.
\]

In order to treat the term \( E_{m^2} = \E [(\sup_{t \in B_m} |(\tilde{\Theta}_n, t)|^2 - 8 \Phi_2^2 D_m \mu \cdot \sum_{k=0}^{\infty} \beta_k) \mathbb{I}_{X \in \Xi^*}] \) we use the decomposition \( (\tilde{\Theta}_n, t) \mathbb{I}_{X \in \Xi^*} = (\tilde{\nu}_1^*(t) + \tilde{\nu}_2^*(t))/2 \cdot \mathbb{I}_{X \in \Xi^*} \) where

\[
\tilde{\nu}_1^*(t) = \frac{1}{p_n} \sum_{t=0}^{p_n-1} \frac{1}{q_n} \sum_{t=2(t+1)q_n+1} \{ r_t(X_i^*) - \E [r_t(X_i^*)]\},
\]

\[
\tilde{\nu}_2^*(t) = \frac{1}{p_n} \sum_{t=0}^{p_n-1} \frac{1}{q_n} \sum_{t=2(t+1)q_n+1} \{ r_t(X_i^*) - \E [r_t(X_i^*)]\},
\]

and \( r_t \) is defined as in the proof of Lemma C.1 (note that it is possible to replace \( X_i \) with \( X_i^* \) by the definition of \( \Xi^* \)). By Lemma C.3 we have for \( i = 1, 2 \)

\[
\E \left[ \left( \sup_{t \in B_m} |\tilde{\nu}_i^*(t)|^2 - \frac{8 \Phi_2^2 D_m \mu \cdot \sum_{k=0}^{\infty} \beta_k}{n} \right)_+ \right] \leq c_1 \left\{ \frac{D_m'}{n} \exp (-c_2 q_n) + \frac{D_m'}{p_n^2} \exp (-c_3 \sqrt{n}) \right\}.
\]

Hence,

\[
E_{m^2} \leq \frac{1}{2} \E \left[ \left( \sup_{t \in B_m} |\tilde{\nu}_1^*(t)|^2 - \frac{8 \Phi_2^2 D_m \mu \cdot \sum_{k=0}^{\infty} \beta_k}{n} \right)_+ \right] + \frac{1}{2} \E \left[ \left( \sup_{t \in B_m} |\tilde{\nu}_2^*(t)|^2 - \frac{8 \Phi_2^2 D_m \mu \cdot \sum_{k=0}^{\infty} \beta_k}{n} \right)_+ \right] \leq \frac{D_m'}{n} \cdot \exp (-c_2 q_n) + \frac{D_m'}{p_n^2} \exp (-c_3 \sqrt{n}),
\]

and therefore

\[
\E \| \hat{\lambda} - \lambda \|^2 \mathbb{I}_{X \in \Xi^*} \leq \min_{m \in M_n} \{ \| \lambda_m - \lambda \|^2 + \text{pen}(m) \} + \sum_{m' \in M_n} \left[ \frac{D_m'}{n} \exp (-2 \log (n + 2)) + \frac{D_m'}{n^2} \exp (-K'_2 \sqrt{n}) \right] + \sum_{m' \in M_n} \left[ \frac{D_m'}{n} \cdot \exp (-c_2 q_n) + \frac{D_m'}{p_n^2} \exp (-c_3 \sqrt{n}) \right] + \frac{1}{n}.
\]

Finally, by exploiting Assumption 3.4 on the class of models and the definition of \( q_n \) we obtain

\[
\E \| \hat{\lambda} - \lambda \|^2 \mathbb{I}_{X \in \Xi^*} \leq \min_{m \in M_n} \{ \| \lambda_m - \lambda \|^2 + \text{pen}(m) \} + \frac{1}{n}.
\]

**Upper bound for T2:** As in the proof of Theorem 3.10 one derives the estimate

\[
\E \| \hat{\lambda} - \lambda \|^2 \mathbb{I}_{X \in \Xi^*} \leq 2 \Phi_2^2 \epsilon n T_4 (\| \lambda \|^2 \mathbb{P}(\Xi^* \cap \Xi^* + 1/2 + 2 \| \lambda \|^2 \mathbb{P}(\Xi^* \cap \Xi^*).
\]

Grant to this estimate it suffices to show \( \mathbb{P}(\Xi^* \cap \Xi^*) \leq n^{-1} \). Proceeding as in the proof of Theorem 3.10, we obtain

\[
\mathbb{P}(\Xi^* \cap \Xi^*) \leq \sum_{j=1}^{M} \mathbb{P}(\{ \nu_n (\varphi_j) \geq \xi \} \cap \Xi^*) + \mathbb{P}(\{ \nu_n (\varphi_j) \geq \xi \} \cap \Xi^*)
\]

\[
+ \mathbb{P}(\{ \tilde{\nu}_n (\varphi_j) \geq \xi \} \cap \Xi^*) + \mathbb{P}(\{ \tilde{\nu}_n (\varphi_j) \geq \xi \} \cap \Xi^*)
\]
Lemma C.1. Let \( \nu_n(\varphi_j) \) and \( \tilde{\nu}_n(\varphi_j) \) be defined as in the proof of Theorem 3.10. The probabilities \( \mathbb{P}( \{ \pm \nu_n(\varphi_j) \geq \xi \} \cap \Xi^* ) \) can be handled with exactly as in the proof of Theorem 3.10 by conditioning on the covariates \( X_i \). In order to bound the terms \( \mathbb{P}( \{ \pm \tilde{\nu}_n(\varphi_j) \geq \xi \} \cap \Xi^* ) \), we consider the \( \tilde{\nu}_n^*(\varphi_j) \) defined exactly as \( \tilde{\nu}_n(\varphi_j) \) with \( X_i \) replaced with \( X_i^* \). Note that by construction \( \mathbb{P}( \{ \pm \tilde{\nu}_n(\varphi_j) \geq \xi \} \cap \Xi^* ) = \mathbb{P}( \{ \pm \tilde{\nu}_n^*(\varphi_j) \geq \xi \} \cap \Xi^* ) \). We have the decomposition

\[
\tilde{\nu}_n^*(\varphi_j) = \frac{1}{2} \tilde{\nu}_{1n}^*(\varphi_j) + \frac{1}{2} \tilde{\nu}_{2n}^*(\varphi_j)
\]

where

\[
\tilde{\nu}_{1n}^*(\varphi_j) = \frac{1}{p_n} \sum_{\ell=0}^{p_n-1} \frac{1}{q_n} \sum_{i=2\ell q_n+1}^{(2\ell+1)q_n} \lambda(X_i^*) \varphi_j(X_i^*) - \int_{\mathbb{X}} \lambda(x) \varphi_j(x) \mathbb{P}(dx),
\]

\[
\tilde{\nu}_{2n}^*(\varphi_j) = \frac{1}{p_n} \sum_{\ell=0}^{p_n-1} \frac{1}{q_n} \sum_{i=(2\ell+1)q_n+1}^{(2\ell+2)q_n} \lambda(X_i^*) \varphi_j(X_i^*) - \int_{\mathbb{X}} \lambda(x) \varphi_j(x) \mathbb{P}(dx).
\]

Note that the separate summands of the outer sum are independent. Hence, applying Bernstein’s inequality (see Lemma D.1) yields for \( i = 1, 2 \) that

\[
\mathbb{P}( \{ \pm \tilde{\nu}_n^*(\varphi_j) \geq \xi \} \cap \Xi^* ) \leq \exp \left( -\frac{c \mathbb{P}_n}{64M} \right)
\]

which implies under the stated assumptions that

\[
\mathbb{P}( \Xi^c \cap \Xi^* ) \leq 4M \exp \left( -\frac{c \mathbb{P}_n}{64M} \right) \lesssim n^{-4}.
\]

Upper bound for \( T_3 \): Following along the lines of the proof of the upper bound for term \( T_2 \) in the proof of Theorem 3.10 one can show

\[
\mathbb{E} \| \tilde{\lambda}_m - \lambda \|^2 1_{\Xi^c} \leq 2\Phi^2 c_m n T_4(\|\lambda\|_{\infty})^{1/2} \mathbb{P}(\Xi^c)^{1/2} + 2\|\lambda\|^2 \mathbb{P}(\Xi^c)
\]

with \( T_4 \) being the fourth Touchard polynomial, and it suffices again to show that \( \mathbb{P}(\Xi^c) \lesssim n^{-4} \). It holds

\[
\mathbb{P}(\Xi^c) \leq 2p_n \beta_{q_n} = 2nq_n^{-1} \beta_{q_n} \lesssim n^{-4}
\]

where the last estimate holds due to Assumption 4.3 both in the geometric and the arithmetic case.

Appendix C. Technical Lemmata

Lemma C.1. For all \( m \in M_n \), we have

\[
\mathbb{E} \left[ \left( \sup_{t \in B_m} |\langle \tilde{\Theta}_n, t \rangle |^2 - 3\mu \frac{\Phi^2 D_m}{n} \right)^+ \right] \leq K_1 \left\{ \frac{1}{n} \exp(-K_2 D_m) + \frac{D_m}{n^2} \exp(-K_3 \sqrt{n}) \right\}
\]

with strictly positive numerical constants \( K_1 \), \( K_2 \), and \( K_3 \).

Proof. With \( t \in B_m \), we associate the function

\[
r_t(x) := \sum_{\eta \in I_m} \tau_{\eta} \lambda(x) \varphi_{\eta}(x)
\]

where the \( \tau_{\eta} = \int_{\mathbb{X}} \varphi_{\eta}(x) t(x) \mathbb{P}(dx) \) denote for \( \eta \in I_m \) the coefficients of the function \( t \) in terms of the basis given by the \( \varphi_{\eta} \). Evidently, we have \( \mathbb{E}[r_t(X)] = \sum_{\eta \in I_m} \tau_{\eta} \beta_{\eta} \). Consequently, one has the identity

\[
\langle \tilde{\Theta}_n, t \rangle = \frac{1}{n} \sum_{i=1}^{n} r_t(X_i) - \mathbb{E}[r_t(X_i)],
\]

and \( \langle \tilde{\Theta}_n, t \rangle \) will take the role of \( \nu_n \) in Lemma E.1. We now check the preconditions concerning the existence of suitable constants \( M_1 \), \( H \) and \( v \) in the statement of Lemma E.1.
Condition concerning $M_1$: We have
\[
\sup_{t \in \mathcal{B}_m} \|r_t\|_\infty^2 = \sup_{t \in \mathcal{B}_m} \|r_t(\cdot)\|_{L^2}^2 \leq \sup_{t \in \mathcal{B}_m} \sup_{y \in [0,1]} |r_t(y)|^2 \leq \sup_{t \in \mathcal{B}_m} \sup_{y \in [0,1]} \left( \sum_{\eta \in \mathcal{I}_m} \tau_\eta^2 \right) \left( \sum_{\eta \in \mathcal{I}_m} \lambda^2(y) \cdot \varphi_\eta^2(y) \right)
\]
\[
\leq \|\lambda\|_\infty^2 \Phi^2 D_m \leq \mu \Phi^2 D_m,
\]
and we can put $M_1 := (\mu \Phi^2 D_m)^{1/2}$.

Condition concerning $H$: We have
\[
\mathbb{E}\left[ \sup_{t \in \mathcal{B}_m} |\tilde{\Theta}_n(t)|^2 \right] \leq \frac{1}{n^2} \mathbb{E}\left[ \sup_{t \in \mathcal{B}_m} \left( \sum_{\eta \in \mathcal{I}_m} \tau_\eta^2 \right) \left( \sum_{\eta \in \mathcal{I}_m} \left| \sum_{i=1}^n \varphi_\eta(X_i) \lambda(X_i) - \theta_\eta \right| \right)^2 \right]
\]
\[
= \frac{1}{n} \sum_{\eta \in \mathcal{I}_m} \mathbb{Var}(\varphi_\eta(X_1) \lambda(X_1)) \leq \frac{1}{n} \sum_{\eta \in \mathcal{I}_m} \mathbb{E}\left[ \left( \varphi_\eta(X_1) \lambda(X_1) \right)^2 \right]
\]
\[
\leq \frac{\Phi^2 D_m}{n} \|\lambda\|_\infty^2 \leq \frac{\mu \Phi^2 D_m}{n},
\]
and thus by Jensen’s inequality we can put $H := \left( \frac{\mu \Phi^2 D_m}{n} \right)^{1/2}$.

Condition concerning $\nu$: For arbitrary $t \in \mathcal{B}_m$, it holds
\[
\mathbb{Var}(r_t(X)) = \mathbb{Var}\left( \sum_{\eta \in \mathcal{I}_m} \tau_\eta \lambda(X) \varphi_\eta(X) \right) \leq \mathbb{E}\left[ \left( \sum_{\eta \in \mathcal{I}_m} \tau_\eta \lambda(X) \varphi_\eta(X) \right)^2 \right] \leq \mu.
\]

Thus, we can take $\nu := \mu$ and the statement of the lemma follows now by applying Lemma E.1 with $\varepsilon = \frac{1}{2}$. \hfill \qed

Lemma C.2. With the notation from the proof of Theorem 3.7 it holds for all $m \in \mathcal{M}_n$
\[
\mathbb{E}\left[ \left( \sup_{t \in \mathcal{B}_m} |\tilde{\Theta}_n(t)|^2 - 50 \mu \cdot \frac{\Phi^2 D_m \log(n + 2)}{n} \right)^+ \right] \leq K'_1 \left\{ \frac{D_m}{n} \exp(-2 \log(n + 2)) + \frac{D_m}{n^2} \exp(-K'_2 \sqrt{n}) \right\}
\]
with strictly positive numerical constants $K'_1$ and $K'_2$.

Proof. Given $X = (X_1, \ldots, X_n)$, we can write $Y_i$ as $\int_0^1 dN_i(s)$ where $N_i$ is a Poisson process with homogeneous intensity equal to $\lambda(X_i)$. Thus, conditional on $X$, it holds
\[
\langle \tilde{\Theta}_n, t \rangle = \frac{1}{n^2} \sum_{\eta \in \mathcal{I}_m} \tau_\eta \sum_{i=1}^n \left\{ \int_0^1 \varphi_\eta(X_i) dN_i(s) - \varphi_\eta(X_i) \cdot \lambda(X_i) \right\}
\]
\[
= \frac{1}{n} \sum_{i=1}^n \left\{ \int_0^1 r_t(s) dN_i(s) - \int_0^1 r_t(s) \lambda(X_i) ds \right\}
\]
where $r_t$ is the function given by $r_t(s) := \sum_{\eta \in \mathcal{I}_m} \tau_\eta \varphi_\eta(X_i)$ (note that, given $X$, this is a constant function). We now check the preconditions concerning the existence of suitable constants $M_1$, $H$ and $\nu$ from Lemma E.2.

Condition concerning $M_1$: We have
\[
\sup_{t \in \mathcal{B}_m} \|r_t\|_\infty^2 = \sup_{t \in \mathcal{B}_m} \left( \sum_{\eta \in \mathcal{I}_m} \tau_\eta \varphi_\eta(X_i) \right)^2 \leq \sup_{t \in \mathcal{B}_m} \left( \sum_{\eta \in \mathcal{I}_m} \tau_\eta^2 \right) \cdot \left( \sum_{\eta \in \mathcal{I}_m} \varphi_\eta^2(X_i) \right) \leq \Phi^2 D_m,
\]
and we can take $M_1 = \sqrt{\mu \cdot \Phi^2 D_m}$. 
Condition concerning $H$: It holds
\[
\mathbb{E}[\sup_{t \in \mathcal{B}_m} |\tilde{\Theta}_n(t)|^2 |X] \\
\leq \sup_{t \in \mathcal{B}_m} \left( \sum_{\eta \in \mathcal{I}_m} \frac{\tau^2_{\eta}}{n} \right) \mathbb{E} \left[ \sum_{\eta \in \mathcal{I}_m} \left\{ \int_0^1 \varphi_\eta(X_1)[dN_i(s) - \lambda(X_1)ds] \right\}^2 \right] \\
\leq \frac{1}{n} \sum_{\eta \in \mathcal{I}_m} \text{Var} \left( \int_0^1 \varphi_\eta(X_1)dN_1(s) \right) \lambda(X_1) \\
= \frac{1}{n} \sum_{\eta \in \mathcal{I}_m} \int_0^1 \varphi^2_\eta(X_1) \lambda(X_1)ds \\
\leq \frac{\Phi^2 \mathbb{D}_m}{n} \cdot \|\lambda\|_\infty \leq \frac{\Phi^2 \mathbb{D}_m}{n} \cdot \mu.
\]
Thus, we can put $H := \left( \frac{\Phi^2 \mathbb{D}_m \mu \log(n+2)}{n} \right)^{1/2}$ (the additional enlargement of $H$ by the logarithmic term is needed in the proof of Theorem 3.7).

Condition concerning $v$: For arbitrary $m \in \mathcal{M}_n$ and $t \in \mathcal{B}_m$ it holds
\[
\text{Var} \left( \int_0^1 r_t(s)dN_k(s)|X_k \right) = \int_0^1 |r_t(s)|^2 \lambda(X_k)ds \leq \|\lambda\|_\infty \cdot |r_t|^2_{\infty} \leq \|\lambda\|_\infty \Phi^2 \mathbb{D}_m,
\]
and we can put $v := \mu \Phi^2 \mathbb{D}_m$.

We can apply Lemma E.2 with $\varepsilon = 12$ which yields
\[
\mathbb{E} \left[ \left( \sup_{t \in \mathcal{B}_m} |\tilde{\Theta}_n(t)|^2 - 50\mu \cdot \frac{\Phi^2 \mathbb{D}_m \log(n+2)}{n} \right) + \left|X\right| \right] \leq \]
\[
K' \left\{ \frac{\Phi^2 \mathbb{D}_m \mu}{n} \exp(-2\log(n+2)) + \frac{\Phi^2 \mathbb{D}_m \mu}{n^2} \exp(-K' \sqrt{n \log(n+2)}) \right\}.
\]
Since the right-hand side of the last estimate does not depend on $X$, taking expectations on both sides implies the assertion of the lemma. \hfill \Box

Lemma C.3. For $\tilde{\nu}_i^* (\cdot)$ defined as in the proof of Theorem 4.4 it holds for $i = 1, 2$ that
\[
\mathbb{E} \left[ \sup_{t \in \mathcal{B}_m} \left| \tilde{\nu}_i^*(t) - \frac{8\Phi^2 \mathbb{D}_m \mu (\sum_{k=0}^{\infty} \beta_k)}{n} \right| \right] \leq c_1 \left\{ \frac{\mathbb{D}_m}{n} \exp(-c_2 q_n) + \frac{\mathbb{D}_m}{p_n} \exp(-c_3 \sqrt{n}) \right\}
\]
with strictly positive numerical constants $c_1$, $c_2$, and $c_3$.

Proof. We state the proof for $i = 1$ only. We want to apply Lemma E.1 to $\tilde{\nu}_1 = \frac{1}{\nu_n} \sum_{\ell=0}^{p_n-1} Z_\ell$ where $Z_\ell = \frac{1}{q_n} \sum_{i=2^{\ell+1}q_n+1}^{(2^{\ell+1}+1)q_n} \{r_t(X_i^*) - \mathbb{E}r_t(X_i^*) \}$, and the function $r_t$ is defined as in Lemma C.1. Note that $Z_\ell$ and $Z_{\ell'}$ are independent for $\ell \neq \ell'$ by construction. Thus, it remains to find constants $M_1, H$ and $v$ satisfying the preconditions of Lemma E.1.

Condition concerning $M_1$: For each $\ell = 0, \ldots, p_n - 1$ it holds
\[
\sup_{t \in \mathcal{B}_m} \left\| \frac{1}{q_n} \sum_{i=2^{\ell+1}q_n+1}^{(2^{\ell+1}+1)q_n} r_t \right\|_\infty \leq \sup_{t \in \mathcal{B}_m} |r_t|_{\infty} \leq \sqrt{\mu \Phi^2 \mathbb{D}_m}
\]
where the last step has already been shown in the proof of Lemma C.1.

Condition concerning $H$: By Lemma B.1 we have
\[
\mathbb{E} \left[ \sup_{t \in \mathcal{B}_m} \left| \frac{1}{p_n q_n} \sum_{\ell=0}^{p_n-1} \sum_{i=2^{\ell+1}q_n+1}^{(2^{\ell+1}+1)q_n} \{r_t(X_i^*) - \mathbb{E}r_t(X_i^*) \} \right|^2 \right]
\]
Lemma D.1 (Bernstein’s inequality, [BLM16], Corollary 2.11)

We refer to [Cha13].

A consequence from Talagrand’s inequality and is taken from [CL15]. For a detailed proof, we refer to [Cha13].

Condition concerning $\nu$: We have

$$\text{Var} \left( \frac{1}{q_n} \sum_{i=2q_n+1}^{(2\ell+1)q_n} r_i(X_i^*) \right) = \frac{1}{q_n^2} \text{Var} \left( \sum_{i=2q_n+1}^{(2\ell+1)q_n} \tau_{i\eta} \lambda(X_i^*) \varphi_{i\eta}(X_i^*) \right)$$

$$\leq \frac{4 \Phi^2 D_m}{p_n q_n} \left( \sum_{k=0}^{n} \beta_k \right) \cdot \|\lambda\|_\infty^2$$

Now, application of Lemma E.1 yields

$$\mathbb{E} \left[ \left( \sup_{t \in \mathcal{B}_m} |\hat{p}_t(t)|^2 - \frac{8 \Phi^2 D_m \mu(\sum_{k=0}^{\infty} \beta_k)}{n} \right) \right] \leq c_4 \left\{ \frac{D_m}{n} \cdot \exp(-c_2 q_n) + \frac{D_m}{q_n^2} \exp(-c_3 \sqrt{n}) \right\}.$$
that are conditionally independent given \( \eta \).

Let \( \text{Lemma E.2.} \)

Concentration inequalities for point processes. The following lemma is taken from \[\text{Kro17}\].

**Lemma E.2.** Let \( N_1, \ldots, N_n \) be independent Cox processes driven by finite random measures \( \eta_1, \ldots, \eta_n \) (that is, given \( \eta_i \), \( N_i \) is a Poisson point process with intensity measure \( \eta_i \)) that are conditionally independent given \( \eta_1, \ldots, \eta_n \). Set \( \nu_n(r) = \frac{1}{n} \sum_{k=1}^{n} \{ \int_{X} r(x) dN_k(x) - \int_{X} r(x) d\eta_k(x) \} \) for \( r \) contained in a countable class of real-valued measurable functions.

Then, for any \( \varepsilon > 0 \), there exist constants \( c_1, c_2 = \frac{1}{6} \), and \( c_3 \) such that

\[
\mathbb{E} \left[ \left( \sup_{r \in \mathcal{R}} |\nu_n(r)|^2 - c(\varepsilon)H^2 \right)^2 \right] + \mathbb{E} \left[ \left( \sup_{r \in \mathcal{R}} |\nu_n(r)| \right)^2 \right] \leq c_1 \mathbb{E} \left[ \left( \sup_{r \in \mathcal{R}} |\nu_n(r)| \right)^2 \right] \leq c_1 \left\{ \frac{\varepsilon}{n} \exp \left( -c_2 \varepsilon \frac{nH^2}{\nu} \right) + \frac{M^2}{C^2(\varepsilon)n^2} \exp \left( -c_3 C(\varepsilon) \sqrt{\frac{nH}{M_1}} \right) \right\},
\]

where \( \mathbf{\eta} = (\eta_1, \ldots, \eta_n) \), \( C(\varepsilon) = (\sqrt{1 + \varepsilon} - 1) \wedge 1 \), \( c(\varepsilon) = 2(1 + 2\varepsilon) \) and \( M_1, H \) and \( \nu \) are such that

\[
\sup_{r \in \mathcal{R}} \|r\|_\infty \leq M_1, \quad \mathbb{E}[\sup_{r \in \mathcal{R}} |\nu_n(r)|] \leq H, \quad \sup_{r \in \mathcal{R}} \mathbb{V}ar \left( \int_{X} r(x) dN_k(x) \right) \leq \nu \quad \forall k.
\]

The following lemma is a Bernstein type inequality for point processes and taken from \[\text{RB03}\].

**Lemma E.3** (\[\text{RB03}\], Proposition 7). Let \( N \) be a Poisson point process on some measurable space \( (X, \mathcal{F}) \) with finite intensity measure \( \mu \). Let \( g \) be a measurable function on \( (X, \mathcal{F}) \), essentially bounded, such that \( \int_{X} g^2(x) \mu(dx) > 0 \). Then

\[
\mathbb{P} \left( \int_{X} g(x)(dN(x) - \mu(dx)) \geq t \right) \leq \exp \left( -\frac{t^2}{2 \left( \int_{X} g^2(x) \mu(dx) \right) + \|g\|_\infty t/3} \right), \quad t > 0.
\]

**References**

[AC10] Arlot, S. and Celisse, A. A survey of cross-validation procedures for model selection. *Stat. Surv.* 4 (2010), 40–79.

[AJ16] Asin, N. and Johannes, J. Adaptive non-parametric estimation in the presence of dependence. *ArXiv preprint* (2016). arXiv: 1602.00531.

[Ans48] Anscombe, F. J. The transformation of Poisson, binomial and negative-binomial data. *Biometrika* 35 (1948), 246–254.

[Bar00] Baraud, Y. Model selection for regression on a fixed design. *Probab. Theory Related Fields* 117 (2000), 467–493.

[BBM99] Barron, A., Birgé, L., and Massart, P. Risk bounds for model selection via penalization. *Probab. Theory Related Fields* 113 (1999), 301–413.

[BCV01] Baraud, Y., Comte, F., and Viennet, G. Adaptive estimation in autoregression or \( \beta \)-mixing regression via model selection. *Ann. Statist.* 29 (2001), 839–875.

[BFS04] Besbeas, P., Feis, I. de, and Sapatinas, T. A Comparative Simulation Study of Wavelet Shrinkage Estimators for Poisson Counts. *International Statistical Review / Revue Internationale de Statistique* 72 (2004), 209–237.
[BLM16] Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: a nonasymptotic theory of independence*. Oxford University Press, Oxford, 2016.

[BM08] Berk, R. and MacDonald, J. M. Overdispersion and Poisson Regression. *Journal of Quantitative Criminology* 24 (2008), 269–284.

[BM97] Birgé, L. and Massart, P. From model selection to adaptive estimation. In: *Festschrift for Lucien Le Cam*. Springer, New York, 1997, 55–87.

[Bos98] Bosq, D. *Nonparametric statistics for stochastic processes*. Second. Vol. 110. Lecture Notes in Statistics. Estimation and prediction. Springer-Verlag, New York, 1998, pp. xvi+210.

[BTW07] Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. Aggregation for Gaussian regression. *Ann. Statist.* 35 (2007), 1674–1697.

[Cha13] Chagny, G. Estimation adaptative avec des données transformées ou incomplètes. Application à des modèles de survie. PhD thesis. 2013.

[CL15] Chagny, G. and Lacour, C. Optimal adaptive estimation of the relative density. *TEST* 24 (2015), 605–631.

[Com01] Comte, F. Adaptive estimation of the spectrum of a stationary Gaussian sequence. *Bernoulli* 7 (2001), 267–298.

[CP02] Carota, C. and Parmigiani, G. Semiparametric regression for count data. *Biometrika* 89 (2002), 265–281.

[CT98] Cameron, A. C. and Trivedi, P. K. *Regression analysis of count data*. Vol. 30. Econometric Society Monographs. Cambridge University Press, Cambridge, 1998, pp. xviii+411.

[Dou94] Doukhan, P. *Mixing*. Vol. 85. Lecture Notes in Statistics. Properties and examples. Springer-Verlag, New York, 1994, pp. xii+142.

[DTM98] Diggle, P. J., Tawn, J. A., and Moyeed, R. A. Model-based geostatistics. *J. Roy. Statist. Soc. Ser. C* 47 (1998). With discussion and a reply by the authors, 299–350.

[Fry08] Fryzlewicz, P. Data-driven wavelet-Fisz methodology for nonparametric function estimation. *Electron. J. Stat.* 2 (2008), 863–896.

[FZ99] Fan, J. and Zhang, W. Statistical estimation in varying coefficient models. *Ann. Statist.* 27 (1999), 1491–1518.

[GL11] Goldenshluger, A. and Lepski, O. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist.* 39 (2011), 1608–1632.

[GLT16a] Guilloux, A., Lemler, S., and Taupin, M.-L. Adaptive estimation of the baseline hazard function in the Cox model by model selection, with high-dimensional covariates. *J. Statist. Plann. Inference* 171 (2016), 38–62.

[GLT16b] Guilloux, A., Lemler, S., and Taupin, M.-L. Adaptive kernel estimation of the baseline function in the Cox model with high-dimensional covariates. *J. Multivariate Anal.* 148 (2016), 141–159.

[HT93] Hastie, T. and Tibshirani, R. Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* 55 (1993). With discussion and a reply by the authors, 757–796.

[IPR16] Ivanoff, S., Picard, F., and Rivoirard, V. Adaptive Lasso and group-Lasso for functional Poisson regression. *J. Mach. Learn. Res.* 17 (2016), Paper No. 55, 46.

[KR60] Kolmogorov, A. N. and Rozanov, J. A. On a strong mixing condition for stationary Gaussian processes. *Teor. Veroiatnost. i Primenen.* 5 (1960), 222–227.

[Kro17] Kroll, M. Concentration inequalities for Poisson point processes with applications to non-parametric statistics. PhD thesis. 2017.
[KYS13] Kamo, K.-I., Yanagihara, H., and Satoh, K. Bias-corrected AIC for selecting variables in Poisson regression models. *Comm. Statist. Theory Methods* 42 (2013), 1911–1921.

[Lac08] Lacour, C. Adaptive estimation of the transition density of a particular hidden Markov chain. *J. Multivariate Anal.* 99 (2008), 787–814.

[Lep91] Lepski, O. V. On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* 35 (1991), 454–466.

[LM09] Lecué, G. and Mendelson, S. Aggregation via empirical risk minimization. *Probab. Theory Related Fields* 145 (2009), 591–613.

[LM16] Lacour, C. and Massart, P. Minimal penalty for Goldenshluger-Lepski method. *Stochastic Process. Appl.* 126 (2016), 3774–3789.

[LMR17] Lacour, C., Massart, P., and Rivoirard, V. Estimator selection: a new method with applications to kernel density estimation. *Sankhya A* 79 (2017), 298–335.

[Mas07] Massart, P. *Concentration inequalities and model selection*. Vol. 1896. Lecture Notes in Mathematics. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. Springer, Berlin, 2007, pp. xiv+337.

[Mok90] Mokkadem, A. Propriétés de mélange des processus autorégressifs polynomiaux. *Ann. Inst. H. Poincaré Probab. Statist.* 26 (1990), 219–260.

[Nak+05] Nakaya, T. et al. Geographically weighted Poisson regression for disease association mapping. *Stat. Med.* 24 (2005), 2695–2717.

[Neu98] Neumann, M. H. Strong approximation of density estimators from weakly dependent observations by density estimators from independent observations. *Ann. Statist.* 26 (1998), 2014–2048.

[NT06] Neumann, M. H. and Thorarinsdottir, T. L. Asymptotic minimax estimation in nonparametric autoregression. *Math. Methods Statist.* 15 (2006), 374–397.

[RB03] Reynaud-Bouret, P. Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Related Fields* 126 (2003), 103–153.

[RB04] Ridout, M. S. and Besbeas, P. An empirical model for underdispersed count data. *Stat. Model.* 4 (2004), 77–89.

[Rei93] Reiss, R.-D. *A course on point processes*. Springer Series in Statistics. Springer-Verlag, New York, 1993, pp. xii+253.

[Roo03] Roos, B. Improvements in the Poisson approximation of mixed Poisson distributions. *J. Statist. Plann. Inference* 113 (2003), 467–483.

[RT12] Rigollet, P. and Tsybakov, A. B. Sparse estimation by exponential weighting. *Statist. Sci.* 27 (2012), 558–575.

[Sau13] Saumard, A. Optimal model selection in heteroscedastic regression using piecewise polynomial functions. *Electron. J. Stat.* 7 (2013), 1184–1223.

[Tsy09] Tsybakov, A. B. *Introduction to nonparametric estimation*. Springer Series in Statistics. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. Springer, New York, 2009, pp. xii+214.

[VHB07] Ver Hoef, J. M. and Boveng, P. L. Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology* 88 (2007), 2766–2772.

[Vie97] Viennet, G. Inequalities for absolutely regular sequences: application to density estimation. *Probab. Theory Related Fields* 107 (1997), 467–492.

[Win08] Winkelmann, R. *Econometric analysis of count data*. Springer Verlag, Berlin, 2008.