The L_2 Rate of Convergence for Event History Regression with Time-dependent Covariates

JIANHUA Z. HUANG

The Wharton School, University of Pennsylvania

CHARLES J. STONE

University of California, Berkeley

ABSTRACT. Consider repeated events of multiple kinds that occur according to a right-continuous semi-Markov process whose transition rates are influenced by one or more time-dependent covariates. The logarithms of the intensities of the transitions from one state to another are modelled as members of a linear function space, which may be finite- or infinite-dimensional. Maximum likelihood estimates are used, where the maximizations are taken over suitably chosen finite-dimensional approximating spaces. It is shown that the L_2 rates of convergence of the maximum likelihood estimates are determined by the approximation power and dimension of the approximating spaces. The theory is applied to a functional ANOVA model, where the logarithms of the intensities are approximated by functions having the form of a specified sum of a constant term, main effects (functions of one variable), and interaction terms (functions of two or more variables). It is shown that the curse of dimensionality can be ameliorated if only main effects and low-order interactions are considered in functional ANOVA models.

Key words: ANOVA decomposition, intensity functions, maximum likelihood, semi-Markov process, tensor product splines, time-dependent covariates

1. Introduction

Event history analysis is closely related to survival analysis and heavily used in economics, sociology and other social sciences. See, for example, Allison (1984), Hamerle (1989), Mayer & Tuma (1990) and Yamaguchi (1991) for general discussions and applications of this subject. Typically, an "event history" can be characterized by a multi-state stochastic process moving among a finite number of states as time progresses with events corresponding to transitions between states. Event history analysis is concerned with the (possibly repeated) occurrence and duration of events and, typically, their dependence on some explanatory variables (covariates). Censoring and time-dependent covariates are the two typical features among the event history data that create difficulties in the statistical analysis.

In this paper, we consider events that occur according to a right-continuous semi-Markov process. As illustrated on p. 59 of Allison (1984), transitions are allowed from a state to itself. For example, the transitions could be job changes and the states could be employer types: academic positions in "high-quality" departments, academic positions in other departments, and non-academic positions; and one can make a transition from a non-academic employer to another non-academic employer. This setup has general applicability in event history analysis and, in particular, it includes survival and competing risk models as special cases. Since the intrinsic time scale of the semi-Markov process is duration rather than "calendar" time, we model the log-intensity of a transition from one state to another as a function of duration in the former state.

Let $1, \ldots, K$ index the possible states (kinds of events) of the system; set $T_0 = 0$ and let $T_1, T_2, \ldots$ denote the successive transitions of the system; and let $Y(t)$ and $\mathbf{X}(t)$ denote,

respectively, the state of the system and the value of the covariates at time $t \geqslant 0$. It is assumed that $\mathbf{X}(t)$ ranges over a known subset $\mathscr{X}$ of some Euclidean space. We refer to the successive transitions $T_1, T_2, \ldots$ and the state process $Y = Y(\cdot)$ as together forming the *event history system.* The *covariate process* $\mathbf{X} = \mathbf{X}(\cdot)$ is assumed to be *external,* that is, not directly involved with the event history system [see p. 123 of Kalbfleisch & Prentice (1980)]. It is also assumed that the log-intensity for a transition from state w to state y at time t has the form $\alpha_{wy}(t - T_{\nu-1}, \mathbf{X}(t)) = \log \lambda_{wy}(t - T_{\nu-1}, \mathbf{X}(t))$ for $T_{\nu-1} \leqslant t < T_\nu$. Then the log-intensity for a transition from state w at time t is given by

$$\begin{aligned}\alpha_w(t - T_{\nu-1}, \mathbf{X}(t)) &= \log \lambda_w(t - T_{\nu-1}, \mathbf{X}(t)) \\ &= \log\left(\sum_y \lambda_{wy}(t - T_{\nu-1}, \mathbf{X}(t))\right), \quad T_{\nu-1} \leqslant t < T_\nu.\end{aligned}$$

Here if a transition from state w to state y is impossible, then $\alpha_{wy} = -\infty$ and $\lambda_{wy} = 0$; otherwise, it is assumed that $\alpha_{wy}(t, \mathbf{x})$ is finite for all $t \geqslant 0$ and $\mathbf{x} \in \mathscr{X}$. The state w is said to be absorbing if $\alpha_w = -\infty$ and $\lambda_w = 0$ and non-absorbing otherwise. Which states are absorbing and which are non-absorbing is assumed to be known. From now on, we ignore pairs w, y with $\alpha_{wy} = -\infty$ and let $\boldsymbol{\alpha} = (\alpha_{wy})$ denote the collection of finite log-intensity functions; we refer to the functions α_{wy} as the *constituents* of $\boldsymbol{\alpha}$. In traditional applications to survival analysis, there are two states ($K = 2$), one of which is absorbing and the other non-absorbing, the semi-Markov process starts out in the non-absorbing state and, at the one and only one transition, it moves to the absorbing state, which corresponds to death or other form of failure.

Let the system be observed during the time period $[0, C]$, where C is the censoring time. It is assumed that the censoring time and the event history system are conditionally independent given the covariate process. Set $\delta_\nu = \text{ind}\,(T_\nu \leqslant C \text{ and } T_\nu < \infty)$ and write $\min(t, c)$ as $t \wedge c$. Then, conditioned on the covariate process, the log-likelihood corresponding to the candidate $\boldsymbol{\alpha} = (\alpha_{wy})$ for $\boldsymbol{\alpha}$ equals

$$\sum_\nu \left(\delta_\nu a_{Y(T_{\nu-1}),Y(T_\nu)}(T_\nu - T_{\nu-1}, \mathbf{X}(T_\nu)) - \int_{T_{\nu-1}\wedge C}^{T_\nu \wedge C} \exp\{\alpha_{Y(T_{\nu-1})}(t - T_{\nu-1}, \mathbf{X}(t))\}\, dt\right);$$

here $\exp(\alpha_w) = \sum_y \exp(\alpha_{wy})$, $\exp(\alpha_{wy})$ and $\exp(\alpha_w)$ are defined to be zero when w is an absorbing state, and the indicated integral over $[T_{\nu-1} \wedge C, T_\nu \wedge C]$ is defined to be zero when $T_{\nu-1} > C$. Observe that the log-likelihood is concave in $\mathbf{a}$. Let $\Lambda(\mathbf{a})$ denote the expected value of this log-likelihood.

Consider n pairs of event history systems and covariate processes that are independent of each other and identically distributed as described above, which we refer to as the random sample of size n. Let T_{iv}, δ_{iv}, $Y_i(t)$, $\mathbf{X}_i(t)$, and C_i denote the values of T_ν, δ_ν, $Y(t)$, $\mathbf{X}(t)$, and C for the ith such case. Then the scaled log-likelihood corresponding to the candidate $\mathbf{a}$ for $\boldsymbol{\alpha}$ and the random sample is given by

$$l(\mathbf{a}) = \frac{1}{n}\sum_i l_i(\mathbf{a}),$$

where

$$l_i(\mathbf{a}) = \sum_{\nu} \Bigg(\delta_{i\nu} a_{Y_i(T_{i,\nu-1}),Y_i(T_{i\nu})}(T_{i\nu} - T_{i,\nu-1}, \mathbf{X}_i(T_{i\nu})) - \int_{T_{i,\nu-1}\wedge C_i}^{T_{i\nu}\wedge C_i} \exp\{a_{Y_i(T_{i,\nu-1})}(t - T_{i,\nu-1}, \mathbf{X}_i(t))\}\, dt \Bigg).$$

In this paper, it is assumed that the constituents of $\boldsymbol{\alpha}$ (that is, the log-intensity functions) belong to a linear function space H, which specifies the functional form of the log-intensity functions. Maximum likelihood estimation is used to fit the data, where the maximization is carried out over the vectors whose constituents lie in a suitably chosen finite-dimensional approximating subspace G of H. Let $\hat{\boldsymbol{\alpha}}_n = (\hat{\alpha}_{wy})$, $\hat{\alpha}_{wy} \in G$, denote the maximum-likelihood estimate. We can think of $\hat{\boldsymbol{\alpha}}_n$ as an estimate of $\boldsymbol{\alpha}$. In general, the constituents of $\boldsymbol{\alpha}$ need not belong to H. In that case, we can think of $\hat{\boldsymbol{\alpha}}_n$ as estimating the best approximation $\boldsymbol{\alpha}^* = (\alpha^*_{wy})$ to $\boldsymbol{\alpha}$, whose constituents α^*_{wy} are chosen to maximize the expected log-likelihood among all vectors with constituents in H. In fact, we shall see that the rate of convergence of $\hat{\boldsymbol{\alpha}}_n$ to $\boldsymbol{\alpha}^*$ is determined by the approximation power and dimension of G.

The general setup in the previous paragraph allows us to give a unified treatment of the parametric and non-parametric approaches in estimation. If H is finite-dimensional, we can chose $G = H$; this corresponds to the parametric approach. On the other hand, if H is infinite-dimensional, we chose G as a finite-dimensional subspace of H whose dimension goes to infinity with sample size, for example, a space of polynomials or splines; this corresponds to the non-parametric approach. The unrestricted non-parametric approach is well known to be subject to the curse of dimensionality when there are many covariates. A remedy is to restrict the form of functions in H. A natural approach is to consider the functional ANOVA model, where H consists of functions having the form of a specified sum of a constant term, main effects (functions of one variable), and selected interaction terms (functions of two or more variables); see Stone *et al.* (1997). As special cases, in an additive model only the constant term and the main effects are considered, while the saturated model includes all possible interaction terms in addition to the terms considered by an additive model. We shall see that the rates of convergence in functional ANOVA models are determined by the smoothness of the components in the ANOVA decomposition of α^*_{wy} and the highest order of interactions considered. By considering only main effects and low-order interactions, we can ameliorate the curse of dimensionality that the saturated model suffers.

Recently, Kooperberg *et al.* (1995a) developed an adaptive methodology (HARE) using polynomial splines and their tensor products to estimate the conditional log-hazard function in the context of survival analysis. The rate of convergence for the corresponding non-adaptive procedure, which provides a guide for the adaptive methodology, was given in Kooperberg *et al.* (1995b). We extend this result in several directions: first, we can handle repeated events of multiple kinds; secondly, we allow time-dependent covariates; finally, we use virtually arbitrary linear spaces of functions and their tensor products as building blocks for the approximating space. The implication of our result is that HARE methodoloy can be extended to handle the event history data with time-dependent covariates, and moreover, that linear spaces other than splines can be used in the estimation procedure.

The rest of the paper is organized as follows. Section 2.1 discusses the existence of the best approximation to the target log-intensity functions in the model space. Section 2.2 provides a general result on the rate of convergence of the maximum likelihood estimate. The functional ANOVA model is considered in section 2.3. We provide some useful preliminary results in section 3. The proofs of the theorems are given in sections 4, 5, and 6. Some technical details are collected in the appendix.

For any function a on $\mathscr{T} \times \mathscr{X}^{\cdot}$, set $\|a\|_\infty = \sup_{t\in\mathscr{T},\mathbf{x}\in\mathscr{X}^{\cdot}} |a(t, \mathbf{x})|$; for $\mathbf{a} = (a_{wy})$, with each a_{wy} a function on $\mathscr{T} \times \mathscr{X}^{\cdot}$, set $\|\mathbf{a}\|_\infty = \sum_{w,y} \|a_{wy}\|_\infty$. Given positive numbers a_n and b_n for $n \geqslant 1$, let $a_n \asymp b_n$ mean that a_n/b_n is bounded away from zero and infinity. Given random variables W_n for $n \geqslant 1$, let $W_n = O_P(b_n)$ mean that $\lim_{c\to\infty} \limsup_n P(|W_n| \geqslant cb_n) = 0$.

2. Statement of results

2.1. Existence of the best approximation

Let $\Lambda(\cdot)$ now denote the restriction of the expected log-likelihood function to vectors $\mathbf{a} = (a_{wy})$ with constituents in H, which we refer to as the expected log-likelihood function corresponding to H. The first goal is to prove that this expected log-likelihood function has a maximum and that this maximum equals the vector $\boldsymbol{\alpha}$ of the true log-intensities if the constituents of α are in H. Suppose that $P(C \leqslant \tau) = 1$, where τ is a fixed positive number, and set $\mathscr{T} = [0, \tau]$.

Condition 1

(i) $P(Y(0) = w|\mathbf{X})$ *is bounded away from zero uniformly in* $\mathbf{X}$ *and non-absorbing states* w;
(ii) $P(C = \tau|\mathbf{X})$ *is bounded away from zero uniformly in* $\mathbf{X}$;
(iii) *the density function of* $\mathbf{X}(t)$ *exists and is bounded away from zero and infinity on* $\mathscr{X}^{\cdot}$ *uniformly over* $t \in \mathscr{T}$;
(iv) *the constituents of* $\boldsymbol{\alpha}$ *are bounded on* $\mathscr{T} \times \mathscr{X}^{\cdot}$.

In the context of hazard regression, condition 1 is implied by cond. 1 and 2 of Kooperberg *et al.* (1995b). The first part of condition 1 is needed to estimate $\alpha_{wy}(t)$ efficiently near time τ. According to condition 1(ii), censoring automatically occurs at time τ if it does not occur before this time.

Condition 2

The space H *is closed in the following sense: if* $h_n \in H$ *and* $h_n \to h$ *in measure, then* $h \in H$.

This condition is satisfied by finite-dimensional spaces H and, in the interesting multivariate situation, it is also satisfied by the space of functions having the form of a specified sum of functions of selected variables; see lem. 4.1 of Stone (1994).

Theorem 2.1

Suppose conditions 1 *and* 2 *hold. Then there exists an essentially uniquely determined function* $\boldsymbol{\alpha}^* = (\alpha^*_{wy})$ *whose constituents are in* H *that maximizes the expected log-likelihood function corresponding to* H. *If the constituents of* $\boldsymbol{\alpha}$ *are in* H, *then* $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}$ *almost everywhere*.

In the statement of theorem 2.1, "essentially uniquely determined" means that any two such functions are equal almost everywhere.

2.2. Maximum likelihood estimation

We first introduce the theoretical and empirical inner products that will be used later on to define the ANOVA decompositions of functions. The corresponding norms will be used to measure the distance between the estimator and the target function.

Given a non-absorbing state w, set $\gamma_{\nu w} = \text{ind}\,(T_{\nu-1} < \infty \text{ and } Y(T_{\nu-1}) = w)$. The corresponding theoretical inner product and norm for functions on $\mathscr{T} \times \mathscr{X}$ are defined by

$$\langle a_1, a_2 \rangle_w = E \sum_\nu \gamma_{\nu w} \int_{T_{\nu-1} \wedge C}^{T_\nu \wedge C} a_1(t - T_{\nu-1}, \mathbf{X}(t)) a_2(t - T_{\nu-1}, \mathbf{X}(t))\, dt$$

and $\|a\|_w^2 = \langle a, a \rangle_w$ for square-integrable functions a_1, a_2, a on $\mathscr{T} \times \mathscr{X}$. Given $a = (a_{wy})$, where each a_{wy} is square-integrable, set $\|\mathbf{a}\|^2 = \sum_{w,y} \|a_{wy}\|_w^2$. Similarly, let $\langle \cdot, \cdot \rangle_{nw}$ and $\|\cdot\|_{nw}^2$ denote the empirical inner product and squared norm corresponding to the random sample of size n defined by

$$\langle a_1, a_2 \rangle_{nw} = \frac{1}{n} \sum_i \sum_\nu \gamma_{i\nu w} \int_{T_{i,\nu-1} \wedge C_i}^{T_{i\nu} \wedge C_i} a_1(t - T_{i,\nu-1}, \mathbf{X}_i(t)) a_2(t - T_{i,\nu-1}, \mathbf{X}_i(t))\, dt$$

and $\|a\|_{nw}^2 = \langle a, a \rangle_{nw}$, where $\gamma_{i\nu w} = \text{ind}\,(T_{i,\nu-1} < \infty \text{ and } Y_i(T_{i,\nu-1}) = w)$. For any vector of functions $\mathbf{a} = (a_{wy})$, set $\|\mathbf{a}\|_n^2 = \sum_{w,y} \|a_{wy}\|_{nw}^2$.

Let $|\mathscr{X}|$ denote the Lebesgue measure of $\mathscr{X}$. For any square-integrable functions a_1, a_2, a on $\mathscr{T} \times \mathscr{X}$, define the inner product and norm induced by Lebesgue measure as $\langle a_1, a_2 \rangle_{L_2} = (\int_{\mathscr{T}} \int_{\mathscr{X}} a_1(t, \mathbf{x}) a_2(t, \mathbf{x})\, d\mathbf{x}\, dt / (\tau |\mathscr{X}|))^{1/2}$ and $\|a\|_{L_2}^2 = \langle a, a \rangle_{L_2}$. According to lemma 3.1, $\|\cdot\|_{L_2}$ is equivalent to the theoretical norm $\|\cdot\|_w$.

Next, we present a general theorem on rates of convergence. Let $G = G_n \subset H$ be a finite-dimensional linear space of bounded functions on $\mathscr{X}$ having dimension $N_n \geqslant 1$. We require that the space G to be *theoretically identifiable* in that, if $g \in G$ and $\sum_w \|g\|_w = 0$, then it identically equals zero. Since we hope to choose G such that the functions in H can be well approximated by functions in G, we refer to G as the approximating space. The space G is said to be *empirically identifiable* if the only function g in the space such that $\sum_w \|g\|_{nw} = 0$ is the function that identically equals zero.

The log-likelihood function, restricted to vectors whose constituents are in G, is concave. Suppose G is identifiable. Then this log-likelihood is strictly concave; thus, there is at most one maximum-likelihood estimate corresponding to G. According to lemma 5.5, the maximum-likelihood estimate $\hat{\boldsymbol{\alpha}}_n$ exists except on an event whose probability tends to zero as $n \to \infty$.

Set $A_n = \sup_{g \in G} \{\|g\|_\infty / \|g\|_{L_2}\}$. Set $\rho_{wy} = \inf_{g \in G} \|g - \alpha_{wy}^*\|_\infty$ and $\rho_n = \sum_{w,y} \rho_{wy}$.

Theorem 2.2

Suppose conditions 1 *and* 2 *hold and that* $\lim_n A_n^2 N_n / n = 0$ *and* $\lim_n A_n \rho_n = 0$. *Then* $\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^*\|^2 = O_P(\rho_n^2 + N_n/n)$ *and* $\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^*\|_n^2 = O_P\big(\rho_n^2 + N_n/n\big)$.

2.3. Functional ANOVA models

Suppose $\mathscr{X}$ is the Cartesian product of compact sets $\mathscr{X}_1, \ldots, \mathscr{X}_L$. For $\mathbf{x} \in \mathscr{X}$, write $\mathbf{x} = (x_1, \ldots, x_L)$, where $x_l \in \mathscr{X}_l$ for $1 \leqslant l \leqslant L$. Given a non-empty subset s of $\{0, 1, \ldots, L\}$, let H_s denote the space of functions on $[0, \infty) \times \mathscr{X}$ that depend on the variable t if $0 \in s$ and on the variable x_l for $l \in s \cap \{1, \ldots, L\}$ and on no other variables. Let H_0 denote the space of constant functions on $[0, \infty) \times \mathscr{X}$. Let $\mathscr{S}$ be a non-empty collection of subsets of $\{0, 1, \ldots, L\}$. It is assumed that $\mathscr{S}$ is *hierarchical*: if $s \in \mathscr{S}$ and $r \subset s$, then $r \in \mathscr{S}$. Set $H = \{\sum_{s \in \mathscr{S}} a_s : a_s \in H_s\}$. By lem. 4.1 of Stone (1994), the space H satisfies condition 2. It follows from theorem 2.1 that if condition 1 holds, then the best approximation $\boldsymbol{\alpha}^*$ to $\boldsymbol{\alpha}$ always exists.

To introduce the notion of ANOVA decompositions, it is necessary to restrict our attention to square-integrable functions. For $s \in \mathscr{S}$, let H_s^2 denote the space of square-integrable functions

in H_s. Let w be an arbitrary non-absorbing state. For $a \in H_s^2$, let $a \perp_w H_r^2$ mean that $\langle a, a_r \rangle_w = 0$ for $a_r \in H_r^2$. Set $H^2 = \{\sum_{s \in \mathscr{S}} a_s : a_s \in H_s^2\}$. It follows from lemma 3.3 that, under condition 1, every function $a \in H^2$ can be written in an essentially unique manner as $\sum_{s \in \mathscr{S}} a_{ws}$, where $a_{ws} \in H_s^2$ and $a_{ws} \perp_w H_r^2$ for every proper subset r of s. We refer to $\sum_{s \in \mathscr{S}} a_{ws}$ as the *theoretical ANOVA decomposition* of a corresponding to the inner product $\langle \cdot, \cdot \rangle_w$.

We now construct the approximating space G. Let $G_\varnothing$ denote the space of constant functions on $\mathscr{T} \times \mathscr{X}$, which has dimension $N_\varnothing = 1$. Given $l \in \{0, 1, \ldots, L\}$, let $G_l \subset H_l$ denote a linear space of bounded functions, which varies with sample size and has finite, positive dimension N_l. Given a subset s of $\{0, 1, \ldots, L\}$, let G_s be the tensor product of G_l, $l \in s$, which is the space of functions on $\mathscr{T} \times \mathscr{X}$ spanned by functions g of the form $g(t, \mathbf{x}) = \prod_{l \in s} g_l(x_l)$, where $x_0 = t$ and $g_l \in G_l$ for $l \in s$. Then the dimension N_s of G_s is $\prod_{l \in s} N_l$. Set $G = \{\sum_{s \in \mathscr{S}} g_s : g_s \in G_s\}$. Then the dimension N_n of G satisfies $\max_{s \in \mathscr{S}} N_s \leqslant N_n \leqslant \sum_{s \in \mathscr{S}} N_s \leqslant \#(\mathscr{S}) \max_{s \in \mathscr{S}} N_s$. Suppose G is empirically identifiable and let g be a member of this space. As a consequence of lemma 3.4, g can be written uniquely in the form $\sum_{s \in \mathscr{S}} g_{ws}$, where $g_{ws} \in G_s$ and $g_{ws} \perp_{nw} G_r$ for every proper subset r of s. We refer to $\sum_{s \in \mathscr{S}} g_{ws}$ as the *empirical ANOVA decomposition* of g corresponding to the inner product $\langle \cdot, \cdot \rangle_{nw}$.

Set $A_s = \sup_{g \in G_s} \{\|g\|_\infty / \|g\|_{L_2}\}$ for $s \in \mathscr{S}$. Suppose the constituents of $\boldsymbol{\alpha}^*$ are members of H^2 and let $\sum_{s \in \mathscr{S}} \alpha^*_{wys}$ be the ANOVA decomposition of α^*_{wy} corresponding to the inner product $\langle \cdot, \cdot \rangle_w$. Set $\rho_{wys} = \inf_{g \in G_s} \|g - \alpha^*_{wys}\|_\infty$ for $s \in \mathscr{S}$ and $\overline{\rho}_n = \sum_{w,y} \sum_s \rho_{wys}$.

Theorem 2.3
Suppose conditions 1 *and* 2 *hold and that* $\lim_n A_s \rho_{wys'} = 0$ *and* $\lim_n A_s^2 N_{s'}/n = 0$ *for each pair* $s, s' \in \mathscr{S}$ *and each pair* w, y. *Then the results of theorem* 1 *hold with* ρ_n *replaced by* $\overline{\rho}_n$.

Proof. Since $A_n \leqslant (\sum_{s \in \mathscr{S}} A_s^2)^{1/2}$, $N_n \asymp \sum_{s \in \mathscr{S}} N_s$, and $\rho_{wy} \leqslant \sum_{s \in \mathscr{S}} \rho_{wys}$, the desired results follow from theorem 2.2.

The ANOVA decomposition of $\hat{\alpha}_{nwy}$ has the form $\hat{\alpha}_{nwy} = \sum_{s \in \mathscr{S}} \hat{\alpha}_{wys}$, where $\hat{\alpha}_{wys} \in G_s$ and $\hat{\alpha}_{wys} \perp_{nw} G_r$ for every proper subset r of s. The next result demonstrates that $\hat{\alpha}_{wys}$ provides consistent estimate of α^*_{wys} for each $s \in \mathscr{S}$.

Theorem 2.4
Suppose conditions 1 *and* 2 *hold and that* $\lim_n A_s^2 N_{s'}/n = 0$ *and* $\lim_n A_s \rho_{wys'} = 0$ *for each pair* $s, s' \in \mathscr{S}$ *and each pair* w, y. *Then, for each* $s \in \mathscr{S}$, $\|\hat{\alpha}_{wys} - \alpha^*_{wys}\|_w^2 = O_P(\overline{\rho}^2 + N_n/n)$ *and* $\|\hat{\alpha}_{wys} - \alpha^*_{wys}\|_{nw}^2 = O_P(\overline{\rho}_n^2 + N_n/n)$.

This result is parallel to th. 3 of Huang (1998) for regression. We can obtain rate of convergence results when polynomials, trigonometric polynomials, splines, or wavelets and their tensor products are used as building blocks for the approximating spaces. To get such results, we need only find upper bounds for the constants A_s and ρ_{wys} by employing results from approximation theory literature.

Suppose $\mathscr{X}_l \subset \mathbb{R}^{d_l}$ with $d_l \geqslant 1$. Set $d = \max_{s \in \mathscr{S}} \sum_{l \in s} d_l$. If $d_l = 1$ for $1 \leqslant l \leqslant L$, then $d = \max_{s \in \mathscr{S}} \#(s)$, where $\#(s)$ denotes the number of members of s. Typically, the spaces G_l are chosen such that $\overline{\rho}_n \asymp N_n^{-p/d}$ and $N_n \asymp n^{d/(2p+d)}$, where p is a suitable defined measure of smoothness of α^*_{wys}. Correspondingly, the rate of convergence in the theorem has the order $\overline{\rho}^2 + N_n/n = O(n^{-2p/(2p+d)})$, which is of the standard form; see Stone (1982, 1994), and Huang (1998).

Example (univariate splines). Assume that $\mathscr{X}$ is the Cartesian product of compact intervals $\mathscr{X}_1, \ldots, \mathscr{X}_L$ in $\mathbb{R}$. Without loss of generality, it is assumed that each of these intervals equals [0, 1]. Let $k \geqslant 0$ be an integer and $0 < \beta \leqslant 1$. Set $p = k + \beta$. As in Stone (1994), a function on $\mathscr{X}$ is said to be *p-smooth* if it is k times continuously differentiable on $\mathscr{X}$ and its partial derivatives of total order k satisfy a Hölder condition with exponent β. We assume that conditions 1 and 2 hold. In addition, we assume that α^*_{wys} is p-smooth for each $s \in \mathscr{S}$ and each pair w, y. Set $d = \max_{s \in \mathscr{S}} \#(s)$.

Let $m \geqslant p - 1$ be an integer. For $l = 1, \ldots, L$, let G_l be the space of splines of degree m with $J = J_n$ equally spaced interior knots. Then $A_s \asymp J^{\#(s)}$, $N_s \asymp J^{\#(s)}$ and $\rho_{wys} \asymp J^{-p}$; see Huang (1998). Suppose $p > d/2$ and $J^{2d} = o(n)$. Then the conditions in theorems 2 and 3 are satisfied. Thus, $\|\hat{\alpha}_{wys} - \alpha^*_{wy}\|^2_w = O_P(J^d/n + J^{-2p})$ for $s \in \mathscr{S}$ and $\|\hat{\alpha}_{wy} - \alpha^*_{wy}\|^2_w = O_P(J^d/n + J^{-2p})$. Taking $J \asymp n^{1/(2p+d)}$, we get that $\|\hat{\alpha}_{wys} - \alpha^*_{wys}\|^2_w = O_P(n^{-2p/(2p+d)})$ for $s \in \mathscr{S}$ and $\|\hat{\alpha}_{wy} - \alpha^*_{wy}\|^2_w = O_P(n^{-2p/(2p+d)})$. The rate $n^{-2p/(2p+d)}$ is the optimal rate for estimating a p-smooth, d-dimensional function; see Stone (1982). This result generalizes that of Kooperberg *et al.* (1995b), where there are only two states, one and only one of which is absorbing, and the covariates do not depend on time.

The result from the previous example tells us that the rates of convergence are determined by the smoothness of the ANOVA components of α^*_{wy} and the highest order of interactions included in the model. It also demonstrates that, by using models with only main effects and low-order interactions, we can ameliorate the curse of dimensionality that the saturated model suffers. In particular, by considering additive models ($d = 1$) or by allowing interactions involving only two factors ($d = 2$), we can get faster rates of convergence than by using the saturated model ($d = L$).

Using univariate functions and their tensor products to model α^*_{wy} restricts the domain of α^*_{wy} to be a hyper-rectangle. By allowing bivariate or multivariate functions and their tensor products to model α^*_{wy}, we gain flexibility, especially when some predictor variable is of spatial type. Our theorems also apply to these cases when the approximating spaces are built with multivariate splines and their tensor products or more general finite element spaces and their tensor products. We can proceed as in Huang (1998) to check the conditions of theorems 2.3 and 2.4 and get the rate of convergence results. Similar results for other statistical contexts, such as regression, are available in Hansen (1994) and Huang (1998). Recently, Hansen *et al.* (1996) developed the Triogram method for function estimation using piecewise planar, bivariate splines based on adaptively constructed triangulation. Hopefully, their method can be used to model interactions of two-dimensional components in the context of event history analysis.

3. Preliminaries

Let $\rho_v(\cdot, y|w, \mathbf{X})$ denote the density function of

$$P(T_v \in \cdot,\ Y(T_v) = y | Y(0) = w, \mathbf{X}).$$

Then

$$\int_B \rho_v(t, y|w, \mathbf{X})\, dt = P(T_v \in B,\ Y(T_v) = y | Y(0) = w, \mathbf{X}), \quad B \subset \mathscr{T}.$$

In particular,

$$\int_0^\tau \rho_v(t, y|w, \mathbf{X})\, dt = P(T_v \leqslant \tau,\ Y(T_v) = y | Y(0) = w, \mathbf{X}).$$

Now

$$\rho_1(t,\, y|w,\, \mathbf{X}) = \lambda_{wy}(t,\, \mathbf{X}(t))\exp\left(-\int_0^t \lambda_w(u,\, \mathbf{X}(u))\, du\right);$$

by condition 1, this is bounded away from zero and infinity uniformly over $t \in \mathscr{T}$, $\mathbf{X}$, and states w, y (where a direct transition from w to y is possible).

Let $\nu \geqslant 2$. It follows from the semi-Markov property that

$$\rho_\nu(t,\, y|w,\, \mathbf{X}) = \sum_v \int_0^t \rho_{\nu-1}(s,\, v|w,\, \mathbf{X})\, \lambda_{vy}(t-s,\, \mathbf{X}(t)) \exp\left(-\int_0^{t-s} \lambda_v(u,\, \mathbf{X}(s+u))\, du\right) ds. \tag{3.1}$$

Consequently,

$$\sum_y \rho_\nu(t,\, y|w,\, \mathbf{X}) = \sum_v \int_0^t \rho_{\nu-1}(s,\, v|w,\, \mathbf{X})\, \lambda_v(t-s,\, \mathbf{X}(t)) \exp\left(-\int_0^{t-s} \lambda_v(u,\, \mathbf{X}(s+u))du\right) ds,$$

so

$$\int_0^\tau \sum_y \rho_\nu(t,\, y|w,\, \mathbf{X})\, dt = \sum_v \int_0^\tau \rho_{\nu-1}(s,\, v|w,\, \mathbf{X}) \left[1 - \exp\left(-\int_0^{\tau-s} \lambda_v(u,\, \mathbf{X}(s+u))\, du\right)\right] ds.$$

Thus, by condition 1, there is a fixed number ϵ with $0 < \epsilon < 1$ such that

$$\int_0^\tau \sum_y \rho_\nu(t,\, y|w,\, \mathbf{X})\, dt \leqslant (1-\epsilon) \int_0^\tau \sum_y \rho_{\nu-1}(t,\, y|w,\, \mathbf{X})\, dt.$$

Therefore,

$$\int_0^\tau \sum_y \rho_\nu(t,\, y|w,\, \mathbf{X})\, dt \leqslant (1-\epsilon)^{\nu-1} \int_0^\tau \sum_y \rho_1(t,\, y|w,\, \mathbf{X})\, dt, \quad \nu \geqslant 1.$$

Thus, by (3.1) and another application of condition 1 (for $\nu \geqslant 2$), there is a fixed positive number M_1 such that

$$\sum_y \rho_\nu(t,\, y|w,\, \mathbf{X}) \leqslant M_1(1-\epsilon)^\nu, \quad \nu \geqslant 1 \text{ and } t \in \mathscr{T}.$$

Integrating with respect to t, we get that

$$P(T_\nu \leqslant \tau | Y(0) = w,\, \mathbf{X}) \leqslant M_1 \tau (1-\epsilon)^\nu, \quad \nu \geqslant 1.$$

Hence we have proved the following result.

Proposition 3.1
Suppose condition 1 holds. Then there is a fixed number ϵ with $0 < \epsilon < 1$ and a fixed positive number M_1 such that

$$\sum_y \rho_\nu(t,\, y|w,\, \mathbf{X}) \leqslant M_1(1-\epsilon)^\nu, \quad \nu \geqslant 1 \text{ and } t \in \mathscr{T}.$$

As a consequence,

$$P(T_\nu \leqslant \tau | Y(0) = w,\, \mathbf{X}) \leqslant M_1 \tau (1-\epsilon)^\nu, \quad \nu \geqslant 1.$$

Recall that $\delta_\nu = \text{ind}\,(T_\nu \leqslant C \text{ and } T_\nu < \infty)$ and $\gamma_{\nu w} = \text{ind}\,(T_{\nu-1} < \infty \text{ and } Y(T_{\nu-1}) = w)$. Set $\eta_{\nu y} = \text{ind}\,(T_\nu < \infty \text{ and } Y(T_\nu) = y)$. Let $f_{\mathbf{X}(t)}$ denote the density function of $\mathbf{X}(t)$. Set

$$B_{1w}(t, \mathbf{x}) = f_{\mathbf{X}(t)}(\mathbf{x}) E\left[P(Y(0) = w|\mathbf{X}) P(C \geqslant t|\mathbf{X}) \exp\left(-\int_0^t \lambda_w(u, \mathbf{X}(u))\, du \right) \middle| \mathbf{X}(t) = \mathbf{x} \right]$$

and, for $v \geqslant 2$, set

$$B_{vw}(t, \mathbf{x}) = \int_0^{\tau - t} f_{\mathbf{X}(t+s)}(x) \sum_v E\Big[P(Y(0) = v|\mathbf{X}) P(C \geqslant t + s|\mathbf{X})$$
$$\rho_{v-1}(s, w|v, \mathbf{X}) \exp\left(-\int_0^t \lambda_w(u, \mathbf{X}(u + s))\, du \right) \Big| \mathbf{X}(t + s) = \mathbf{x} \Big]\, ds.$$

Then we have the following result.

Proposition 3.2.
For any square-integrable functions a, a_1 and a_2 on $\mathscr{T} \times \mathscr{X}^{\cdot}$,

$$E[\delta_v \lambda_{vw} \eta_{vy} a(T_v - T_{v-1}, \mathbf{X}(T_v))] = \int_{\mathscr{T}} \int_{\mathscr{X}^{\cdot}} a(t, \mathbf{x}) \lambda_{wy}(t, \mathbf{x}) B_{vw}(t, \mathbf{x})\, d\mathbf{x}\, dt$$

and

$$E\left[\lambda_{vw} \int_{T_{v-1} \wedge C}^{T_v \wedge C} a_1(t - T_{v-1}, \mathbf{X}(t)) a_2(t - T_{v-1}, \mathbf{X}(t))\, dt \right]$$
$$= \int_{\mathscr{T}} \int_{\mathscr{X}^{\cdot}} a_1(t, \mathbf{x}) a_2(t, \mathbf{x}) B_{vw}(t, \mathbf{x})\, d\mathbf{x}\, dt.$$

Proof. The proof is deferred to the appendix.

Throughout the remaining part of this section, let w be a non-absorbing state. The next lemma shows that the norm $\|\cdot\|_{L_2}$ is equivalent to the theoretical norm $\|\cdot\|_w$ for any such a w.

Lemma 3.1
Suppose condition 1 *holds. Then there are positive constants M_2 and M_3 such that for any square-integrable function f on $\mathscr{T} \times \mathscr{X}^{\cdot}$,*

$$M_2 \|f\|_{L_2} \leqslant \|f\|_w \leqslant M_3 \|f\|_{L_2}.$$

Proof. By proposition 3.2, $\|f\|_w^2 = \int_{\mathscr{T}} \int_{\mathscr{X}^{\cdot}} f^2(t, \mathbf{x}) B_w(t, \mathbf{x})\, d\mathbf{x}\, dt$, where $B_w(t, \mathbf{x}) = \sum_v B_{vw}(t, \mathbf{x})$. It follows from condition 1 and propositions 3.1 that $B_w(t, \mathbf{x})$ is bounded away from zero and infinity on $\mathscr{T} \times \mathscr{X}^{\cdot}$.

The following lemma tells us that the empirical inner product is uniformly close to the theoretical inner product on the approximating space G. As a consequence, the empirical and theoretical norms are equivalent over G.

Lemma 3.2
Suppose $\lim_n A_n^2 N_n / n = 0$ and let $t > 0$. Then, except on an event whose probability tends to zero as $n \to \infty$,

$$|\langle f, g \rangle_{nw} - \langle f, g \rangle_w| \leqslant t \|f\|_w \|g\|_w, \quad f, g \in G.$$

Consequently, except on an event whose probability tends to zero as $n \to \infty$,

$$\tfrac{1}{2} \|g\|_w^2 \leqslant \|g\|_{nw}^2 \leqslant 2 \|g\|_w^2, \quad g \in G.$$

Proof. Given functions a_1, a_2 on $\mathscr{T} \times \mathscr{X}$. Set

$$\mathbf{\Psi}_w(a_1, a_2, \mathbf{X}, Y) = \sum_\nu \lambda_{\nu w} \int_{T_{\nu-1}\wedge C}^{T_\nu \wedge C} a_1(t - T_{\nu-1}, \mathbf{X}(t)) a_2(t - T_{\nu-1}, \mathbf{X}(t))\, dt.$$

Then $\mathbf{\Psi}_w(a_1, a_2, \mathbf{X}, Y)$ is bilinear in a_1 and a_2,

$$\langle a_1, a_2 \rangle_{nw} = \frac{1}{n} \sum_i \mathbf{\Psi}_w(a_1, a_2, \mathbf{X}_i, Y_i),$$

and $\langle a_1, a_2 \rangle_w = E[\mathbf{\Psi}_w(a_1, a_2, \mathbf{X}, Y)]$. Now $|\mathbf{\Psi}_w(a_1, a_2, \mathbf{X}, Y)| \leqslant \tau \|a_1\|_\infty \|a_2\|_\infty$ since $P(C \leqslant \tau) = 1$. It is shown in the appendix that

$$\operatorname{var}[\mathbf{\Psi}_w(a_1, a_2, \mathbf{X}, Y)] \leqslant \tau \|a_1\|_w^2 \|a_2\|_\infty^2. \tag{3.2}$$

Thus, the desired results follow from lem. 10 of Huang (1998).

Corollary 3.1
Suppose $\lim_n A_n^2 N_n / n = 0$. *Then, except on an event whose probability tends to zero as* $n \to \infty$, *G is empirically identifiable.*

Let M_2, M_3 be defined as in lemma 3.1. Set $\epsilon_1 = 1 - \sqrt{1 - (M_2/M_3)^2}$.

Lemma 3.3
Suppose condition 1 *holds. Then* $\|a\|_w^2 \geqslant \epsilon_1^{\#(\mathscr{S})-1} \sum_{s\in\mathscr{S}} \|a_{ws}\|_w^2$ *for all* $a = \sum_{s\in\mathscr{S}} a_{ws}$, *where* $a_{ws} \in H_s$ *and* $a_{ws} \perp_w H_r^2$ *for every proper subset* r of s.

Proof. This lemma can be proved by using lemma 3.1 and arguing as in the proof of lem. 3.1 of Stone (1994).

The previous lemma reveals that the components of the theoretical ANOVA decomposition are not too confounded. As a consequence, each function in H^2 has a unique ANOVA decomposition with respect to norm $\|\cdot\|_w$.

According to the next result, the components of the empirical ANOVA decomposition are not too confounded, either empirically or theoretically.

Lemma 3.4
If G is empirically identifiable, then each $g \in G$ *has a unique representation* $g = \sum_{s\in\mathscr{S}} g_{ws}$, *where* $g_{ws} \in G_s$ and $g_{ws} \perp_{nw} G_r$ *for every proper subset r of s. Suppose condition* 1 *holds and that* $\lim_n A_n^2 N_n / n = 0$. *Let* $0 < \epsilon_2 < \epsilon_1$. *Then, except on an event whose probability tends to zero as* $n \to \infty$, $\|g\|_w^2 \geqslant \epsilon_2^{\#(\mathscr{S})-1} \sum_{s\in\mathscr{S}} \|g_{ws}\|_w^2$ and $\|g\|_{nw}^2 \geqslant \epsilon_2^{\#(\mathscr{S})-1} \sum_{s\in\mathscr{S}} \|g_{ws}\|_{nw}^2$ *for all* $g = \sum_{s\in\mathscr{S}} g_{ws}$, *where* $g_{ws} \in G_s$ *and* $g_{ws} \perp_{nw} G_r$ *for every proper subset r of s.*

Proof. The first part of the lemma was proved in lem. 3.2 of Stone (1994). The second part of the lemma can be proved by using lemmas 3.1 and 3.2 and arguing as in the proof of lem. 3.1 of Stone (1994).

4. Proof of theorem 2.1

By proposition 3.2,

$$\Lambda(\mathbf{a}) = \sum_{w,y} \int_{\mathscr{T}} \int_{\mathscr{X}} [a_{wy}(t, \mathbf{x}) \lambda_{wy}(t, \mathbf{x}) - \exp\{a_{wy}(t, \mathbf{x})\}] \sum_\nu B_{\nu w}(t, \mathbf{x})\, d\mathbf{x}\, dt. \tag{4.1}$$

By condition 1, the functions λ_{wy} are bounded away from zero and infinity. Thus there are positive numbers A_1 and A_2 such that

$$a_{wy}\lambda_{wy} - \exp(a_{wy}) \leqslant A_1 - A_2|a_{wy}|.$$

Hence, $\Lambda(\mathbf{a}) \leqslant A_1 \sum_{wy} \int_{\mathscr{T}} \int_{\mathscr{X}} \sum_{v} B_{vw} - A_2 \sum_{wy} \int_{\mathscr{T}} \int_{\mathscr{X}} |a_{wy}| \sum_{v} B_{vw}$. Now by condition 1 and proposition 3.1, $\sum_{v} B_{vw}$ is bounded away from zero and infinity. Consequently, if $\int_{\mathscr{T}} \int_{\mathscr{X}} |a_{wy}| = \infty$ for some w, y, then $\Lambda(\mathbf{a}) = -\infty$. Moreover, the function $\Lambda(\cdot)$ is bounded above by some constant A_3. Therefore, the numbers $\Lambda(\mathbf{a})$, $\mathbf{a} = (a_{wy})$, $a_{wy} \in H$, have a finite least upper bound L. Choose $\mathbf{a}_k = (a_{kwy})$, $a_{kwy} \in H$, such that $\Lambda(\mathbf{a}_k) > -\infty$ and $\Lambda(\mathbf{a}_k) \to L$ as $k \to \infty$. Observe that the numbers $\int_{\mathscr{T}} \int_{\mathscr{X}} |a_{kwy}|$, $k \geqslant 1$, are bounded.

Let $\boldsymbol{\alpha}_1 = (\alpha_{1wy})$ and $\boldsymbol{\alpha}_2 = (\alpha_{2wy})$ be vectors of functions with α_{1wy}, $\alpha_{2wy} \in H$ such that $\Lambda(\boldsymbol{\alpha}_1) > -\infty$ and $\Lambda(\boldsymbol{\alpha}_2) > -\infty$. For $u \in [0, 1]$, set $\boldsymbol{\alpha}^{(u)} = (1-u)\boldsymbol{\alpha}_1 + u\boldsymbol{\alpha}_2$ and $\boldsymbol{\Phi}(u) = \Lambda(\boldsymbol{\alpha}^{(u)})$. Then $\boldsymbol{\Phi}(\cdot)$ is a concave function. It follows from the argument of th. 4.1 in Stone (1994) that there is a vector of integrable functions $\boldsymbol{\alpha}^* = (\alpha^*_{wy})$ such that $a_{kwy} \to a^*_{wy}$ in measure as $k \to \infty$. Since H satisfies condition 2, we can assume that $\alpha^*_{wy} \in H$. It follows from Fatou's Lemma that $\Lambda(\mathbf{a}_k) \to \Lambda(\boldsymbol{\alpha}^*) = L = \max_{\mathbf{a} \in H} \Lambda(\mathbf{a})$ as $k \to \infty$. Furthermore, if $a_{wy} \in H$ and $\Lambda(\mathbf{a}) = \Lambda(\boldsymbol{\alpha}^*)$, then it follows from the concavity described above that $\mathbf{a} = \boldsymbol{\alpha}^*$ almost everywhere. Hence the first statement of the theorem is valid. The second statement follows from (4.1) and the fact that $a\lambda - e^a$, as a function of a, has a unique maximum at $a = \log \lambda$.

5. Proof of theorem 2.2

When it exists, we refer to $\boldsymbol{\alpha}^*_n = \arg\max\{\Lambda(\mathbf{g})\colon \mathbf{g} = (g_{wy}),\ g_{wy} \in G\}$ as the best approximation to $\boldsymbol{\alpha}$ with constituents in G. We have the decomposition $\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^* = (\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^*_n) + (\boldsymbol{\alpha}^*_n - \boldsymbol{\alpha}^*)$, where the first term on the right side of the equation is referred to as the approximation error and the second term as the estimation error. In lemmas 5.2–5.5, we assume the conditions in theorem 2.2 hold.

5.1. Approximation error

Lemma 5.1

Suppose condition 1 *holds and* $\|\boldsymbol{\alpha}^*\|_\infty < \infty$. *Let* U *be a positive constant. Then there are positive constants* M_4 *and* M_5 *such that*

$$-M_4\|\mathbf{a} - \boldsymbol{\alpha}^*\|^2 \leqslant \Lambda(\mathbf{a}) - \Lambda(\boldsymbol{\alpha}^*) \leqslant -M_5\|\mathbf{a} - \boldsymbol{\alpha}^*\|^2$$

for all $\mathbf{a} = (a_{wy})$ *with* $a_{wy} \in H$ *and* $\|\mathbf{a}\|_\infty \leqslant U$.

Proof. Given $\mathbf{a} = (a_{wy})$ with $a_{wy} \in H$ and $\|\mathbf{a}\|_\infty \leqslant U$ and given $u \in [0, 1]$, set $a^{(u)}_{wy} = (1-u)\alpha^*_{wy} + ua_{wy}$. Then $d\Lambda(\mathbf{a}^{(u)})/du|_{u=0} = 0$ and hence, by integration by parts,

$$\Lambda(\mathbf{a}) - \Lambda(\boldsymbol{\alpha}^*) = \int_0^1 (1-u)\frac{d^2}{du^2}\Lambda(\mathbf{a}^{(u)})\, du.$$

Observe that

$$\frac{d^2}{du^2}\Lambda(\mathbf{a}^{(u)}) = -\sum_{w,y} E\sum_{v} \lambda_{vw}\int_{T_{v-1}\wedge C}^{T_v \wedge C} [a_{wy} - \alpha^*_{wy}]^2(t - T_{v-1}, \mathbf{X}(t))$$

$$\times \exp\{[(1-u)\alpha^*_{wy} + ua_{wy}](t - T_{v-1}, \mathbf{X}(t))\}\, du.$$

The desired result now follows from condition 1 and the definition of $\|\cdot\|$.

Lemma 5.2
The best approximation $\boldsymbol{\alpha}_n^*$ to $\boldsymbol{\alpha}$ with constituents in G exists for n sufficiently large and satisfies $\|\boldsymbol{\alpha}_n^* - \boldsymbol{\alpha}^*\|^2 = O(\rho_n^2)$ and $\|\boldsymbol{\alpha}_n^* - \boldsymbol{\alpha}^*\|_n^2 = O_P(\rho_n^2)$. Moreover, for any positive number $U > \|\boldsymbol{\alpha}^*\|_\infty$, $\|\boldsymbol{\alpha}_n^*\|_\infty \leqslant U$ holds for n sufficiently large,

Proof. Since α_{wy}^* is bounded, by a compactness argument, there is a function $g_{wy}^* \in G$ such that $\|g_{wy}^* - \alpha_{wy}^*\|_\infty = \rho_{wy}$. Write $\mathbf{g}^* = (g_{wy}^*)$. Then $\|\mathbf{g}^* - \boldsymbol{\alpha}^*\| \leqslant \rho_n$ and, for n sufficiently large,

$$\|\mathbf{g}^*\|_\infty \leqslant \|\mathbf{g}^* - \boldsymbol{\alpha}^*\|_\infty + \|\boldsymbol{\alpha}^*\|_\infty \leqslant \rho_n + \|\boldsymbol{\alpha}^*\|_\infty \leqslant 1 + \|\boldsymbol{\alpha}^*\|_\infty.$$

Let c denote a positive constant (to be determined later). Choose $\mathbf{g} = (g_{wy})$ with $g_{wy} \in G$ and $\|\mathbf{g} - \boldsymbol{\alpha}^*\| = c\rho_n$. Then, by lemma 3.1 and the triangle inequality,

$$\|\mathbf{g} - \mathbf{g}^*\|_\infty \leqslant A_n M_2^{-1} \|\mathbf{g} - \mathbf{g}^*\| \leqslant A_n M_2^{-1}(c\rho_n + \rho_n).$$

Thus, for n sufficiently large,

$$\|\mathbf{g}\|_\infty \leqslant A_n M_2^{-1}(c\rho_n + \rho_n) + \rho_n + \|\boldsymbol{\alpha}^*\|_\infty \leqslant 1 + \|\alpha^*\|_\infty.$$

Now applying lemma 5.1 with $U = 1 + \|\boldsymbol{\alpha}^*\|_\infty$, we get that, for n sufficiently large,

$$\Lambda(\mathbf{g}^*) - \Lambda(\boldsymbol{\alpha}^*) \geqslant -M_4 \rho_n^2 \tag{5.1}$$

and

$$\Lambda(\mathbf{g}) - \Lambda(\boldsymbol{\alpha}^*) \leqslant -M_5 c^2 \rho_n^2 \tag{5.2}$$

for all $\mathbf{g} = (g_{wy})$ with $g_{wy} \in G$ and $\|\mathbf{g} - \boldsymbol{\alpha}^*\| = c\rho_n$. Let c be chosen such that $c > \sqrt{M_4/M_5}$. Then $\|\mathbf{g}^* - \boldsymbol{\alpha}^*\| < c\rho_n$, and it follows from (5.1) and (5.2) that, for n sufficiently large,

$$\Lambda(\mathbf{g}) < \Lambda(\mathbf{g}^*) \quad \text{for all } \mathbf{g} = (g_{wy}) \text{ with } g_{wy} \in G \text{ and } \|\mathbf{g} - \boldsymbol{\alpha}^*\| = c\rho_n.$$

Therefore, by the definition of $\boldsymbol{\alpha}_n^*$ and the concavity of $\Lambda(\mathbf{g})$ as a function of $\mathbf{g}$, $\boldsymbol{\alpha}_n^*$ exists and satisfies $\|\boldsymbol{\alpha}_n^* - \boldsymbol{\alpha}^*\| < c\rho_n$ for n sufficiently large. Thus, $\|\boldsymbol{\alpha}_n^* - \boldsymbol{\alpha}^*\|^2 = O(\rho_n^2)$. This result together with the triangle inequality and lemma 3.2 implies that $\|\boldsymbol{\alpha}_n^* - \boldsymbol{\alpha}^*\|_n^2 = O_P(\rho_n^2)$. The last part of the lemma follows from the first part, $\lim_n A_n \rho_n = 0$, and the inequality $\|\boldsymbol{\alpha}_n^*\|_\infty \leqslant \|\boldsymbol{\alpha}^*\|_\infty + A_n \|\boldsymbol{\alpha}_n^* - g^*\| + \|\boldsymbol{\alpha}^* - g^*\|_\infty$.

5.2. Estimation error

Let $\{\phi_{wj}, 1 \leqslant j \leqslant N_n\}$ be an orthonormal basis of G relative to the theoretical inner product $\|\cdot\|_w$. Then each $g \in G$ can be represented uniquely as $g = \sum_j \beta_{wj}\phi_{wj}$, where $\beta_{wj} = \langle g, \phi_{wj}\rangle_w$ for $j = 1, \ldots, N_n$. Let $\boldsymbol{\beta}_w$ denote the N_n-dimensional vector with entries β_{wj}. To indicate the dependence of g on $\boldsymbol{\beta}_w$, we write $g(\cdot, \cdot) = g(\cdot, \cdot; \boldsymbol{\beta}_w)$. Let $|\cdot|$ denote the Euclidean norm of vectors. Then $\|g(\cdot, \cdot; \boldsymbol{\beta}_w)\|_w = |\boldsymbol{\beta}_w|$.

Given a (column) vector $\boldsymbol{\beta}_{wy}$ having entries β_{wyj}, $j = 1, \ldots, N_n$, set $\boldsymbol{\beta} = (\boldsymbol{\beta}_{wy})$ and $\mathbf{g}(\cdot, \cdot; \boldsymbol{\beta}) = (g_{wy}(\cdot, \cdot; \boldsymbol{\beta}_{wy}))$, where $g_{wy}(\cdot, \cdot; \boldsymbol{\beta}_{wy}) = \sum_j \beta_{wyj}\phi_{wj}$. Then the log-likelihood function corresponding to $\mathbf{g}$, viewed as a function of $\boldsymbol{\beta}$, is concave. Recall that $\hat{\boldsymbol{\alpha}}_n$ is the maximum likelihood estimate and $\boldsymbol{\alpha}_n^*$ is the best approximation to $\boldsymbol{\alpha}$ with constituents in G. Let $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$ be given by the equations $\hat{\boldsymbol{\alpha}}_n(\cdot, \cdot) = \mathbf{g}(\cdot, \cdot; \hat{\boldsymbol{\beta}})$ and $\boldsymbol{\alpha}_n^*(\cdot, \cdot) = \mathbf{g}(\cdot, \cdot; \boldsymbol{\beta}^*)$. Then $\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_n^*\| = |\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|$.

Set $\eta_{ivy} = \text{ind}\,(T_{iv} < \infty \text{ and } Y_i(T_{iv}) = y)$. Then the (scaled) log-likelihood function corresponding to the candidate $\mathbf{g} = (g_{wy})$ for $\boldsymbol{\alpha}$ can be written as

$$l(\mathbf{g}) = \frac{1}{n}\sum_i l_i(\mathbf{g}), \quad \text{where } l_i(\mathbf{g}) = \sum_v l_{iv}(\mathbf{g}) = \sum_v [l_{iv1}(\mathbf{g}) + l_{iv2}(\mathbf{g})]$$

with

$$l_{iv1}(\mathbf{g}) = \delta_{iv} \sum_{w,y} \gamma_{ivw}\eta_{ivy} g_{wy}(T_{iv} - T_{i,v-1}, \mathbf{X}_i(T_{iv}))$$

and

$$l_{iv2}(\mathbf{g}) = -\sum_{w,y} \gamma_{ivw} \int_{T_{i,v-1}\wedge C_i}^{T_{iv}\wedge C_i} \exp\{g_{wy}(t - T_{i,v-1}, \mathbf{X}_i(t))\} dt.$$

Let

$$\mathbf{S}(\boldsymbol{\beta}) = \frac{\partial l(\mathbf{g})}{\partial \boldsymbol{\beta}} = \frac{1}{n}\sum_i \sum_v \frac{\partial l_{iv}(\mathbf{g})}{\partial \boldsymbol{\beta}}$$

denote the score at $\boldsymbol{\beta}$; that is, the vector whose entries are given by

$$\frac{\partial l_{iv1}(\mathbf{g})}{\partial \beta_{wyj}} = \delta_{iv}\gamma_{ivw}\eta_{ivy}\phi_{wj}(T_{iv} - T_{i,v-1}, \mathbf{X}_i(T_{iv}))$$

and

$$\frac{\partial l_{iv2}(\mathbf{g})}{\partial \beta_{wyj}} = -\gamma_{ivw}\int_{T_{i,v-1}\wedge C_i}^{T_{iv}\wedge C_i} \phi_{wj}(t - T_{i,v-1}, \mathbf{X}_i(t)) \exp\{g_{wy}(t - T_{i,v-1}, \mathbf{X}_i(t))\}\, dt.$$

Let

$$\mathbf{D}(\boldsymbol{\beta}) = \frac{\partial^2 l(\mathbf{g})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\mathrm{T}}} = \frac{1}{n}\sum_i \sum_v \frac{\partial^2 l_{iv}(\mathbf{g})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\mathrm{T}}}$$

denote the Hessian of $l(\mathbf{g})$; that is, the matrix whose possibly non-zero entries are given by

$$\frac{\partial^2 l_{iv}(\mathbf{g})}{\partial \beta_{wyj_1}\partial \beta_{wyj_2}} = -\gamma_{ivw}\int_{T_{i,v-1}\wedge C_i}^{T_{iv}\wedge C_i} \phi_{wj_1}(t - T_{i,v-1}, \mathbf{X}_i(t)) \phi_{wj_2}(t - T_{i,v-1}, \mathbf{X}_i(t)) \exp\{g_{wy}(t - T_{i,v-1}, \mathbf{X}_i(t))\}\, dt. \tag{5.3}$$

Then the following identity holds:

$$l(\boldsymbol{\beta}) = l(\boldsymbol{\beta}^*) + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{S}(\boldsymbol{\beta}^*) + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^{\mathrm{T}}\left[\int_0^1 (1-t)\,\mathbf{D}(\boldsymbol{\beta}^* + t(\boldsymbol{\beta} - \boldsymbol{\beta}^*))\, dt\right](\boldsymbol{\beta} - \boldsymbol{\beta}^*). \tag{5.4}$$

Lemma 5.3
For any positive constant M,

$$\lim_{c\to\infty} \limsup_{n\to\infty} P\left(|\mathbf{S}(\boldsymbol{\beta}^*)| \geqslant Mc\left(\frac{N_n}{n}\right)^{1/2}\right) = 0.$$

Proof. By the definition of $\boldsymbol{\beta}^*$, $E(\partial l(\boldsymbol{\alpha}_n^*)/\partial \beta_{wyj}) = 0$. Hence

$$E\left[\left(\frac{\partial l(\boldsymbol{\alpha}_n^*)}{\partial \beta_{wyj}}\right)^2\right] = \frac{1}{n} E\left[\left(\frac{\partial l_i(\boldsymbol{\alpha}_n^*)}{\partial \beta_{wyj}}\right)^2\right]$$

$$\leqslant \frac{2}{n}\left\{E\left[\left(\sum_\nu \frac{\partial l_{i\nu 1}(\boldsymbol{\alpha}_n^*)}{\partial \beta_{wyj}}\right)^2\right] + E\left[\left(\sum_\nu \frac{\partial l_{i\nu 2}(\boldsymbol{\alpha}_n^*)}{\partial \beta_{wyj}}\right)^2\right]\right\}.$$

We claim that there are positive constants M_6, M_7 such that

$$E\left[\left(\sum_\nu \frac{\partial l_{i\nu 1}(\boldsymbol{\alpha}_n^*)}{\partial \beta_{wyj}}\right)^2\right] \leqslant M_6 \|\phi_{wj}\|_w^2 \tag{5.5}$$

and

$$E\left[\left(\sum_\nu \frac{\partial l_{i\nu 2}(\boldsymbol{\alpha}_n^*)}{\partial \beta_{wyj}}\right)^2\right] \leqslant M_7 \|\phi_{wj}\|_w^2. \tag{5.6}$$

[The proofs of (5.5) and (5.6) are deferred to the appendix.] Consequently,

$$E|\mathbf{S}(\boldsymbol{\beta}^*)|^2 = \sum_{w,y}\sum_j E\left[\left(\frac{\partial l(\boldsymbol{\alpha}_n^*)}{\partial \beta_{wyj}}\right)^2\right] \leqslant 2(M_6 + M_7)K^2 \frac{N_n}{n}.$$

The desired result follows.

Lemma 5.4

There is a positive constant M_8 such that, for any fixed $c > 0$,

$$(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^{\mathrm{T}}\left[\int_0^1 (1-u)\mathbf{D}(\boldsymbol{\beta}^* + u(\boldsymbol{\beta} - \boldsymbol{\beta}^*))\, du\right](\boldsymbol{\beta} - \boldsymbol{\beta}^*)$$

$$\leqslant -M_8|\boldsymbol{\beta} - \boldsymbol{\beta}^*|^2 \quad \text{for all } \boldsymbol{\beta} = (\beta_{wyj}) \text{ with } |\boldsymbol{\beta} - \boldsymbol{\beta}^*| = c\left(\frac{N_n}{n}\right)^{1/2}$$

on an event $\Omega_n(c)$ with $\lim_n P(\Omega_n(c)) = 1$.

Proof. Choose $\mathbf{a}(\cdot, \cdot) = \mathbf{a}(\cdot, \cdot; \boldsymbol{\beta})$ such that $|\boldsymbol{\beta} - \boldsymbol{\beta}^*| = c(N_n/n)^{1/2}$. Then by the triangle inequality and lemma 3.1, $\|\mathbf{a}\|_\infty \leqslant A_n M_2^{-1} c(N_n/n)^{1/2} + \|\boldsymbol{\alpha}_n^*\|_\infty$. Thus by lemma 5.2,

$$\lim_{n\to\infty} P(\|\mathbf{a}\|_\infty \leqslant 1 + \|\boldsymbol{\alpha}^*\|_\infty \quad \text{for all } \mathbf{a} = (a_{wy})$$

$$\text{with } a_{wy} \in G \text{ and } \|\mathbf{a} - \boldsymbol{\alpha}_n^*\| = c(N_n/n)^{1/2}) = 1. \tag{5.7}$$

Recall that $\boldsymbol{\alpha}_n^* = \mathbf{g}(\cdot, \cdot; \boldsymbol{\beta}^*) = (\alpha_{nwy}^*)$. It follows from (5.3) that, for $u \in [0, 1]$,

$$(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{D}(\boldsymbol{\beta}^* + u(\boldsymbol{\beta} - \boldsymbol{\beta}^*))(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$$

$$= -\frac{1}{n}\sum_i \sum_\nu \sum_{w,y} \gamma_{i\nu w} \int_{T_{i,\nu-1}\wedge C_i}^{T_{i,\nu}\wedge C_i} [a_{wy} - \alpha_{nwy}^*]^2 (t - T_{i,\nu-1}, \mathbf{X}_i(t))$$

$$\times \exp\{[\alpha_{nwy}^* + u(a_{wy} - \alpha_{nwy}^*)](t - T_{i,\nu-1}, \mathbf{X}_i(t))\}\, dt.$$

Therefore, by (5.7) and condition 1, except on an event whose probability tends to zero as $n \to \infty$,

$$(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{D}(\boldsymbol{\beta}^*+u(\boldsymbol{\beta}-\boldsymbol{\beta}^*))(\boldsymbol{\beta}-\boldsymbol{\beta}^*) \leqslant -M_8 \sum_{w,y} \|a_{wy}-\alpha^*_{nwy}\|^2_{nw},$$

for all $\boldsymbol{\beta}=(\beta_{wyj})$ with $|\boldsymbol{\beta}-\boldsymbol{\beta}^*| = c(N_n/n)^{1/2}$ and all $u \in [0, 1]$. Hence, it follows from lemma 3.2 that, except on an event whose probability tends to zero as $n \to \infty$,

$$(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^{\mathrm{T}}\left[\int_0^1 (1-u)\mathbf{D}(\boldsymbol{\beta}^*+u(\boldsymbol{\beta}-\boldsymbol{\beta}^*))\,du\right](\boldsymbol{\beta}-\boldsymbol{\beta}^*)$$
$$\leqslant -M_8 \sum_{w,y} \|a_{wy}-\alpha^*_{nwy}\|^2_w = -M_8\|\mathbf{a}-\boldsymbol{\alpha}^*_n\|^2 = -M_8|\boldsymbol{\beta}-\boldsymbol{\beta}^*|^2,$$

for all $\boldsymbol{\beta}=(\beta_{wyj})$ with $|\boldsymbol{\beta}-\boldsymbol{\beta}^*| = c(N_n/n)^{1/2}$. This completes the proof of the lemma.

Lemma 5.5
The maximum-likelihood estimate $\hat{\boldsymbol{\alpha}}_n$ exists except on an event whose probability tends to zero as $n \to \infty$. Moreover $\|\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}^*_n\|^2 = O_P(N_n/n)$.

Proof. Since $l(\boldsymbol{\beta})$ is a concave function of $\boldsymbol{\beta}$, we conclude from (5.4) and lemma 5.4 that,

$$\left\{\hat{\boldsymbol{\beta}} \text{ exists and } |\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*| < c\left(\frac{N_n}{n}\right)^{1/2}\right\} \cap \Omega_n(c)$$
$$\supset \left\{l(\boldsymbol{\beta})-l(\boldsymbol{\beta}^*)<0 \quad \text{for all } \boldsymbol{\beta}=(\beta_{wyj}) \text{ with } |\boldsymbol{\beta}-\boldsymbol{\beta}^*| = c\left(\frac{N_n}{n}\right)^{1/2}\right\} \cap \Omega_n(c)$$
$$\supset \left\{|\mathbf{S}(\boldsymbol{\beta}^*)| < M_8 c\left(\frac{N_n}{n}\right)^{1/2}\right\} \cap \Omega_n(c).$$

Hence, by lemma 5.3,

$$\lim_{c\to\infty} \liminf_{n\to\infty} P\left(\hat{\boldsymbol{\beta}} \text{ exists and } |\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*| < c\left(\frac{N_n}{n}\right)^{1/2}\right)$$
$$\geqslant \lim_{c\to\infty} \liminf_{n\to\infty} \left\{P\left(|\mathbf{S}(\boldsymbol{\beta}^*)| < M_8 c\left(\frac{N_n}{n}\right)^{1/2}\right) + P(\Omega_n(c)) - 1\right\} = 1.$$

The conclusion of the lemma follows.

Theorem 2.2 follows from lemmas 3.2, 5.2 and 5.5.

6. Proof of theorem 2.4

The estimation error has the ANOVA decomposition $\hat{\alpha}_{wy}-\alpha^*_{nwy} = \sum_{s\in\mathscr{S}}(\hat{\alpha}_{wys}-\alpha^*_{nwys})$. It follows from theorem 3 and lemmas 3.4 and 5.5 that, for each $s \in \mathscr{S}$, $\|\hat{\alpha}_{wys}-\alpha^*_{nwys}\|^2_w = O_P(N_n/n)$ and $\|\hat{\alpha}_{wys}-\alpha^*_{nwys}\|^2_{nw} = O_P(N_n/n)$.

The approximation error has the ANOVA decomposition $\alpha^*_{nwy}-\alpha^*_{wy} = \sum_{s\in\mathscr{S}}(\alpha^*_{nwys}-\alpha^*_{wys})$. By the same argument as in lemma 7 of Huang (1998), for each $s \in \mathscr{S}$, there are functions $g_s \in G_s$, with $g_s \perp_{nw} G_r$ for every proper subset r of s, such that

$$\|\alpha^*_{wys}-g_s\|^2_w = O_P\left(\sum_{r\subset s, r\neq s} \frac{N_s}{n}+\rho^2_{wys}\right) \tag{6.1}$$

and

$$\|\alpha^*_{wys} - g_s\|^2_{nw} = O_P\left(\sum_{r\subset s, r\neq s} \frac{N_s}{n} + \rho^2_{wys}\right). \tag{6.2}$$

Write $g = \sum_{s\in\mathscr{S}} g_s$. Then $\|g - \alpha^*_{wy}\|^2_w = O_P(\sum_{s\in\mathscr{S}} N_n/n + \sum_{s\in\mathscr{S}} \rho^2_{wys})$. Thus, by the triangle inequality and lemma 5.2,

$$\|g - \alpha^*_{nwy}\|^2_w \leqslant 2\|g - \alpha^*_{wy}\|^2_w + 2\|\alpha^*_{wy} - \alpha^*_{nwy}\|^2_w = O_P\left(\sum_{s\in\mathscr{S}} \frac{N_s}{n} + \overline{\rho}^2_n\right).$$

Therefore, by lemma 3.4, except on an event whose probability tends to zero as $n \to \infty$,

$$\|g_s - \alpha^*_{nwys}\|^2_w \leqslant \epsilon_2^{1-\#(s)}\|g - \alpha^*_{nwy}\|^2_w = O_P\left(\sum_{s\in\mathscr{S}} \frac{N_s}{n} + \overline{\rho}^2_n\right).$$

Hence, it follows from (6.1), (6.2), the triangle inequality and lemma 3.2 that, for each $s \in \mathscr{S}$, $\|\alpha^*_{nwys} - \alpha^*_{wys}\|^2_w = O_P(N_n/n + \overline{\rho}^2_n)$ and $\|\alpha^*_{nwys} - \alpha^*_{wys}\|^2_{nw} = O_P(N_n/n + \overline{\rho}^2_n)$. This completes the proof of theorem 2.4.

Acknowledgement

This work was supported in part by NSF Grant DMS-9504463.

References

Allison, P. D. (1984). *Event history analysis: regression for longitudinal event data.* Sage, Beverly Hills, CA.

Andersen, P. K., Borgan, Ø., Gill, R. D. & Keiding, N. (1993). *Statistical models based on counting processes.* Springer-Verlag, New York.

Hamerle, A. (1989). Multiple-spell regression models for duration data. *Appl. Statist.* **38**, 127–138.

Hansen, M. (1994). Extended linear models, multivariate splines, and ANOVA. PhD Dissertation, University of California at Berkeley.

Hansen, M., Kooperberg, C. & Sardy, S. (1996). Triogram models. Technical Report 304, Department of Statistics, University of Washington, Seattle.

Huang, J. Z. (1999). Projection estimation in multiple regression with application to functional ANOVA models. *Ann. Statist.* **26**, 242–272.

Kalbfleisch, J. D. & Prentice R. L. (1980). *The statistical analysis of failure time data.* Wiley, New York.

Kooperberg, C., Stone, C. J. & Truong, Y. K. (1995a). Hazard regression. *J. Amer. Statist. Assoc.* **90**, 78–94.

Kooperberg, C., Stone, C. J. & Truong, Y. K. (1995b). The L_2 rate of convergence for hazard regression. *Scand. J. Statist.* **22**, 143–157.

Mayer, K. U. & Tuma, N. B. (1990). *Event history analysis in life course research.* University of Wisconsin Press, Madison.

Schumaker, L. L. (1981). *Spline functions: basic theory.* Wiley, New York.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **8**, 1348–1360.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *Ann. Statist.* **22**, 118–171.

Stone, C. J., Hansen, M., Kooperberg, C. & Truong, Y. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *Ann. Statist.* **25**, 1371–1470.

Yamaguchi, K. (1991). *Event history analysis.* Sage, Newbury Park, CA.

Received September 1996, in final form September 1997

Jianhua Huang, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6302, USA.

Charles J. Stone, Department of Statistics, University of California, Berkeley, CA 94720-3860, USA.

Appendix

Proof of proposition 3.2. We prove only the first equality. The proof for the second equality is similar. For $\nu > 1$, by conditioning, we get that

$$E\left[\delta_\nu \gamma_{\nu w} \eta_{\nu y} a(T_\nu - T_{\nu-1}, \mathbf{X}(T_\nu))\right] = \sum_v E\big[P(Y(0) = v|\mathbf{X}) E(\mathrm{ind}\,(T_\nu \leqslant C)\,\mathrm{ind}\,(Y(T_{\nu-1}) = w)$$
$$\mathrm{ind}(Y(T_\nu) = y)\, a(T_\nu - T_{\nu-1}, \mathbf{X}(T_\nu))|\, Y(0) = v, \mathbf{X})\big].$$

By the semi-Markov property of the process Y, the last expression equals

$$\sum_v E\left[P(Y(0) = v|\mathbf{X}) \int_0^\tau ds \int_s^\tau dt \rho_{\nu-1}(s, w|\nu, \mathbf{X}) \lambda_{wy}(t - s, \mathbf{X}(t))\right.$$
$$\left. \exp\left(-\int_0^{t-s} \lambda_w(u, \mathbf{X}(u + s))\, du\right) a(t - s, \mathbf{X}(t)) P(C \geqslant t|\mathbf{X})\right].$$

Then by change of variable and changing the order of integration, we can see that the previous expression equals $\int_{\mathscr{T}} \int_{\mathscr{X}} a(t, \mathbf{x}) \lambda_{wy}(t, \mathbf{x}) B_{\nu w}(t, \mathbf{x})\, d\mathbf{x}\, dt$. The case for $\nu = 1$ can be proved similarly.

Proof of (3.2). We have that

$$\mathrm{var}\,[\Psi_w(a_1, a_2, \mathbf{X}, \mathbf{Y})] \leqslant E\left[\int_0^\tau \sum_\nu \gamma_{\nu w}\, \mathrm{ind}\,(T_{\nu-1} \wedge C \leqslant t \leqslant T_\nu \wedge C)\right.$$
$$\left. a_1(t - T_{\nu-1}, \mathbf{X}(t)) a_2(t - T_{\nu-1}, \mathbf{X}(t))\, dt\right]^2.$$

By the Cauchy–Schwartz inequality, the last expression is bounded above by

$$\tau E \sum_\nu \gamma_{\nu w} \int_{T_{\nu-1} \wedge C}^{T_\nu \wedge C} a_1^2(t - T_{\nu-1}, \mathbf{X}(t)) a_2^2(t - T_{\nu-1}, \mathbf{X}(t))\, dt \leqslant \tau \|a_1\|_w^2 \|a_2\|_\infty^2.$$

Proof of (5.5). We have

$$E\left[\left(\sum_\nu \frac{\partial l_{i\nu 1}(\boldsymbol{\alpha}_n^*)}{\partial \beta_{wyj}}\right)^2\right] = E\left[\sum_\nu \delta_\nu \gamma_{\nu w} \eta_{\nu y} \phi_{wj}(T_\nu - T_{\nu-1}, \mathbf{X}(T_\nu))\right]^2.$$

By the Cauchy–Schwartz inequality, the right side of this equation is bounded above by

$$E\left[\sum_\nu \delta_\nu\right]\left[\sum_\nu \delta_\nu \gamma_{\nu w} \eta_{\nu y} \phi_{wj}^2(T_\nu - T_{\nu-1}, \mathbf{X}(T_\nu))\right]$$
$$= E \sum_{\nu 2}\left[\sum_{\nu_1 \leqslant \nu_2} + \sum_{\nu_1 > \nu_2}\right]\left\{\delta_{\nu_1} \delta_{\nu_2} \gamma_{\nu_2 w} \eta_{\nu_2 y} \phi_{wj}^2(T_{\nu_2} - T_{\nu_2 - 1}, \mathbf{X}(T_{\nu_2}))\right\}$$
$$= \sum_{\nu_2} E\left[\nu_2 \delta_{\nu_2} \gamma_{\nu_2 w} \eta_{\nu_2 y} \phi_{wj}^2(T_{\nu_2} - T_{\nu_2 - 1}, \mathbf{X}(T_{\nu_2}))\right]$$
$$+ \sum_{\nu_2} \sum_{\nu_1 > \nu_2} E\left[\delta_{\nu_1} \delta_{\nu_2} \gamma_{\nu_2 w} \eta_{\nu_2 y} \phi_{wj}^2(T_{\nu_2} - T_{\nu_2 - 1}, \mathbf{X}(T_{\nu_2}))\right].$$

From the semi-Markov property and proposition 3.1, if $\nu_1 > \nu_2$, then

$$E[\delta_{\nu_1}\delta_{\nu_2}\gamma_{\nu_2 w}\eta_{\nu_2 y}\phi_{wj}^2(T_{\nu_2} - T_{\nu_2-1}, \mathbf{X}(T_{\nu_2}))]$$
$$\leqslant E\Big[\operatorname{ind}(T_{\nu_1} \leqslant \tau)\delta_{\nu_2}\gamma_{\nu_2 w}\eta_{\nu_2 y}\phi_{wj}^2(T_{\nu_2} - T_{\nu_2-1}, \mathbf{X}(T_{\nu_2}))\Big]$$
$$\leqslant M_1\tau(1-\epsilon)^{\nu_2-\nu_1}E\Big[\delta_{\nu_2}\gamma_{\nu_2 w}\eta_{\nu_2 y}\phi_{wj}^2(T_{\nu_2} - T_{\nu_2-1}, \mathbf{X}(T_{\nu_2}))\Big].$$

Thus,

$$E\left[\left(\sum_{\nu}\frac{\partial l_{i\nu 1}(\boldsymbol{\alpha}_n^*)}{\partial\beta_{wyj}}\right)^2\right]$$
$$\leqslant \sum_{\nu_2}\left[\nu_2 + \sum_{\nu_1>\nu_2} M_1\tau(1-\epsilon)^{\nu_1-\nu_2}\right]E\Big[\delta_{\nu_2}\gamma_{\nu_2 w}\eta_{\nu_2 y}\phi_{wj}^2(T_{\nu_2} - T_{\nu_2-1}, \mathbf{X}(T_{\nu_2}))\Big].$$

Applying proposition 3.2, we can write the right side of this inequality as

$$\sum_{\nu_2}\left[\nu_2 + M_1\tau\frac{1-\epsilon}{\epsilon}\right]\int_{\mathscr{T}}\int_{\mathscr{X}}\phi_{wj}^2(t, \mathbf{x})\lambda_{wy}(t, \mathbf{x})B_{\nu_2 w}(t, \mathbf{x})\,d\mathbf{x}\,dt. \tag{7.1}$$

Proposition 3.1 and condition 1 together imply that

$$\sum_{\nu_2}\left[\nu_2 + M_1\tau\frac{1-\epsilon}{\epsilon}\right]\lambda_{wy}(t, \mathbf{x})B_{\nu_2 w}(t, \mathbf{x})$$

is bounded above uniformly in $t \in \mathscr{T}$ and $\mathbf{x} \in \mathbf{X}$. Therefore, the expression (7.1) is less than a multiple of $\|\phi_{wj}\|_{L_2}$, and hence, by lemma 3.1, by a multiple of $\|\phi_{wj}\|_w^2$.

Proof of (5.6). Plugging into the expression for

$$\frac{\partial l_{i\nu 2}(\boldsymbol{\alpha}_n^*)}{\partial\beta_{wyj}},$$

we have that

$$E\left[\left(\sum_{\nu}\frac{\partial l_{i\nu 2}(\boldsymbol{\alpha}_n^*)}{\partial\beta_{wyj}}\right)^2\right] = E\left[\int_0^{\tau}\sum_{\nu}\gamma_{\nu w}\operatorname{ind}(T_{\nu-1}\wedge C \leqslant t \leqslant T_{\nu}\wedge C)\right.$$
$$\left.\phi_{wj}(t - T_{\nu-1}, \mathbf{X}(t))\exp(\alpha_{nwy}^*(t - T_{\nu-1}, \mathbf{X}(t)))\,dt\right]^2.$$

Applying the Cauchy–Schwartz inequality, we bound the last expression by

$$\tau E\int_0^{\tau}\sum_{\nu}\gamma_{\nu w}\operatorname{ind}(T_{\nu-1}\wedge C \leqslant t \leqslant T_{\nu}\wedge C)$$
$$\phi_{wj}^2(t - T_{\nu-1}, \mathbf{X}(t))\exp(2\alpha_{nwy}^*(t - T_{\nu-1}, \mathbf{X}(t)))\,dt.$$

By condition 3, α_{nwy}^* is bounded. Thus, the above expression is bounded above by a constant multiple of

$$E\sum_{\nu}\gamma_{\nu w}\int_{T_{\nu-1}\wedge C}^{T_{\nu}\wedge C}\phi_{wj}^2(t - T_{\nu-1}, \mathbf{X}(t))\,dt = \|\phi_{wj}\|_w^2.$$