*The Annals of Statistics*
2000, Vol. 28, No. 4, 961–999

# FUNCTIONAL ANOVA MODELING FOR PROPORTIONAL HAZARDS REGRESSION

By Jianhua Z. Huang,[1] Charles Kooperberg,[2]
Charles J. Stone[3] and Young K. Truong[4]

*University of Pennsylvania, Fred Hutchinson Cancer Research Center,
University of California, Berkeley,
University of North Carolina, Chapel Hill and
National University of Singapore*

The logarithm of the relative risk function in a proportional hazards model involving one or more possibly time-dependent covariates is treated as a specified sum of a constant term, main effects, and selected interaction terms. Maximum partial likelihood estimation is used, where the maximization is taken over a suitably chosen finite-dimensional estimation space, whose dimension increases with the sample size and which is constructed from linear spaces of functions of one covariate and their tensor products. The $L_2$ rate of convergence for the estimate and its ANOVA components is obtained. An adaptive numerical implementation is discussed, whose performance is compared to (full likelihood) hazard regression both with and without the restriction to proportional hazards.

**1. Introduction.** In survival analysis, a popular approach for treating the relationship between the survival time and the covariates is through the use of Cox's proportional hazards regression model [Cox (1972)]. Kalbfleisch and Prentice (1980), Cox and Oakes (1984), Fleming and Harrington (1991) and Andersen, Borgan, Gill and Keiding (1993) discuss this approach in considerable detail. An important issue that has been overshadowed by the popularity of the Cox model is the assumption of linearity of the covariate effects, which may or may not be valid in practice. If it is violated, then the corresponding estimates will be biased. One remedy is the use of nonparametric modeling, in which covariate effects are estimated without imposing a prespecified parametric form. Nonparametric modeling provides more flexibility than Cox's original approach, but when there are multiple covariates its direct implementation is subject to the "curse of dimensionality," which refers to the difficulty caused by data sparseness in high dimensions. This difficulty can be ameliorated by using functional analysis of variance (ANOVA) models, where the overall effect of the covariates is modeled as a specified sum of a constant effect, main effects (functions of one covariate) and selected low-order

Received April 27, 1998; revised July 2000.

[1] Supported by University of Pennsylvania Research Foundation.

[2] Supported in part by NIH Grant CA74841.

[3] Supported in part by NSF Grant DMS-95-04463.

[4] Supported in part by NSF Grant DMS-94-03800 and NIH Grant CA61937.

*AMS* 1991 *subject classifications*. Primary 62N05; secondary 62G05.

*Key words and phrases*. Convergence rates, Cox model, partial likelihood, splines, tensor product.

interactions (functions of a few covariates). Such ANOVA models can be used to achieve dimensionality reduction and at the same time retain the flexibility of nonparametric modeling.

A comprehensive review of one approach to functional ANOVA modeling is given in Stone, Hansen, Kooperberg and Truong (1997). The main objective of the present paper is to conduct a theoretical investigation of this approach in the context of proportional hazards regression. To this end, let $T$, $C$ and $\mathbf{X} = \mathbf{X}(\cdot)$ have a joint distribution, where $T$ and $C$ are nonnegative random variables and $\mathbf{X}(t)$ is a random vector of possibly time-dependent covariates $X_1(t), \dots, X_L(t)$ (some of which could also be vector valued). In survival analysis, $T$ and $C$ are referred to as the survival time (or failure time) and censoring time, respectively. Set $Y = \min(T, C)$ and $\delta = \operatorname{ind}(T \le C)$. The indicator random variable $\delta$ equals 1 if failure occurs on or before the censoring time (i.e., if $T \le C$), and it equals 0 otherwise. The observable time $Y$ is said to be uncensored or censored according as $\delta = 1$ or $\delta = 0$. Instead of observing the triple $(T, C, \mathbf{X})$, we only observe $(Y, \delta, \mathbf{X})$.

The covariate process $\mathbf{X}$ is assumed to be external, that is, not directly involved with the failure mechanism [see page 123 of Kalbfleisch and Prentice (1980)]. For identifiability, it is assumed that $T$ and $C$ are conditionally independent given $\mathbf{X}$. Let $F(t \mid \mathbf{X}) = P(T \le t \mid \mathbf{X})$ and $\overline{F}(t \mid \mathbf{X}) = P(T > t \mid \mathbf{X})$ denote the conditional distribution function and conditional survival function, respectively, of $T$ given $\mathbf{X}$. The conditional density function $f(t \mid \mathbf{X})$ is assumed to exist, and the conditional hazard function is given by $\lambda(t \mid \mathbf{X}) = f(t \mid \mathbf{X})/\overline{F}(t \mid \mathbf{X})$. The proportional hazards assumption is that the log-hazard function has the form $\log \lambda(t \mid \mathbf{X}) = \alpha_0(t) + \alpha(\mathbf{X}(t))$, where $\lambda_0 = \exp \alpha_0$ and $\alpha$ are referred to, respectively, as the baseline hazard function and log relative risk function (log of the relative risk function). Here the log-hazard at time $t$ is assumed to depend only on the current values of the covariates. Note that in the above expression for $\log \lambda(t \mid \mathbf{X})$, if one adds a constant to $\alpha_0(t)$ and subtracts the same constant form $\alpha(\mathbf{X}(t))$, then $\log \lambda(t \mid \mathbf{X})$ does not change. Hence we say that the function $\alpha(\cdot)$ is not identifiable and some identifiability constraint needs to be imposed to make $\alpha(\mathbf{X}(t))$ uniquely defined. [In the Cox model the usual constraint is $\alpha(\mathbf{0}) = 0$.] The goal is to estimate the log relative risk function $\alpha(\cdot)$.

The functional ANOVA approach involves the choice of a special form for the log relative risk function. Suppose, for example, that $L = 3$ and consider the model

$$\alpha(\mathbf{X}(t)) = \alpha_1(X_1(t)) + \alpha_2(X_2(t)) + \alpha_3(X_3(t)) + \alpha_{1,2}(X_1(t), X_2(t)), \tag{1}$$

which omits the three-factor interaction and contains exactly one of the three possible two-factor interactions. In order to make the representation in (1) unique, we need to impose some identifiability constraints on the various components in the representation. Given a random sample, we can use maximum partial likelihood estimation to obtain an estimate $\hat{\alpha}$ having the same form as (1), where the partial log-likelihood is defined in (3) of Section 2 and the maximization is carried out in a suitably chosen finite-dimensional estimation space, whose dimension increases with the sample size. If the components of the ANOVA decomposition of $\alpha$ are accurately estimated by the corresponding, similarly constrained, components of $\hat{\alpha}$, then examination of the components of $\hat{\alpha}$ should shed light on the relationship of the survival time $T$ to $\mathbf{X}$ through the log relative risk function $\alpha$.

In this paper the estimation spaces are constructed from linear spaces of functions of one variable and their tensor products while respecting the ANOVA structure of the target function $\alpha$. In particular, polynomial splines and their tensor products can be used as building blocks for the estimation spaces. We will give conditions under which $\hat{\alpha}$ converges to $\alpha$ and, more important, the components of $\hat{\alpha}$ converge to the corresponding components of $\alpha$. Rates of convergence are also studied.

Technically speaking, the theory developed in this paper does not depend on the validity of the assumed form of the ANOVA model for the log relative risk function. In particular, if (1) is invalid, then $\hat{\alpha}$ can be viewed as an estimate of the function $\alpha^*$ having the form

$$\alpha^*(\mathbf{X}(t)) = \alpha_1^*(X_1(t)) + \alpha_2^*(X_2(t)) + \alpha_3^*(X_3(t)) + \alpha_{12}^*(X_1(t), X_2(t)) \tag{2}$$

that maximizes the expected partial log-likelihood as defined in (4) in Section 2. In fact, the proofs do not even depend on the validity of the proportional hazards assumption. If this assumption is invalid, then (2) corresponds to an approximation to $\lambda(t \mid \mathbf{X})$ having the form given by $\log \lambda^*(t \mid \mathbf{X}) = \alpha_0^*(t) + \alpha^*(\mathbf{X}(t))$.

Cox's model with nonparametric or additive covariate effects has been a very useful exploratory tool for analyzing survival data. There are several methodological approaches. For example, O'Sullivan (1998) and Gray (1992) used smoothing splines. Tibshirani and Hastie (1987) considered the local likelihood method. Sleeper and Harrington (1990) and Gentleman and Crowley (1991) employed regression spline methods. LeBlanc and Crowley (1999) developed an adaptive regression spline method for fitting additive and general interaction models. Gu (1996) developed a smoothing spline approach to (nonproportional) hazard regression that has similarities to the HARE methodology of Kooperberg, Stone and Truong (1995a). Wahba, Wang, Gu, Klein and Klein (1995) discussed ANOVA decompositions for smoothing spline models in a general context, but did not treat hazard regression explicitly. In theoretical papers, O'Sullivan (1993) studied rates of convergence of the penalized methods (smoothing spline approach) for saturated models and Dabrowska (1997) considered a partly linear model in which a kernel method is used to adjust for the nonparametric covariate effects while obtaining the usual $n^{-1/2}$ rate of convergence for the parametric effect. Neither paper considered additive models or more generally functional ANOVA models. The current paper gives a rather complete theoretical account of functional ANOVA models for maximum partial likelihood estimation over finite-dimensional estimation spaces. Similar results for other methods have yet to be established.

An alternative approach to functional ANOVA modeling in proportional hazards regression is to model the entire hazard function, including the baseline hazard and the covariate effects, and to maximize the full likelihood to fit the data. Theoretical accounts of this approach have been given in Kooperberg, Stone and Truong (1995b) and Huang and Stone (1998); Kooperberg, Stone and Truong (1995a) developed an adaptive hazard regression methodology (HARE). In comparison with the full likelihood approach, a theoretical obstacle that arises in maximum partial likelihood estimation is the lack of identifiability of the relative risk function when the baseline hazard function is not modeled. This obstacle causes difficulties in establishing the existence of the best approximation in the model space (see the proof of Theorem 1). The lack of identifiability also causes the lack of negative definiteness of the Hessian of the partial log-likelihood and expected partial log-likelihood functions, which is handled by using orthogonality to constant functions as an identifiability constraint [see (5) and Section 2.2]. To ensure that the identifiability constraint does not destroy the nice structure of the ANOVA decompositions of the model space and estimation space, the theoretical inner product is used to impose the identifiability constraint on the target function while the empirical inner product is used for the estimate (see Section 2.3). The imposition of these constraints necessarily complicates the analysis. Another technical obstacle we overcome here is that, due to the special features of partial log-likelihood, the empirical inner product cannot be expressed as a summation over iid random variables, so the theory in Huang (1998a) cannot be used directly as in Huang and Stone (1998).

The rest of the paper is organized as follows. In Section 2, we present the main theoretical results. Section 2.1 contains the theoretical set-up, in which the partial log-likelihood function is defined and the unknown log relative risk function is modeled in a general linear function space, referred to as the model space. Theorem 1 establishes the existence and uniqueness of the function $\alpha^*$ in the model space that maximizes the expected partial log-likelihood. This function should be viewed as the target of our estimate. Under the assumption of proportional hazards, $\alpha^*$ can be thought of as the best approximation in the model space to the log relative risk function $\alpha$. A general result (Theorem 2) on the rate of convergence of the maximum partial likelihood estimate is given in Section 2.2 Section 2.3 is devoted to functional ANOVA modeling, where the ANOVA decompositions of the estimate and target function are formally defined. The convergence properties of the maximum partial likelihood estimate and its ANOVA components are given in Theorem 3. Section 2.4 treats functional ANOVA modeling when polynomial splines and their tensor products are used as the building blocks in constructing the estimation spaces; here Theorem 3 is applied to obtain the rate of convergence of the corresponding estimate. Section 3 describes some experience with a numerical adaptive implementation of proportional hazards regression, which is compared to HARE both with and without a restriction to consider only proportional hazards models. This material should be regarded as merely suggestive, not as an attempt at a thorough study of the methodological issues that complement the present, mainly theoretical, investigation. Some discussion is given in Section 4. The proofs of Theorems 2 and 3 are given in Section 5 and 6, respectively. The proofs of Theorem 1 and several of the lemmas in Section 5 are collected in the Appendix.

## 2. Main results.

2.1. *Theoretical set-up.* Let $\tau$ be a fixed positive number and suppose that censoring automatically occurs at time $\tau$ if it has not occurred prior to that time; that is, $P(C \le \tau) = 1$. Suppose also that each of the random vectors $\mathbf{X}(t)$, $t \in [0, \tau]$, takes values in a fixed compact set $\mathscr{X}$ of some Euclidean space. Conditions similar to the following one are used in Kooperberg, Stone and Truong (1995b) and Huang and Stone (1998).

CONDITION 1. (i) $P(C = \tau \mid \mathbf{X})$ *is bounded away from zero uniformly in* $\mathbf{X}$; (ii) *the density function of* $\mathbf{X}(t)$ *exists and is bounded away from zero and infinity on* $\mathscr{X}$ *uniformly over* $t \in [0, \tau]$; (iii) *the conditional log-hazard* $\log \lambda(t \mid \mathbf{X})$ *is bounded uniformly over* $t \in [0, \tau]$ *and* $\mathbf{X}$.

Set $Z(t) = \operatorname{ind}(Y \ge t)$ and $N(t) = \operatorname{ind}(\delta = 1 \text{ and } T \le t) = \operatorname{ind}(T \le t \text{ and } T \le C)$. It follows from Condition 1 that $E(Z(t)|\mathbf{X}) = P(Y \ge t|\mathbf{X}) = P(T \ge t \mid \mathbf{X})P(C \ge t \mid \mathbf{X})$ is bounded away from zero uniformly over $t \in [0, \tau]$.

Consider a random sample $(T_1, C_1, \mathbf{X}_1), \dots, (T_n, C_n, \mathbf{X}_n)$ from the distribution of $(T, C, \mathbf{X})$. For $1 \le i \le n$, set $Y_i = \min(T_i, C_i)$, $\delta_i = \operatorname{ind}(T_i \le C_i)$, $Z_i(t) = \operatorname{ind}(Y_i \ge t)$, and $N_i(t) = \operatorname{ind}(\delta_i = 1 \text{ and } T_i \le t) = \operatorname{ind}(T_i \le t \text{ and } T_i \le C_i)$. Also, set $\overline{N}(t) = n^{-1} \sum_i N_i(t)$. Then, up to a term that does not depend on $h$, the normalized partial log-likelihood $\ell(h)$ corresponding to the candidate $h$ for the log relative risk function can be written as

$$\ell(h) = \frac{1}{n} \sum_i \int_0^\tau h(\mathbf{X}_i(t))\, dN_i(t) - \int_0^\tau \log\left[\frac{1}{n} \sum_i Z_i(t) \exp h(\mathbf{X}_i(t))\right] d\overline{N}(t) \tag{3}$$

[see Cox (1972), Andersen and Gill (1982) and O'Sullivan (1993)]. The asymptotic value of $\ell(h)$ as $n \to \infty$ is given by

$$\Lambda(h) = E \int_0^\tau h(\mathbf{X}(t))\, dN(t) - \int_0^\tau \log E[Z(t) e^{h(\mathbf{x}(t))}] dEN(t), \tag{4}$$

which we refer to as the expected partial log-likelihood. It follows as in the proof in Section A.1 of the second conclusion of Theorem 1 below that, under the assumption of the proportional hazards model, the actual log relative risk function $\alpha$ maximizes $\Lambda(\cdot)$ over all integrable functions $h$.

Let $\mathbb{H}_0$, referred to as the *model space*, be a finite- or infinite-dimensional linear space of integrable functions on $\mathscr{X}$. For convenience in starting the identifiability constraint, we require that this linear space contain the constant functions. We envision that the space $\mathbb{H}_0$ incorporates structural, assumptions on $\alpha$ and we construct our estimate pretending that $\alpha$ is a member of $\mathbb{H}_0$. For example, the original Cox model for the log relative risk function amounts to choosing $\mathbb{H}_0$ to be the space of linear functions on $\mathscr{X}$. The additive model for the log relative risk function amounts to choosing $\mathbb{H}_0$ to be the space of integrable functions of the form $h_1(x_1) + \cdots + h_L(x_L)$, where $\mathscr{X} = \mathscr{X}_1 \times \cdots \times \mathscr{X}_L$ and $x_l$ ranges over $\mathscr{X}_l$ for $1 \le l \le L$. Appropriate choices of $\mathbb{H}_0$ also yield functional ANOVA models that will be discussed in detail in Section 2.3.

Note that $\Lambda(h) = \Lambda(h + c)$ for $c \in \mathbb{R}$ and $h \in \mathbb{H}_0$, so the space $\mathbb{H}_0$ is not identifiable. We impose the following identifiability constraint on functions $h \in \mathbb{H}_0$:

$$(5) \qquad \int_0^\tau \frac{E[h(\mathbf{X}(t))Z(t)]}{E[Z(t)]}\, dE[N(t)] = 0.$$

The identifiability constraint can be imposed in different ways. However, the way used here makes it convenient to study functional ANOVA models (see Section 2.3).

Set $\mathbb{H} = \{h \in \mathbb{H}_0\colon h \text{ satisfies } (5)\}$. The theorem below states that there is an essentially unique function $\alpha^* \in \mathbb{H}$ that maximizes the expected partial log-likelihood over $\mathbb{H}$. When the proportional hazards assumption is valid but $\alpha$ is not a member of $\mathbb{H}$, we think of $\alpha^*$ as the "best" approximation in $\mathbb{H}$ to $\alpha$. In general, we think of $\alpha^*$ as the target of our estimate regardless the validity of the proportional hazards assumption.

THEOREM 1. *Suppose* $\mathbb{H}_0$ *is closed in the following sense*: *if* $h_n \in \mathbb{H}_0$, $h$ *is an integrable function on* $\mathscr{X}$, *and* $h_n \to h$ *in measure as* $n \to \infty$, *then* $h \in \mathbb{H}_0$. *In addition*, *suppose Condition* 1 *holds*. *Then there exists an essentially uniquely determined function* $\alpha^* \in \mathbb{H}$ *such that* $\Lambda(\alpha^*) = \max_{h\in\mathbb{H}} \Lambda(h)$. *If the proportional hazards assumption is valid with* $\alpha \in \mathbb{H}$, *then* $\alpha^* = \alpha$ *almost everywhere*.

In the statement of this theorem, "essentially uniquely determined" means that any two such functions are equal almost everywhere relative to Lebesgue measure on $\mathscr{X}$. The proof of the theorem will be given in Section 7.1. The closedness requirement on $\mathbb{H}_0$ is obviously satisfied by the original Cox model. According to Lemma 4.1 of Stone (1994), additive models, or more generally functional ANOVA models (Section 2.3), also satisfy this closedness requirement.

Before proceeding further, it is convenient to introduce some notations. For any function $h$ on $\mathscr{X}$, set $\|h\|_\infty = \sup_{\mathbf{x}\in\mathscr{X}} |h(\mathbf{x})|$ and if $h$ is square-integrable, set $\|h\|_{L_2}^2 = \int_{\mathscr{X}} h^2(\mathbf{x})\, d\mathbf{x}$. Given positive numbers $a_n$ and $b_n$ for $n \ge 1$, let $a_n \lesssim b_n$ mean that $a_n/b_n$ is bounded and let $a_n \sim b_n$ mean that $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Given random variables $W_n$ for $n \ge 1$, let $W_n = O_P(b_n)$ mean that $\lim_{c\to\infty} \limsup_n P(|W_n| \ge cb_n) = 0$ and let $W_n = o_P(b_n)$ mean that $\limsup_n P(|W_n| \ge cb_n) = 0$ for all $c > 0$. For a random variable $V$, let $E_n$ denote the expectation relative to its empirical distribution, that is, $E_n(V) = n^{-1}\sum_i V_i$, where $V_i$, $1 \le i \le n$, is a random sample from the distribution of $V$. We use $M, M_1, M_2, \ldots,$ to denote generic constants which may vary from one context to another.

2.2. *Maximum partial likelihood estimation.* From now on we restrict our attention to square-integrable functions. Let us first introduce the theoretical and empirical inner products on the space of such functions. The corresponding norms will be used to measure the distance between the estimate and the target function. These inner products will play a critical role in the study of functional ANOVA models below. For a square-integrable functions $h_1$ and $h_2$ on $\mathscr{X}$, set

$$(E_n^Z h_1 h_2)(t) = \frac{E_n[h_1(\mathbf{X}(t))h_2(\mathbf{X}(t))Z(t)]}{E_n[Z(t)]}$$

and

$$(E^Z h_1 h_2)(t) = \frac{E[h_1(\mathbf{X}(t))h_2(\mathbf{X}(t))Z(t)]}{E[Z(t)]}.$$

Define the empirical and theoretical inner products by

$$\langle h_1, h_2\rangle_n = \int_0^\tau (E_n^Z h_1 h_2)(t)\, d\bar{N}(t) \quad \text{and} \quad \langle h_1, h_2\rangle = \int_0^\tau (E^Z h_1 h_2)(t)\, dE[N(t)].$$

The corresponding norms are given by $\|h\|_n^2 = \langle h, h\rangle_n$ and $\|h\|^2 = \langle h, h\rangle$. It will be shown in Lemma 1 (Section 5) that $\|\cdot\|$ is equivalent to $\|\cdot\|_{L_2}$, the $L_2$ norm relative to Lebesgue measure on $\mathscr{X}$. Note that the empirical inner product and norm are determined by the data and thus do not depend on unknown quantities. For a square-integrable function $h$, the identifiability constraint (5) can be written in terms of the theoretical inner product as $\langle h, 1\rangle = 0$.

We estimate $\alpha$ by using maximum partial likelihood over an appropriately chosen space. Let $\mathbb{G}_0 = \mathbb{G}_{0,n} \subset \mathbb{H}_0$ be a finite-dimensional linear space of bounded functions on $\mathscr{X}$, which we refer to as the *estimation space*. It is assumed that this space contains all constant functions and hence has dimension $N_n \geq 1$. We also require that it be *theoretically identifiable*: if $g \in \mathbb{G}_0$ and $\|g\| = 0$, then $g$ identically equals zero. This requirement is used to rule out the pathological cases. The space $\mathbb{G}_0$ is said to be *empirically identifiable* if $g \in \mathbb{G}_0$ and $\|g\|_n = 0$ implies that $g$ identically equals zero.

Note that $\ell(g) = \ell(g + c)$ for $c \in \mathbb{R}$ and $g \in \mathbb{G}_0$. We need to restrict our attention to a subspace of the estimation space to get a unique maximizer of $\ell(\cdot)$. To this end, an identifiability constraint is introduced similar to that for the model space, but defined in terms of the empirical inner product so that it is determined by the data. Specifically, set $\mathbb{G} = \{g \in \mathbb{G}_0\colon \langle g, 1\rangle_n = 0\}$. Also, set $\hat{\alpha}_n = \operatorname{argmax}_{g\in\mathbb{G}} \ell(g)$, which we refer to as the maximum partial likelihood estimate.

The function $\alpha^*$, shown to exist in Theorem 1, is guaranteed to be integrable, but not necessarily square-integrable or smooth. In the context of the next theorem we require that $\alpha^*$ be bounded and we set $\rho_n = \inf_{g\in\mathbb{G}_0} \|g - \alpha^*\|_\infty$ and $A_n = \sup_{g\in\mathbb{G}_0}\{\|g\|_\infty / \|g\|_{L_2}\} \geq 1$.

THEOREM 2. *Suppose Condition* 1 *holds and that*

$$\lim_n A_n \rho_n = 0 \quad \text{and} \quad \lim_n A_n^2 \max(N_n, \log n)/n = 0.$$

*Then, except on an event whose probability tends to zero as $n \to \infty$, $\mathbb{G}_0$ is empirically identifiable and $\hat{\alpha}_n$ exists and is uniquely defined. Moreover, $\|\hat{\alpha}_n - \alpha^*\|^2 = O_P(\rho_n^2 + N_n/n)$ and $\|\hat{\alpha}_n - \alpha^*\|_n^2 = O_P(\rho_n^2 + N_n/n)$.*

REMARKS.

1. If $\mathbb{H}_0$ is the space of linear functions on $\mathscr{X}$, we can choose $\mathbb{G}_0 = \mathbb{H}_0$. Then $A_n$ and $N_n$ are constants not depending on $n$ and $\rho_n = 0$. Applying Theorem 2, we get that $\|\hat{\alpha}_n - \alpha^*\|^2 = O_P(1/n)$, which is the parametric rate of convergence.
2. For nonparametric cases, the estimation space $\mathbb{G}_0$ is usually chosen such that $\log n \lesssim N_n$; if so, then the assumption $\lim_n A_n^2 \max(N_n, \log n)/n = 0$ reduces to $\lim_n A_n^2 N_n/n = 0$.
3. The conclusion that $\mathbb{G}_0$ is empirically identifiable does not depend on the assumption that $\lim_n A_n \rho_n = 0$; see Lemma 3 in Section 5.1.

Theorem 2 gives a very general treatment of the rate of convergence for the maximum partial likelihood estimate, which is parallel to the result on least squares estimation for the regression context established in Theorem 1 of Huang (1998a). The two terms that govern the magnitude of the error $\hat{\alpha}_n - \alpha^*$ have intuitively appealing explanations: $N_n/n$ is just the inverse of the number of observations per parameter, and $\rho_n$ is the best possible approximation rate in the estimation space to the target function. In the statement of the theorem, the constant $A_n$ is a measure of the irregularity of the estimation space, while for a specific choice of the estimation space, the rate of decay of $\rho_n$ reflects a smoothness assumption on $\alpha^*$. As elaborated in the cited paper, the magnitudes of the constants $A_n$ and $\rho_n$ are easily found for the linear estimation spaces that are commonly used in approximation theory. As a consequence, we can obtain rates of convergence when the estimation spaces are built up from polynomials, trigonometric polynomials, and polynomial splines. We will use this theorem to study the rates of convergence for functional ANOVA models in the next subsection.

2.3. *Functional ANOVA models.* In this subsection we first give the precise definition of functional ANOVA decompositions and then state the main result for functional ANOVA models. The ANOVA decomposition of the target function is constructed in such a way that each nonconstant component is orthogonal to the proper lower-order components relative to the theoretical inner product defined in the previous section. The ANOVA decomposition of the estimate is defined by a similar orthogonality requirement relative to the empirical inner product defined in that section. We shall see that not only does the estimate converge to the target function, but the components of the estimate also converge to the corresponding components of the target function. In addition, we shall see that, when the ANOVA model for the target function is invalid, the estimate will converge to its best approximation having the indicated ANOVA form. This result is important since in practice the functional ANOVA model is usually only an approximation.

Suppose $\mathscr{X}$ is the Cartesian product of compact sets $\mathscr{X}_1, \dots, \mathscr{X}_L$. For $\mathbf{x} \in \mathscr{X}$, write $\mathbf{x} = (x_1, \dots, x_L)$, where $x_l \in \mathscr{X}_l$ for $1 \le l \le L$. It is convenient to use subset notation to denote the various components in an ANOVA decomposition. Let $\mathbb{H}_\varnothing$ denote the space of constant functions on $\mathscr{X}$. Given a nonempty subset $s$ of $\{1, \dots, L\}$, let $\mathbb{H}_s$ denote the space of square-integrable functions on $\mathscr{X}$ that depend only on the variables $x_l$, $l \in s$. Also, set

$$\mathbb{H}_s^0 = \{h \in \mathbb{H}_s\colon h \perp \mathbb{H}_r \text{ for every proper subset } r \text{ of } s\};$$

here $h \perp \mathbb{H}_r$ means that $\langle h, h_r \rangle = 0$ for $h_r \in \mathbb{H}_r$.

Let $\mathscr{S}_0$ be a nonempty collection of subsets of $\{1, \dots, L\}$. It is assumed that $\mathscr{S}_0$ is *hierarchical*: if $s \in \mathscr{S}_0$ and $r \subset s$, then $r \in \mathscr{S}_0$. Set $\mathbb{H}_0 = \{\sum_{s \in \mathscr{S}_0} h_s\colon h_s \in \mathbb{H}_s\}$. Under Condition 1, every function $h \in \mathbb{H}_0$ can be written in an essentially unique manner as $\sum_{s \in \mathscr{S}_0} h_s$, where $h_s \in \mathbb{H}_s^0$ [see Lemma 3.1 of Stone (1994)]. We refer to $\sum_{s \in \mathscr{S}} h_s$ as the *theoretical ANOVA decomposition* of $h$ corresponding to the inner product $\langle \cdot, \cdot \rangle$ and to $h_s$, $s \in \mathscr{S}$ as the components of $h$ is this decomposition. The component $h_s$ is referred to as the constant component if $\#(s) = 0$, as a main effect component if $\#(s) = 1$ and as an interaction component if $\#(s) \ge 2$; here, $\#(B)$ denote the cardinality (number of members) of a set $B$.

In order to use $\mathbb{H}_0$ to model the log relative risk function, we impose the identifiability constraint (5). This turns out to be very convenient, since imposing (5) is equivalent to setting the constant component in the theoretical ANOVA decomposition of $h \in \mathbb{H}_0$ to be zero. (This is actually the motivation for choosing that particular form of identifiability constraint.) Now, set $\mathscr{S} = \mathscr{S}_0 \setminus \{\varnothing\}$ and $\mathbb{H} = \{h \in \mathbb{H}_0\colon h \perp 1\} = \{\sum_{s \in \mathscr{S}} h_s\colon h_s \in \mathbb{H}_s^0\}$. We say that $\mathscr{S}$ specifies a functional ANOVA model. Different choices of $\mathscr{S}$ provide different models. For example, choosing $\mathscr{S} = \{\{1\}, \dots, \{L\}\}$ gives an additive model. A square-integrable function given by (1) could be described as a member of $\mathbb{H}$ by setting $\mathscr{S} = \{\{1\}, \{2\}, \{3\}, \{1, 2\}\}$.

Let $\mathbb{G}_\varnothing$ denote the space of constant functions on $\mathscr{X}$, which has dimension $N_\varnothing = 1$. Given $l \in \{1, \dots, L\}$, let $\mathbb{G}_l \subset \mathbb{H}_l$ denote a linear space of bounded functions, which can vary with the sample size and has finite, positive dimension $N_l$. Given a subset $s$ of $\{1, \dots, L\}$, let $\mathbb{G}_s$ denote the tensor product of $\mathbb{G}_l$, $l \in s$, which is the linear space of functions on $\mathscr{X}$ spanned by functions $g$ of the form $g(\mathbf{x}) = \prod_{l \in s} g_l(x_l)$, where $g_l \in \mathbb{G}_l$ for $l \in s$. Also, let $N_s = \prod_{l \in s} N_l$ denote the dimension of $\mathbb{G}_s$, set $N_s^0 = \prod_{l \in s}(N_l - 1)$ and observe that $N_s = \sum_{r \subset s} N_r^0$. Set

$$\mathbb{G}_s^0 = \{g \in \mathbb{G}_s\colon g \perp_n G_r \text{ for every proper subset } r \text{ of } s\};$$

here $g \perp_n \mathbb{G}_r$ means that $\langle g, g_r \rangle_n = 0$ for $g \in \mathbb{G}_r$. If $\mathbb{G}_s$ is identifiable, then $\mathbb{G}_s^0$ is identifiable and it has dimension $N_s^0$; moreover, $\mathbb{G}_s$ is the direct sum of $\mathbb{G}_r^0$, $r \subset s$.

Consider the estimation space $\mathbb{G}_0 = \{\sum_{s \in \mathscr{S}_0} g_s\colon g_s \in \mathbb{G}_s \text{ for } s \in \mathscr{S}_0\}$, which has dimension $N_n = \sum_{s \in \mathscr{S}_0} N_s^0$. Observe that $\max_{s \in \mathscr{S}_0} N_s \le N_n \le \#(\mathscr{S}_0) \max_{s \in \mathscr{S}_0} N_s$, where $\#(\mathscr{S}_0)$ denotes the number of members of $\mathscr{S}_0$. Suppose $\mathbb{G}_0$ is empirically identifiable. Then $\mathbb{G}_s$ is empirically identifiable for

$s \in \mathscr{S}_0$ and $\mathbb{G}_0$ is the direct sum of $\mathbb{G}_s^0$, $s \in \mathscr{S}_0$. We refer to $g = \sum_{s \in \mathscr{S}_0} g_s$, where $g_s \in \mathbb{G}_s^0$ for $s \in \mathscr{S}_0$, as the *empirical ANOVA decomposition* of $g \in \mathbb{G}_0$. Set $\mathbb{G} = \{g \in \mathbb{G}_0 \colon g \perp_n 1\}$, which is the direct sum of $\mathbb{G}_s^0$, $s \in \mathscr{S}$.

Suppose the maximum partial likelihood estimate $\hat{\alpha}_n$ in $\mathbb{G}$ exists and write its empirical ANOVA decomposition as $\sum_{s \in \mathscr{S}} \hat{\alpha}_s$. Suppose also that $\alpha^* = \operatorname{argmax}_{h \in \mathbf{H}} \Lambda(h)$ exists as a member of $\mathbb{H}$ and write its theoretical ANOVA decomposition as $\sum_{s \in \mathscr{S}} \alpha_s^*$. The similarity of the structures of $\mathbb{G}$ and $\mathbb{H}$ is crucial in ensuring the convergence of $\hat{\alpha}_s$ to $\alpha_s^*$ for $s \in \mathscr{S}$.

In the next theorem it is required that the components $\alpha_s^*$, $s \in \mathscr{S}$, of the target function $\alpha^*$ be bounded. Set $\rho_s = \inf_{g \in \mathbb{G}_s} \|g - \alpha_s^*\|_\infty$ for $s \in \mathscr{S}$ and $A_s = \sup_{g \in \mathbb{G}_s} \{\|g\|_\infty / \|g\|_{L_2}\} \geq 1$, and set $\bar{\rho}_n = \sum_{s \in \mathscr{S}} \rho_s$.

THEOREM 3. *Suppose Condition* 1 *holds and that*

$$\lim_n A_s \rho_{s'} = 0 \quad \text{and} \quad \lim_n A_s^2 \max(N_{s'}, \log n)/n = 0, \qquad s, s' \in \mathscr{S}.$$

*Then, except on an event whose probability tends to zero as $n \to \infty$, $\mathbb{G}_0$ is empirically identifiable and $\hat{\alpha}_n$ exists. Moreover, $\|\hat{\alpha}_n - \alpha^*\|^2 = O_P(\bar{\rho}_n^2 + N_n/n)$, $\|\hat{\alpha}_n - \alpha^*\|_n^2 = O_P(\bar{\rho}_n^2 + N_n/n)$ and $\|\hat{\alpha}_s - \alpha_s^*\|^2 = O_P(\bar{\rho}_n^2 + N_n/n)$ and $\|\hat{\alpha}_s - \alpha_s^*\|_n^2 = O_P(\bar{\rho}_n^2 + N_n/n)$ for $s \in \mathscr{S}$.*

The results in the above theorem about the convergence properties of the maximum partial likelihood estimate and its ANOVA components are parallel to that in Huang (1998a) for the regression context. Using this theorem, we can obtain rate of convergence results when polynomials, trigonometric polynomials, or splines and their tensor products are used as building blocks for the estimation spaces. To get such results, we need only find upper bounds for the constants $A_s$ and $\rho_s$ by employing results from approximation theory literature. We shall illustrate this in the next subsection when the estimation spaces are built from spline functions.

Note that we do not restrict any of the explanatory variables $x_1, \ldots, x_L$ to be one-dimensional, so any of the main effect components (components depending on one variable) in the ANOVA decomposition of the log relative risk function could be bivariate or multivariate. This provides additional flexibility in functional ANOVA modeling.

Suppose $\mathscr{X}_l \subset \mathbb{R}^{d_l}$ with $d_l \geq 1$. Set $d = \max_{s \in \mathscr{S}} \sum_{l \in s} d_l$. If $d_l = 1$ for $1 \leq l \leq L$, then $d = \max_{s \in \mathscr{S}} \#(s)$. Typically, the spaces $\mathbb{G}_l$ are chosen such that $\bar{\rho}_n \asymp N_n^{-p/d}$ and $N_n \asymp n^{d/(2p+d)}$, where $p$ is a suitable defined measure of smoothness of $\alpha_s^*$. Correspondingly, the rate of convergence in the theorem is given by $\bar{\rho}_n^2 + N_n/n = O(n^{-2p/(2p+d)})$, which is of the standard form; see Stone (1982, 1994) and Huang (1998a).

2.4. *Application to spline estimation.* Let $\mathscr{X}$ be the Cartesian product of compact intervals $\mathscr{X}_1, \ldots, \mathscr{X}_L$ in $\mathbb{R}$. Without further loss of generality, we assume that each of these intervals equals [0, 1]. Let $0 < \beta \leq 1$. A function $h$ on $\mathscr{X}$ is said to satisfy a Hölder condition with exponent $\beta$ if there is a positive number $\gamma$ such that $|h(\mathbf{x}) - h(\mathbf{x}_0)| \le \gamma|\mathbf{x} - \mathbf{x}_0|^{\beta}$ for $\mathbf{x}_0, \mathbf{x} \in \mathscr{X}$; here $|\mathbf{x}| = (\sum_{l=1}^{L} x_l^2)^{1/2}$ is the Euclidean norm of $\mathbf{x} = (x_1, \dots, x_L) \in \mathscr{X}$. Given an $L$-tuple $i = (i_1, \dots, i_L)$ of nonnegative integers, set $[i] = i_1 + \cdots + i_L$ and let $D^i$ denote the differential operator defined by

$$D^i = \frac{\partial^{[i]}}{\partial x_1^{i_1} \cdots \partial x_L^{i_L}}.$$

Let $k$ be a nonnegative integer and set $p = k + \beta$. A function on $\mathscr{X}$ is said to be *p-smooth* if it is $k$ times continuously differentiable on $\mathscr{X}$ and $D^i$ satisfies a Hölder condition with exponent $\beta$ for all $i$ with $[i] = k$.

Let $J$ be a positive integer, and let $t_0, t_1, \dots, t_J, t_{J+1}$ be real numbers with $0 = t_0 < t_1 < \cdots < t_J < t_{J+1} = 1$. Partition [0, 1] into $J + 1$ subintervals $I_j = [t_j, t_{j+1})$, $j = 0, \dots, J-1$, and $I_J = [t_J, t_{J+1}]$. Let $m \ge 0$ be an integer. A function on [0, 1] is a spline of degree $m$ with knots $t_1, \dots, t_J$ if the following hold: (i) it is a polynomial of degree $m$ or less on each interval $I_j$, $j = 0, \dots, J$; (ii) (for $m \ge 1$) it is $(m-1)$-times continuously differentiable on [0, 1]. Such spline functions constitute a linear space of dimension $J + m + 1$. A spline of degree $m = 0$ is just a piecewise constant function. The splines are called linear, quadratic or cubic splines according as $m = 1, 2$ or 3. For detailed discussions of univariate splines, see de Boor (1978), Schumaker (1981) and DeVore and Lorentz (1993).

Assume that the components $\alpha_s^*, s \in \mathscr{S}$ of the target function $\alpha^*$ are $p$-smooth for some positive number $p$. Let $m \ge p - 1$ be an integer. For $l = 1, \dots, L$, let $\mathbb{G}_l$ be the space of splines of degree $m$ with $J = J_n$ knots. Suppose that

$$\frac{\max_{0 \le j \le J-m}(t_{j+m+1} - t_j)}{\min_{0 \le j \le J-m}(t_{j+m+1} - t_j)} \le \eta \tag{6}$$

for some positive constant $\eta$. Then $A_s \asymp J^{\#(s)}/2$, $N_s \asymp J^{\#(s)}$ and $\rho_s \asymp J^{-p}$ for $s \in \mathscr{S}$; see Huang (1998a).

Set $d = \max_{s \in \mathscr{S}} \#(s)$ and suppose that $p > d/2$ and $J^{2d} = o(n)$. Then the conditions in Theorem 3 are satisfied. Thus, $\|\hat{\alpha}_s - \alpha_s^*\|^2 = O_p(J^d/n + J^{-2p})$ for $s \in \mathscr{S}$ and $\|\hat{\alpha}_n = \alpha^*\|^2 = O_P(J^d/n + J^{-2p})$. Taking $J \asymp n^{1/(2p+d)}$, we get that $\|\hat{\alpha}_s - \alpha_s^*\|^2 = O_P(n^{-2p/(2p+d)})$ for $s \in \mathscr{S}$ and $\|\hat{\alpha}_n - \alpha^*\|^2 = O_P(n^{-2p/(2p+d)})$. The rate $n^{-2p/(2p+d)}$ is the optimal rate for estimating a $p$-smooth, $d$-dimensional function; see Stone (1982). This result shows that, by using models with only main effects and low-order interactions, we can ameliorate the curse of dimensionality that adversely affects the saturated model ($d = L$). For instance, by considering additive models ($d = 1 < L$) or by allowing interactions involving only two factors ($d = 2 < L$), we can get faster rates of convergence than by using the saturated model. The results in this section are parallel to those in Kooperberg, Stone and Truong (1995b) and Huang and Stone (1998), where the baseline hazard is modeled as a smooth function and the maximum (full) likelihood estimate is used.

**3. Software implementation and examples.** For a fixed linear estimation space $\mathbb{G}$, the problem of finding the maximum partial likelihood estimate is reduced to the estimation of finite number of coefficient parameters after a basis of $\mathbb{G}$ is chosen. Specifically, given a collection $\{B_j, j = 1, \dots, p\}$ of basis functions that span $\mathbb{G}$, the estimate has the form

$$\hat{\alpha}(\mathbf{X}(t)) = \sum_{j=1}^{p} \hat{\beta}_j B_j(\mathbf{X}(t)), \tag{7}$$

where the coefficients $\beta_1, \dots, \beta_P$ are estimated from the available data by maximum partial likelihood. Thus, if the set of basis functions or, equivalently, the estimation space is manually selected by the user, then the estimate can be obtained by using any statistical package that can fit parametric proportional hazards models by defining new "covariates" that are equal to the basis functions. In practice, the manual selection of appropriate spline basis functions by users is typically infeasible, so additional code for the automatic, adaptive selection of such basis functions is required.

In the following we will describe an adaptive implementation of proportional hazards regression (PHARE), which we have designed as similarly as possible to the hazard regression (HARE) implementation of Kooperberg, Stone and Truong (1995a). In this implementation, linear splines and their tensor products are used to build the estimation spaces. We assume throughout the remainder of this section that the covariates are all time independent. (Time-dependent covariates increase the complexity of the program, but they do not make the calculations more time-consuming.)

*Adaptive implementation.* In the adaptive implementation, one important issue to be resolved is the choice of the linear space $\mathbb{G}$. Following Kooperberg, Stone and Truong (1995a) and Stone, Hansen, Kooperberg and Truong (1997), we select $\mathbb{G}$ from a family $\mathscr{G}$ of allowable spaces through a process of stepwise addition and deletion of basis functions. We refer the reader to these two papers for motivation of the allowable spaces and further discussion of the stepwise procedure; here we restrict ourselves to listing the basis functions that we consider and indicating when these basis functions can be in $\mathbb{G}$. As in HARE, we require that the basis functions all depend on at most two of the covariates. Let $(x)_+ = x$ when $x > 0$ and 0 otherwise. The basis functions that we consider are:

1. Basis functions that are linear in one of the covariates, $B(\mathbf{X}) = X_j$, can always be in $\mathbb{G}$.
2. Basis functions that model a knot in one of the covariates, $B(\mathbf{X}) = (X_j - r)_+$, for some value of $r$ within the range of $X_j$ are only allowed in $\mathbb{G}$ when $B(\mathbf{X}) = X_j$ is also in $\mathbb{G}$.
3. Basis functions that are tensor products of two different linear basis functions, $B(\mathbf{X}) = X_{j_1} X_{j_2}$, are only allowed in $\mathbb{G}$ when $B(\mathbf{X}) = X_{j_1}$ and $B(\mathbf{X}) = X_{j_2}$ are also in $\mathbb{G}$.
4. Basis functions that are a tensor product of a linear basis function and a basis function depending on a knot in a different covariates, $B(\mathbf{X}) = X_{j_1}$

$(X_{j_2}-r)_+$, are only allowed in $\mathbb{G}$ when $B(\mathbf{X}) = X_{j_1}X_{j_2}$ and $B(\mathbf{X}) = (X_{j_2}-r)_+$ are also in $\mathbb{G}$.

5. Basis functions that are a tensor product of two basis functions depending on knots in different covariates, $B(\mathbf{X}) = (X_{j_1} - r_1) + (X_{j_2} - r_2)_+$, are only allowed in $\mathbb{G}$ when $B(\mathbf{X}) = (X_{j_1} - r_1)_+ X_{j_2}$ and $B(\mathbf{X}) = X_{j_1}(X_{j_2} - r_2)_+$ are also in $\mathbb{G}$.

In our adaptive implementation we initially fit the model with $p = 0$ and $\mathbb{G} = \{0\}$. Then we proceed with stepwise addition. Here we successively replace the $(p-1)$-dimensional allowable space $\mathbb{G}_0$ by a $p$-dimensional allowable space $\mathbb{G}$ containing $\mathbb{G}_0$ as a subspace, choosing among the various candidates for a new basis function by a heuristic search designed approximately to maximize the corresponding Rao (score) statistic; see Kooperberg, Stone and Truong (1995a) for details. Upon stopping the stepwise addition stage with $p = P_{\max}$ basis functions we proceed with stepwise deletion. Here we successively replace the $p$-dimensional allowable space $\mathbb{G}$ by a $(p-1)$-dimensional allowable subspace $\mathbb{G}_0$ until we arrive at $p = 0$, at each step choosing the candidate space $\mathbb{G}_0$ so that the Wald statistic for a basis function that is in $\mathbb{G}$ but not in $\mathbb{G}_0$ is smallest in magnitude.

During the combination of stepwise addition and deletion, we get a sequence of models indexed by $\nu$, with the $\nu$th model having $p_\nu$ parameters. Let $\hat{\ell}_\nu$ be the fitted partial log-likelihood for the $\nu$th model, and let $\text{AIC}_{\alpha,\nu} = -2\hat{\ell}_\nu + ap_\nu$ be the AIC with penalty parameter $a$. We select the model corresponding to the value of $\nu$ that minimizes $\text{AIC}_{\alpha,\nu}$. As for HARE we take $a = \log n$, as in BIC [Schwarz (1978)].

By design, the PHARE methodology under consideration is very similar to the HARE methodology, the main differences being that HARE also models the (baseline) hazard function and the parameters in HARE are estimated using the full likelihood function. In practice, this means that HARE also allows for basis functions that are linear splines in time and for basis functions that are tensor products of a basis function in time and a basis function in a covariate. When such tensor product basis functions are included in the model, it is no longer a proportional hazards model. The basis functions in HARE are constructed such that the conditional hazard function $\lambda(t \mid \mathbf{X})$ is constant in the right tail for all $\mathbf{X}$. See Kooperberg, Stone and Truong (1995a) for more details. Kooperberg and Clarkson (1997) extend the HARE methodology to include cubic splines, interval-censored data and time-dependent covariates.

An alternative way to fit proportional hazards model is to use existing HARE software, but to restrict the allowable models in HARE to exclude any tensor product basis functions that depend on time and at least one covariate, which we refer to as HARE-ph. Thus PHARE and HARE-ph have the same collection of basis functions to model the covariate effects, but the latter has additional candidate basis functions to model the baseline hazard. While the parameters in PHARE are estimated using maximum partial likelihood, those in HARE-ph are estimated by maximizing the full likelihood. In the examples below we will see that these two different ways of fitting proportional hazards regression models give surprisingly close results.

EXAMPLES. Our first example uses the breast cancer data of Gray (1992), which was also discussed in Section 9.3 of Kooperberg, Stone and Truong (1995a). The data come from six breast cancer studies conducted by the Eastern Cooperative Oncology Group. There were 2404 patients in this study, for whom the response is survival time (years) since surgery. There are six covariates: estrogen receptor status (ER 0 is negative, 1 is positive), the logarithm of the number of positive axillary lymph nodes, tumor size (in cm), age (in years), body mass index (BMI), and menopause (0 is premenopause, 1 is postmenopause). See Kooperberg, Stone and Truong (1995a) for more details.

Initially we applied three algorithms to the data: HARE, HARE-ph, and PHARE. The results are summarized in Table 1. Observe the similarity of the three fitted models. In particular, the only differences between HARE and HARE-ph are two basis functions that involve both time and a covariate that are in the HARE model but not in the HARE-ph model (in other examples, the proportionality restriction on HARE does lead to major changes in the other basis functions). The fact that nonproportional hazards models fit the data better than proportional hazards models was already observed by Gray (1992) and Kooperberg, Stone and Truong (1995a). The only difference between the HARE-ph model and the PHARE model is the location of the knot in age; even the coefficients for those basis functions that do not involve age or menopause, which is highly correlated with age, are very similar in the two models. When we move the knot from the location of one model to the location of the other model and refit all coefficients, the coefficients for HARE-ph and PHARE are indistinguishable: the largest difference in a coefficient is 0.0033 and the largest difference in a standard error is 0.00004.

Clearly small implementation differences can cause the algorithm that searches for a location of a knot [see Section 11.3 of Kooperberg, Stone and Truong (1995a)] to yield different knot locations. To examine further the effect of knot location, we fitted models similar to the HARE-ph model and the PHARE model in Table 1, in which we varied the location of the knot in age over the observed range of age in the data. In Figure 1 we plot the resulting (partial) log-likelihood as a function of the location of the knot in age; that is, for each value of the location of this knot we estimated all coefficients for a HARE-ph model and a PHARE model like the one in Table 1. We lined up the maximum value of the log-likelihood and the partial log-likelihood in this figure. Note that the two curves are almost identical. HARE-ph almost finds the location of the left mode, while PHARE finds the location of the right mode. The heights of the two modes are nearly equal, so we cannot say that one algorithm does a better job than the other. The dip between the two modes is explained by the presence of a menopause basis function in the model: this basis function is 0 for almost all women under 48, and 1 for almost all women over 55.

TABLE 1
*Summary of three different models for the breast cancer data*

| | HARE | | HARE-ph | | PHARE | |
|---|---|---|---|---|---|---|
| **Basis function** | **Coefficient** | **SE** | **Coefficient** | **SE** | **Coefficient** | **SE** |
| 1 | −3.4048 | 0.4279 | −2.5881 | 0.3990 | NA | NA |
| ER | 1.0600 | 0.2046 | −0.3001 | 0.0629 | −0.3019 | 0.0630 |
| log(nodes) | 0.6879 | 0.0700 | 0.6593 | 0.0700 | 0.6506 | 0.0698 |
| size | 0.1588 | 0.0354 | 0.1801 | 0.0348 | 0.1813 | 0.0347 |
| age | −0.0405 | 0.0092 | −0.0397 | 0.0092 | −0.0214 | 0.0052 |
| $(\text{age} - 43)_+$ | 0.0413 | 0.0115 | 0.0402 | 0.0115 | NA | NA |
| $(\text{age} - 65.25)_+$ | NA | NA | NA | NA | 0.0557 | 0.0152 |
| menopause | 0.4036 | 0.1050 | 0.4063 | 0.1049 | 0.6087 | 0.1075 |
| log(nodes) × size | −0.0653 | 0.0180 | −0.0562 | 0.0181 | −0.0545 | 0.0181 |
| $(0.44 - t)_+$ | −5.7430 | 1.0201 | −5.5785 | 1.0152 | NA | NA |
| $(1.89 - t)_+$ | −0.9662 | 0.1501 | −0.5145 | 0.0894 | NA | NA |
| $(7.95 - t)_+$ | 0.3653 | 0.0351 | 0.1892 | 0.0214 | NA | NA |
| $(1.89 - t)_+ \times \text{size}$ | 0.1050 | 0.0321 | NA | NA | NA | NA |
| $(7.95 - t)_+ \times \text{ER}$ | −0.2586 | 0.0360 | NA | NA | NA | NA |

Thus, for the breast cancer data, for which the underlying model appears to have nonproportional hazards, HARE-ph and PHARE give virtually identical results. This has been our experience with a number of examples: differences between the results from these two methods are either at the noise level or can be traced back to tiny implementation details that differ between the two methodologies.

One might suspect, however, that for more extreme nonproportional hazards models HARE-ph and PHARE would give different results. To investigate this suspicion we generated an extremely high signal to noise dataset (in survival analysis the signal to noise ratio is usually much lower). This dataset involves a single covariate $X$, which is uniformly distributed on [0.5, 2.0], and an uncensored survival time $T$, whose conditional distribution given that $X = x$ coincides with the distribution of $Z^x$ with $Z$ having the exponential distribution with mean 1 (which conditional distribution is a Weibull distribution). The corresponding conditional hazard function is given by

$$\log \lambda(t|X = x) = \left(\frac{1}{x} - 1\right) \log t - \log x.$$

The dataset consisted of a random sample of size $n = 10{,}000$ from the joint distribution of $T$ and $X$. To keep the set-up of this example simple, we did not censor the data.

We applied HARE, HARE-ph and PHARE to this dataset. In Figure 2 we show the true and the fitted hazard rates for the values 0.6, 1.0 and 1.8 of the covariate $X$. Note that both axes are logarithmic, so the true hazard function is a straight line. As can be seen, HARE does a good job of fitting the hazard functions, except in the extreme tails of the distribution. The HARE model includes five knots in time, a knot in $X$ and five basis functions that depend

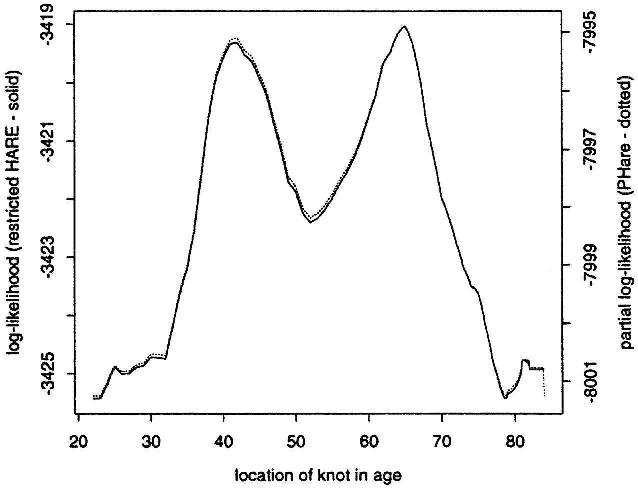

FIG. 1. *Log-likelihood for HARE-ph (solid) and partial log-likelihood for PHARE (dotted) as a function of the location of the knot in age for the model for the breast cancer data.*

on time and $X$. The largest knot in time is beyond the plotted region; beyond this last knot the fitted hazard functions are constant. The HARE-ph model is forced to fit a proportional hazards model, which clearly does not fit the data. The fitted HARE-ph model has six knots in time (again, the last one is beyond the plotted region), but $X$ occurs only linearly in the model. The coefficient of the basis function $X$ is $-0.2600$ with a standard error of 0.0244. The fitted PHARE model has $X$ linearly as its only basis function, models with knots in $X$ have higher AIC values. The coefficient of this linear basis function is $-0.2599$ with a standard error of 0.0245. Since PHARE does not yield an estimate for the baseline hazard function, we cannot provide a plot like that of Figure 2 for this method.

**4. Discussion.** The difference between our general proportional hazards model and Cox's original model is that the log relative risk function is now allowed to have a more flexible form than linear. The role of the baseline hazard in our model is similar to that in Cox's model. After the proportional hazards model is fit, one can use Breslow's estimator [Breslow (1972)] to estimate the baseline hazard.

There are two approaches in fitting a proportional hazards model: maximum partial likelihood with the baseline hazard function not modeled (PHARE) or, maximum full likelihood with the baseline hazard function modeled as a smooth function (HARE-ph). Theoretically speaking, the two approaches cannot be distinguished in terms of rates of convergence of the resulting estimates, but the first approach has the advantage of not requiring a smoothness assumption on the baseline hazard function. More precisely, provided that the

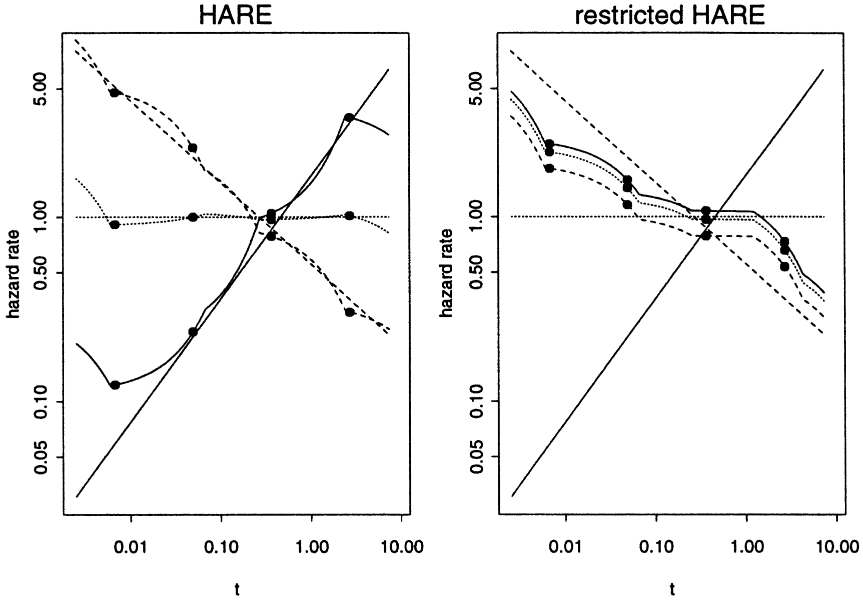

FIG. 2. *True and fitted* (*with bullets*) *hazard functions for the simulated data for three different values x of the covariate X*: $x = 0.6$ (*solid*), $x = 1.0$ (*dotted*) *and* $x = 1.8$ (*dashed*). (*Note that both axes are logarithmic.*)

baseline hazard is reasonably smooth, the same rates of convergence can be achieved by both approaches under the same smoothness conditions on the log relative risk function; compare the results in Section 2.4 with those in Kooperberg, Stone and Truong (1995b). This seems to be consistent with the observation in the data examples in the previous section. For both the breast cancer data and the simulated data the results that we obtained from PHARE and HARE-ph are extremely similar. Maybe this should have been anticipated since Breslow (1974) pointed out that if the hazard function is approximated by a step function with jumps at each observed failure time (which is a spline of order one), then the partial likelihood estimates are identical to the full likelihood estimates.

For proportional hazards models with parametric covariate effects, Efron (1977) showed through an information calculation that using Cox's partial likelihood has full asymptotic efficiency over using full likelihood. A similar result should hold in our nonparametric framework. However, to make a mathematically rigorous statement, one needs to establish an asymptotic distribution theory and especially to determine biases due to spline approximations. This is an interesting area for future research.

There are some computational advantages of PHARE over HARE-ph. An equivalent implementation of PHARE requires far less cpu time than that for HARE. Indeed, a count of the number of addition and multiplication operations suggests that for a dataset the size of the breast cancer data, HARE with linear splines would require 5–10 times as many floating operations as PHARE, due primarily to the integration that is required for HARE. (This integration may be done analytically, but requires many floating point operations. HARE with higher order splines requires numerical integration and is thereby much more computationally intense.) The intellectual, effort to write a program for PHARE is much smaller than that required to write one for HARE, as many more standard routines for fitting Cox models and computing corresponding score functions and Hessians are available. The subtleties of integration are an added complexity for HARE.

However, the computational advantages of PHARE over HARE may not be serious for a user that has access to a computationally efficient version of HARE. (HARE software, written in C with an interface to S-Plus is publically available from the second author `http://bear.fhcrc.org/~clk/soft.html`; PHARE software is currently not available.) Perhaps the main advantage of PHARE over HARE is its familiarity. There is a large body of literature on using Cox's partial likelihood. Within the biomedical community using partial likelihood is conventional, so a statistician may have an easier time convincing a scientist who is interested only in covariate effects to use or accept results from PHARE than from HARE. There are also several advantages of HARE over PHARE. When a (smooth) estimate for the baseline hazard function is desired, using HARE-ph (rather than PHARE and some post hoc technique to estimate a smooth baseline hazard) makes much more sense. In addition, HARE can be used to examine the assumption of proportional hazards.

**5. Rates of convergence.** We prove Theorem 2 in this section and assume that the conditions of this theorem hold throughout. We decompose the error into approximation error and estimation error and then treat them separately. Recall that $\hat{\alpha}_n = \operatorname{argmax}_{g\in\mathbb{G}} \ell(g)$. Since $\mathbb{G}$ is random and may not be contained in $\mathbb{H}$, it is convenient to introduce a proxy of $\mathbb{G}$ in $\mathbb{H}$: $\widetilde{\mathbb{G}} = \{g \in \mathbb{G}_0,\ \langle g, 1\rangle = 0\} \subset \mathbb{H}$. When it exists, we refer to $\alpha_n^* = \operatorname{argmax}_{g\in\widetilde{\mathbb{G}}} \Lambda(g)$ as the best approximation to $\alpha$ in $\widetilde{\mathbb{G}}$. Recall that $\alpha^* = \operatorname{argmax}_{h\in\mathbb{H}} \Lambda(h)$. We have the decomposition $\hat{\alpha}_n - \alpha^* = (\alpha_n^* - \alpha^*) + (\hat{\alpha}_n - \alpha_n^*)$, where the first term on the right side of this equation is referred to as the approximation error and the second term as the estimation error.

5.1. *Preliminary lemmas.* This section collects some useful lemmas. To simplify our presentation, we introduce two ancillary inner products and norms in parallel to the theoretical and empirical ones defined in Section 2. Specifically, we set

$$\langle f_1, f_2\rangle_{0n} = \int_0^\tau E_n^Z[(f_1 - E_n^Z f_1)(f_2 - E_n^Z f_2)]\, d\overline{N}$$

and

$$\langle f_1, f_2\rangle_0 = \int_0^\tau E^Z[(f_1 - E^Z f_1)(f_2 - E^Z f_2)]\, dEN$$

The corresponding squared norms are given by $\|f\|_{0n}^2 = \langle f, f\rangle_{0n}$ and $\|f\|_0^2 = \langle f, f\rangle_0$. In the proof of Theorem 2, these two norms will serve as bridges to build the relationships between the Hessian of the expected log-likelihood and the theoretical norm and between the Hessian of the log-likelihood and the empirical norm.

The following results on the equivalence of norms play important roles in the proofs of Theorems 2 and 3. Lemma 1 establishes the equivalence of the theoretical norm and the $L_2$ norm on $\mathbb{H}_0$. Lemma 2 establishes the equivalence of the ancillary theoretical norm and the $L_2$ norm on $\mathbb{H}$.

LEMMA 1. *Under Condition* 1, $\|f\| \sim \|f\|_{L_2}$ *uniformly in* $f \in \mathbb{H}_0$.

PROOF. It follows from Condition 1 that $E[Z(t)] \sim 1$ and

$$E[f^2(\mathbf{X}(t))Z(t)] = E[f^2(\mathbf{X}(t))E(Z(t)|\mathbf{X})] \sim E[f^2(\mathbf{X}(t))] \sim \|f\|_{L_2}^2.$$

Consequently,

$$\|f\|^2 = \int_0^\tau \frac{E[f^2(\mathbf{X}(t))Z(t)]}{E[Z(t)]}\,dE[N(t)] \sim \|f\|_{L_2}^2 E[N(\tau)] \sim \|f\|_{L_2}^2,$$

uniformly in $f \in \mathbb{H}_0$. □

LEMMA 2. *Under Condition* 1, $\|f\|_0 \sim \|f\|_{L_2}$ *uniformly for* $f \in \mathbb{H}$.

PROOF. Observe that

$$\begin{aligned}\|f\|_0^2 &= \int_0^\tau [E^Z(f - E^Z f)^2](t)\,dE[N(t)]\\ &= \int_0^\tau \inf_{c=c(t)} \frac{E[(f(\mathbf{X}(t)) - c)^2 Z(t)]}{E[Z(t)]}\,dE[N(t)]\\ &\sim \int_0^\tau \inf_c E[(f(\mathbf{X}(t)) - c)^2 Z(t)]\,dE[N(t)]\end{aligned}$$

uniformly in $f \in \mathbb{H}$; here, we use the fact that $E[Z(t)]$ is bounded away from zero and infinity uniformly in $t$. Arguing as in the proof of Lemma 1, we obtain that

$$\inf_c E[(f(\mathbf{X}(t)) - c)^2 Z(t)] \sim \inf_c \int_{\mathscr{X}} (f(\mathbf{x}) - c)^2\,d\mathbf{x}.$$

On the other hand,

$$\langle f, 1\rangle = \int_0^\tau E^Z(f)\,dEN = \int_{\mathscr{X}} f(\mathbf{x})p(\mathbf{x})\,d\mathbf{x},$$

where

$$p(\mathbf{x}) = \int_0^\tau \frac{f_{\mathbf{x}(t)}(\mathbf{x})E(Z(t)|\mathbf{X}(t) = \mathbf{x})}{E[Z(t)]}\,dE[N(t)].$$

Thus $\int_{\mathscr{X}} f(\mathbf{x})p(\mathbf{x})\,d\mathbf{x} = 0$ for $f \in \mathbb{H}$. note that $p(\mathbf{x})$ is bounded away from zero and infinity. Hence

$$\|f\|_0^2 \sim \inf_c \int_{\mathscr{X}} (f(\mathbf{x}) - c)^2 p(\mathbf{x})\,d\mathbf{x} = \int_{\mathscr{X}} f^2(\mathbf{x})p(\mathbf{x})\,d\mathbf{x} \sim \|f\|_{L_2}^2. \qquad \square$$

The equivalence of the empirical and theoretical norms is implied by the next two results. Lemma 3 reveals that the empirical inner product is uniformly close to the theoretical inner product on the estimation space $\mathbb{G}_0$, while Lemma 4 is the analogue of Lemma 3 for the ancillary inner products. The proofs of these two lemmas and of Lemma 5 are given in the Appendix.

LEMMA 3. *Suppose Condition* 1 *holds and that* $\lim_n A_n^2 \max(N_n, \log n)/n = 0$. *Then*

$$\sup_{f_1, f_2 \in \mathbb{G}_0} \left| \frac{\langle f_1, f_2 \rangle_n - \langle f_1, f_2 \rangle}{\|f_1\| \|f_2\|} \right| = o_p(1).$$

*Consequently*, $\mathbb{G}_0$ *is empirically identifiable except on an event whose probability tends to zero as* $n \to \infty$.

LEMMA 4. *Suppose Condition* 1 *holds and that* $\lim_n A_n^2 \max(N_n, \log n)/n = 0$. *Then*

$$\sup_{f_1, f_2 \in \widetilde{\mathbb{G}}} \left| \frac{\langle f_1, f_2 \rangle_{0n} - \langle f_1, f_2 \rangle_0}{\|f_1\|_0 \|f_2\|_0} \right| = o_p(1).$$

LEMMA 5. *Let* $\{\phi_i\}$ *be an orthonormal basis of* $\mathbb{G}_0$ *relative to* $\langle\cdot,\cdot\rangle$, *and let* $h_n$, $n \ge 1$, *be uniformly bounded functions on* $\mathscr{X}$. *Then*

$$\sum_i \int_0^\tau [E_n^Z(\phi_i h_n) - E^Z(\phi_i h_n)]^2\,d\overline{N} = O_P\left(\frac{N_n}{n}\right).$$

LEMMA 6. *Let* $h_n, n \ge 1$, *be uniformly bounded functions on* $\mathscr{X}$. *Then*

$$\sup_{g \in \mathbb{G}_0} \left| \frac{\langle g, h_n \rangle_n - \langle g, h_n \rangle}{\|g\|} \right| = O_p\left(\left(\frac{N_n}{n}\right)^{1/2}\right).$$

PROOF. Note that $\langle g, h_n \rangle_n - \langle g, h_n \rangle = I_1 + I_2$, where

$$I_1 = \int_0^\tau [E_n^Z(gh_n) - E^Z(gh_n)]\,d\overline{N}$$

and

$$I_2 = \int_0^\tau E^Z(gh_n)(d\overline{N} - dEN).$$

Let $\{\phi_i\}$ be an orthonormal basis of $\mathbb{G}_0$ relative to $\langle\cdot,\cdot\rangle$. Observe that if $g \in \mathbb{G}_0$, then $g = \sum_i a_i \phi_i$ and $\|g\|^2 = \sum_i a_i^2$. Thus, by the Cauchy–Schwarz inequality,

$$J_1 := \sup_{g\in\mathbb{G}_0} \frac{|I_1|}{\|g\|} \le \biggl( \sum_i \int_0^\tau [E_n^Z(\phi_i h_n) - E^Z(\phi_i h_n)]^2 \, d\overline{N} \biggr)^{1/2}$$

and

$$J_2 := \sup_{g\in\mathbb{G}_0} \frac{|I_2|}{\|g\|} \le \biggl( \sum_i \Bigl( \int_0^\tau E^Z(\phi_i h_n)(d\overline{N} - dEN) \Bigr)^2 \biggr)^{1/2}.$$

Applying Lemma 5, we get that $J_1 = O_P((N_n/n)^{1/2})$. On the other hand,

$$\begin{aligned} E(J_2^2) &= \frac{1}{n} \sum_i \operatorname{var}\biggl( \int_0^\tau E^Z(\phi_i h_n)\, dN \biggr) \\ &\le \frac{1}{n} \sum_i \int_0^\tau [E^Z(\phi_i h_n)]^2 \, dEN \\ &\lesssim \frac{1}{n} \sum_i \int_0^\tau E^Z(\phi_i^2)\, dEN = \frac{1}{n} \sum_i \|\phi_i\|^2 = \frac{N_n}{n}. \end{aligned}$$

Thus, by the Markov inequality, $J_2 = O_P((N_n/n)^{1/2})$. This completes the proof. □

### 5.2. *Approximation error.*

LEMMA 7. *The best approximation $\alpha_n^*$ to $\alpha$ in $\widetilde{\mathbb{G}}$ exists and is uniquely defined for $n$ sufficiently large, $\|\alpha_n^* - \alpha^*\|^2 = O(\rho_n^2)$, $\|\alpha_n^* - \alpha^*\|_n^2 = O_P(\rho_n^2)$, and $\limsup_n \|\alpha_n^*\|_\infty \le \|\alpha^*\|_\infty$.*

To prove the above lemma, we need the following result.

LEMMA 8. *Let $U$ be a positive constant. Then there are positive constants $M_1$ and $M_2$ such that*

$$-M_1\|h - \alpha^*\|^2 \le \Lambda(h) - \Lambda(\alpha^*) \le -M_2\|h - \alpha^*\|^2, \qquad h \in \mathbb{H} \text{ with } \|h\|_\infty \le U.$$

PROOF. Given $h \in \mathbb{H}$ with $\|h\|_\infty \le U$ and given $u \in [0,1]$, set $h^{(u)} = (1-u)\alpha^* + uh$. Then

$$\frac{d}{du}\Lambda(h^{(u)})\bigg|_{u=0} = 0$$

and, by integration by parts,

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

Now

$$\frac{d^2}{du^2}\Lambda(h^{(u)}) = -\int_0^\tau \left[ \frac{E[(h-\alpha^*)^2(\mathbf{X}(t))\exp(h^{(u)}(\mathbf{X}(t)))Z(t)]}{E[\exp(h^{(u)}(\mathbf{X}(t)))Z(t)]} - \left( \frac{E[(h-\alpha^*)(\mathbf{X}(t))\exp(h^{(u)}(\mathbf{X}(t)))Z(t)]}{E[\exp(h^{(u)}(\mathbf{X}(t)))Z(t)]} \right)^2 \right] dE[N(t)]$$

$$= -\int_0^\tau \inf_{c=c(t)} \frac{E[(h-\alpha^*-c)^2(\mathbf{X}(t))\exp(h^{(u)}(\mathbf{X}(t)))Z(t)]}{E[\exp(h^{(u)}(\mathbf{X}(t)))Z(t)]}\, dE[N(t)].$$

Note that $\alpha^*$ is bounded and $\|h\|_\infty \le U$. Thus, by Condition 1 and Lemmas 1 and 2,

$$\frac{d^2}{du^2}\Lambda(h^{(u)}) \sim -\int_0^\tau \inf_{c=c(t)} \frac{E[(h-\alpha^*-c)^2(\mathbf{X}(t))Z(t)]}{E[Z(t)]}\, dE[N(t)]$$

$$= -\|h-\alpha^*\|_0^2 \sim -\|h-\alpha^*\|^2.$$

The desired result follows. □

PROOF OF LEMMA 7. By the strict concavity of $\Lambda(\cdot)$ on $\mathbb{H}$, $\alpha_n^*$ is uniquely defined if it exists. In fact, if both $\alpha_n^*$ and $\tilde{\alpha}_n^*$ maximize $\Lambda(\cdot)$ over $\widetilde{\mathbb{G}}$, then $(d^2/du^2)\Lambda((1-u)\alpha_n^* + u\tilde{\alpha}_n^*) = 0$, $0 < u < 1$. This implies that $\|\alpha_n^* - \tilde{\alpha}_n^*\| = 0$ and hence $\alpha_n^* = \tilde{\alpha}_n^*$. (See also the proof of Lemma 8.)

Since $\alpha^*$ is bounded, by a compactness argument, there is a function $g_0^* \in \mathbb{G}_0$ such that $\|g_0^* - \alpha^*\|_\infty = \rho_n$. Set $g^* = g_0^* - \langle g_0^*, 1\rangle/\langle 1, 1\rangle$. Then $g^* \in \widetilde{\mathbb{G}}$, $\|g^* - \alpha^*\|_\infty \le 2\rho_n$ (note that $\langle \alpha^*, 1\rangle = 0$), and $\|g^*-\alpha^*\| \le \|g_0^*-\alpha^*\| \le \|g_0^*-\alpha^*\|_\infty = \rho_n$. Thus, for $n$ sufficiently large, $\|g^*\|_\infty \le \|g^* - \alpha^*\|_\infty + \|\alpha^*\|_\infty \le 1 + \|\alpha^*\|_\infty$. Let $c$ denote a positive constant (to be determined later). Choose $g \in \widetilde{\mathbb{G}}$ with $\|g - \alpha^*\| = c\rho_n$. Then, by Lemma 1 and the triangle inequality, $\|g - g^*\|_\infty \le A_n\|g - g^*\| \lesssim A_n\rho_n$. Thus, for $n$ sufficiently large, $\|g\|_\infty \lesssim A_n\rho_n + \|g^*\|_\infty \le 1 + \|\alpha^*\|_\infty$. Now applying Lemma 8 with $U = 1 + \|\alpha^*\|_\infty$, we get that, for $n$ sufficiently large,

$$\Lambda(g^*) - \Lambda(\alpha^*) \ge -M_1\rho_n^2 \tag{8}$$

and

$$\Lambda(g) - \Lambda(\alpha^*) \le -M_2c^2\rho_n^2 \tag{9}$$

for all $g \in \widetilde{\mathbb{G}}$ with $\|g-\alpha^*\| = c\rho_n$. Let $c$ be chosen such that $c > \sqrt{M_1/M_2}$. Then $\|g^* - \alpha^*\| < c\rho_n$, and it follows from (8) and (9) that, for $n$ sufficiently large, $\Lambda(g) < \Lambda(g^*)$ for all $g \in \widetilde{\mathbb{G}}$ with $\|g - \alpha^*\| = c\rho_n$. Therefore, by the definition of $\alpha_n^*$ and the concavity of $\Lambda(g)$ as a function of $g, \alpha_n^*$ exists and satisfies $\|\alpha_n^* - \alpha^*\|^2 < c\rho_n$ for $n$ sufficiently large. Thus, $\|\alpha_n^* - \alpha^*\|^2 = O(\rho_n^2)$. This result together with the inequality $\|g^* - \alpha^*\|_\infty \le 2\rho_n$, the triangle inequality and Lemma 3 implies that $\|\alpha_n^* - \alpha^*\|_n^2 = O_P(\rho_n^2)$. The last part of the lemma follows from the first part, $\lim_n A_n\rho_n = 0$, and the inequalities $\|\alpha_n^* - \alpha^*\|_\infty \le A_n\|\alpha_n^* - g^*\| + \|\alpha^* - g^*\|_\infty \le A_n(\|\alpha_n^* - \alpha^*\| + \|\alpha^* - g^*\|) + \|\alpha^* - g^*\|_\infty$. □

### 5.3. *Estimation error.*

LEMMA 9. *Except on an event whose probability tends to zero as $n \to \infty$, the maximum partial likelihood estimate $\hat{\alpha}_n$ exists. Moreover, $\|\hat{\alpha}_n - \alpha_n^*\|^2 = O_P(N_n/n)$.*

PROOF. Note that the approximation space $\mathbb{G}$ depends on the random sample. It is easier to work with an "estimate" in the space $\widetilde{\mathbb{G}}$, which does not vary with the sample. Thus, suppose that $\tilde{\alpha}_n = \operatorname{argmax}_{g \in \widetilde{\mathbb{G}}} \ell(g)$ exists. Then

$$\hat{\alpha}_n = \tilde{\alpha}_n - \frac{\langle \tilde{\alpha}_n, 1 \rangle_n}{\langle 1, 1 \rangle_n}$$

is the unique maximum partial likelihood estimate in $\mathbb{G}$. [Suppose another estimate $\hat{\alpha}_n^1$ maximizes $\ell(g)$ over $\mathbb{G}$. Then $(d^2/du^2)\ell((1-u)\hat{\alpha}_n + u\hat{\alpha}_n^1) = 0$, $0 < u < 1$, and hence $\|\hat{\alpha}_n - \hat{\alpha}_n^1\|_{0n} = 0$; see (12) below. Thus, by Lemmas 1, 2 and 4, $\hat{\alpha}_n^1 = \hat{\alpha}_n$ except on an event with probability tending to zero.] Since $\langle \tilde{\alpha}_n, 1 \rangle = 0$, we have that

$$\|\hat{\alpha}_n - \tilde{\alpha}_n\| \le \left| \frac{\langle \tilde{\alpha}_n, 1 \rangle_n - \langle \tilde{\alpha}_n, 1 \rangle}{\langle 1, 1 \rangle_n} \right| \le \frac{\|\tilde{\alpha}_n\|}{\langle 1, 1 \rangle_n} \sup_{g \in \mathbb{G}_0} \left| \frac{\langle g, 1 \rangle_n - \langle g, 1 \rangle}{\|g\|} \right|.$$

Suppose that $\|\tilde{\alpha}_n\| = O_P(1)$. Then it follows from Lemma 6 that

$$\|\hat{\alpha}_n - \tilde{\alpha}_n\| = O_P\left(\left(\frac{N_n}{n}\right)^{1/2}\right).$$

Note that $\hat{\alpha}_n - \alpha_n^* = (\hat{\alpha}_n - \tilde{\alpha}_n) + (\tilde{\alpha}_n - \alpha_n^*)$. In light of Lemma 6, it remains to prove that $\tilde{\alpha}_n$ exists and $\|\tilde{\alpha}_n - \alpha_n^*\| = O_P((N_n/n)^{1/2})$.

Let $\{\phi_j,\ 1 \le j \le N_n\}$ be an orthonormal basis of $\widetilde{\mathbb{G}}$ relative to the theoretical inner product $\langle \cdot, \cdot \rangle$. Then each $g \in \widetilde{\mathbb{G}}$ can be represented uniquely as $g = \sum_j \beta_j \phi_j$, where $\beta_j = \langle g, \phi_j \rangle$ for $j = 1, \ldots, N_n$. Let $\boldsymbol{\beta}$ denote the $N_n$-dimensional vector with entries $\beta_j$. To indicate the dependence of $g$ on $\boldsymbol{\beta}$, we write $g(\cdot) = g(\cdot; \boldsymbol{\beta})$. Let $|\cdot|$ denote the Euclidean norm of vectors. Then $\|g(\cdot; \boldsymbol{\beta})\| = |\boldsymbol{\beta}|$. Let $\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$ be given by $\tilde{\alpha}_n(\cdot) = g(\cdot; \tilde{\boldsymbol{\beta}})$ and $\alpha_n^*(\cdot) = g(\cdot; \boldsymbol{\beta}^*)$. Then $\|\tilde{\alpha}_n - \alpha_n^*\| = |\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|$.

The corresponding partial log-likelihood function is given by

$$\begin{aligned} \ell(\boldsymbol{\beta}) = {} & \frac{1}{n} \sum_i \int_0^\tau g(\mathbf{X}_i(t); \boldsymbol{\beta})\, dN_i(t) \\ & - \int_0^\tau \log\left( \frac{1}{n} \sum_i Z_i(t) \exp g(\mathbf{X}_i(t); \boldsymbol{\beta}) \right) d\overline{N}(t). \end{aligned}$$

Let

$$\mathbf{S}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \ell(\boldsymbol{\beta})$$

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

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \frac{1}{n}\sum_i \int_0^\tau \phi_j(\mathbf{X}_i(t))\, dN_i(t) \\ - \int_0^\tau \frac{(1/n)\sum_i \phi_j(\mathbf{X}_i(t)) Z_i(t) \exp g(\mathbf{X}_i(t);\boldsymbol{\beta})}{(1/n)\sum_i Z_i(t)\exp g(\mathbf{X}_i(t);\boldsymbol{\beta})}\, d\overline{N}(t).$$

Let

$$\mathbf{D}(\boldsymbol{\beta}) = \frac{\partial^2 \ell(g)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}$$

denote the Hessian of $\ell(g)$. Then the following identity holds:

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \ell(\boldsymbol{\beta}^*) + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \mathbf{S}(\boldsymbol{\beta}^*) \\ &\quad + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T\left(\int_0^1 (1-u)\mathbf{D}(\boldsymbol{\beta}^* + u(\boldsymbol{\beta} - \boldsymbol{\beta}^*))\, du\right)(\boldsymbol{\beta} - \boldsymbol{\beta}^*).\end{aligned} \tag{10}$$

To complete the proof, we need the following results, the proofs of which will be given shortly.

LEMMA 10. *For any positive constant $M$,*

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

LEMMA 11. *There is a positive constant $M$ such that, for any fixed positive constant $a$,*

$$(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T\left(\int_0^1 (1-u)\mathbf{D}(\boldsymbol{\beta}^* + u(\boldsymbol{\beta} - \boldsymbol{\beta}^*))\, du\right)(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \\ \le -M|\boldsymbol{\beta} - \boldsymbol{\beta}^*|^2 \quad \text{for all } \boldsymbol{\beta} \text{ with } |\boldsymbol{\beta} - \boldsymbol{\beta}^*| = a\left(\frac{N_n}{n}\right)^{1/2}$$

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

Since $\ell(\boldsymbol{\beta})$ is a concave function of $\beta$, we conclude from (10) and Lemma 11 that

$$\left\{\tilde{\boldsymbol{\beta}} \text{ exists and } |\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*| < a\left(\frac{N_n}{n}\right)^{1/2}\right\} \cap \Omega_n(a) \\ \supset \left\{\ell(\boldsymbol{\beta}) - \ell(\boldsymbol{\beta}^*) < 0 \text{ for all } \boldsymbol{\beta} \text{ with } |\boldsymbol{\beta} - \boldsymbol{\beta}^*| = a\left(\frac{N_n}{n}\right)^{1/2}\right\} \cap \Omega_n(a) \\ \supset \left\{|\mathbf{S}(\boldsymbol{\beta}^*)| < Ma\left(\frac{N_n}{n}\right)^{1/2}\right\} \cap \Omega_n(a).$$

Hence, by Lemma 10,

$$\begin{aligned}\lim_{a\to\infty}\liminf_{n\to\infty} P\biggl(\tilde{\boldsymbol\beta} \text{ exists and } |\tilde{\boldsymbol\beta}-\boldsymbol\beta^*| < a\biggl(\frac{N_n}{n}\biggr)^{1/2}\biggr)\\ \geq \lim_{a\to\infty}\liminf_{n\to\infty}\biggl\{P\biggl(|\mathbf{S}(\boldsymbol\beta^*)| < Ma\biggl(\frac{N_n}{n}\biggr)^{1/2}\biggr) + P(\Omega_n(a)) - 1\biggr\} = 1.\end{aligned}$$

Consequently, $\tilde{\alpha}_n = g(\cdot, \tilde{\boldsymbol\beta})$ exists with probability tending to one; moreover,

$$\|\tilde{\alpha}_n - \alpha_n^*\|^2 = O_P(N_n/n).$$

This completes the proof of Lemma 9. □

Proof of Lemma 10. Note that

$$\begin{aligned}S_j(\boldsymbol\beta^*) = \frac{1}{n}\sum_i \int_0^\tau \phi_j(\mathbf{X}_i(t))\,dN_i(t)\\ - \int_0^\tau \frac{\sum_i \phi_j(\mathbf{X}_i(t))Z_i(t)\exp\alpha_n^*(\mathbf{X}_i(t))}{\sum_i Z_i(t)\exp\alpha_n^*(\mathbf{X}_i(t))}\,d\overline{N}(t).\end{aligned}$$

By the definition of $\boldsymbol\beta^*$, $(\partial/\partial\beta_j)\Lambda(\boldsymbol\beta^*) = 0$ for each $j$; that is,

$$E\biggl(\int_0^\tau \phi_j(\mathbf{X}(t))\,dN(t)\biggr) - \int_0^\tau \frac{E[\phi_j(\mathbf{X}(t))Z(t)\exp\alpha_n^*(\mathbf{X}(t))]}{E[Z(t)\exp\alpha_n^*(\mathbf{X}(t))]}\,dE[N(t)] = 0.$$

Thus, $S_j(\boldsymbol\beta^*) = I_{1j} - I_{2j} - I_{3j}$, where

$$\begin{aligned}I_{1j} &= \frac{1}{n}\sum_i \int_0^\tau \phi_j(\mathbf{X}_i(t))\,dN_i(t) - E\biggl(\int_0^\tau \phi_j(\mathbf{X}(t))\,dN(t)\biggr),\\ I_{2j} &= \int_0^\tau \biggl(\frac{\sum_i \phi_j(\mathbf{X}_i(t))Z_i(t)\exp\alpha_n^*(\mathbf{X}_i(t))}{\sum_i Z_i(t)\exp\alpha_n^*(\mathbf{X}_i(t))}\\ &\qquad - \frac{E[\phi_j(\mathbf{X}(t))Z(t)\exp\alpha_n^*(\mathbf{X}(t))]}{E[Z(t)\exp\alpha_n^*(\mathbf{X}(t))]}\biggr)\,d\overline{N}(t)\end{aligned}$$

and

$$I_{3j} = \int_0^\tau \frac{E[\phi_j(\mathbf{X}(t))Z(t)\exp\alpha_n^*(\mathbf{X}(t))]}{E[Z(t)\exp\alpha_n^*(\mathbf{X}(t))]}\,(d\bar{N}(t) - dE[N(t)]).$$

Hence, $|\mathbf{S}(\beta^*)|^2 \lesssim \sum_j (I_{1j}^2 + I_{2j}^2 + I_{3j}^2)$.

Now let us deal with $\sum_j I_{1j}^2$. We have that

$$E(I_{1j}^2) \le \frac{1}{n} E\biggl(\int_0^\tau \phi_j^2(\mathbf{X}(t))\,dN(t)\biggr) \sim \frac{1}{n}\int_0^\tau E[\phi_j^2(\mathbf{X}(t))]\,dt \sim \frac{1}{n}\|\phi_j\|_{L_2}^2.$$

Thus

$$E\biggl(\sum_j I_{1j}^2\biggr) \lesssim \frac{1}{n}\sum_j \|\phi_j\|^2 = \frac{N_n}{n},$$

and hence $\sum_j I_{1j}^2 = O_P(N_n/n)$. To deal with $\sum_j I_{2j}^2$, by the Cauchy–Schwarz inequality,

$$\sum_j I_{2j}^2 \lesssim \sum_j \int_0^\tau \biggl(\frac{\sum_i \phi_j(\mathbf{X}_i(t))Z_i(t)\exp\alpha_n^*(\mathbf{X}_i(t))}{\sum_i Z_i(t)\exp\alpha_n^*(\mathbf{X}_i(t))} - \frac{E[\phi_j(\mathbf{X}(t))Z(t)\exp\alpha_n^*(\mathbf{X}(t))]}{E[Z(t)\exp\alpha_n^*(\mathbf{X}(t))]}\biggr)^2 d\bar{N}(t).$$

Note that $\alpha_n^*$ is bounded uniformly in $n$. Using the same argument in the proof of Lemma 5, we obtain that $\sum_j I_{2j}^2 = O_P(N_n/n)$. The argument used in handling $J_2^2$ in the proof of Lemma 6 leads to $\sum_j I_{3j}^2 = O_P(N_n/n)$. □

PROOF OF LEMMA 11. Choose $M_1 > \|\alpha^*\|_\infty$ and let $a > 0$. It then follows from Lemma 7 that, for $n$ sufficiently large, $\|g\|_\infty \le \|\alpha_n^*\|_\infty + A_n\|g - \alpha_n^*\| \le M_1$ for all $g \in \widetilde{\mathbb{G}}$ with $\|g - \alpha_n^*\| = a(N_n/n)^{1/2}$. We now prove that there is a positive constant $M_2$ such that, except on an event whose probability tends to zero as $n \to \infty$,

$$\frac{d^2}{du^2}\ell(g_1 + u(g_2 - g_1) \le -M_2\|g_1 - g_2\|^2) \tag{11}$$

for $0 < u < 1$ and all $g_1, g_2 \in \widetilde{\mathbb{G}}$ with $\|g_1\|_\infty \le M_1$ and $\|g_2\|_\infty \le M_1$. Indeed,

$$\begin{aligned}
&\frac{d^2}{du^2}\ell(g_1+u(g_2-g_1))\\
&\quad= -\int_0^\tau \biggl[\frac{(1/n)\sum_i Z_i(t)[(g_2-g_1)^2\exp(g_1+u(g_2-g_1))](\mathbf{X}_i(t))}{(1/n)\sum_i Z_i(t)\exp(g_1+u(g_2-g_1))(\mathbf{X}_i(t))}\\
&\qquad\qquad - \biggl(\frac{(1/n)\sum_i Z_i(t)[(g_2-g_1)\exp(g_1+|,u(g_2-g_1))](\mathbf{X}_i(t))}{(1/n)\sum_i Z_i(t)\exp(g_1+u(g_2-g_1)(\mathbf{X}_i(t))}\biggr)^2\biggr]\,d\bar{N}(t) \qquad (12)\\
&\quad= -\int_0^\tau \inf_c \frac{(1/n)\sum_i Z_i(t)[(g_2-g_1-c))^2\exp(g_1+u(g_2-g_1))](\mathbf{X}_i(t))}{(1/n)\sum_i Z_i(t)\exp(g_1+u(g_2-g_1))(\mathbf{X}_i(t))}\,d\bar{N}(t)
\end{aligned}$$

Since $\|g_1\|_\infty \le M_1$ and $\|g_2\|_\infty \le M_1$,

$$\begin{aligned}&\frac{d^2}{du^2}\ell(g_1 + u(g_2 - g_1))\\ &\quad\sim -\int_0^\tau \inf_c \frac{(1/n)\sum_i Z_i(t)[g_2(\mathbf{X}_i(t)) - g_1(\mathbf{X}_i(t)) - c]^2}{(1/n)\sum_i Z_i(t)}\, d\bar{N}(t)\\ &\quad= -\|g_2 - g_1\|_{0n}^2 \sim -\|g_2 - g_1\|_0^2 \sim -\|g_2 - g_1\|^2,\end{aligned}$$

except on an event whose probability tends to zero as $n \to \infty$; here we use Lemmas 1, 2 and 4 to obtain the equivalence of the norms. This completes the proof of (11).

Choose $g \in \widetilde{\mathbb{G}}$ such that $\|g - \alpha_n^*\|^2 = a(N_n/n)^{1/2}$. Then by the definition of $A_n$ and Lemma 1,

$$\|g - \alpha_n^*\|_\infty \le A_n\|g - \alpha_n^*\|_{L_2} \sim A_n\|g - \alpha_n^*\| = A_n a(N_n/n)^{1/2} = o(1).$$

Moreover, $\alpha_n^*$ is bounded uniformly in $n$ by Lemma 7. Hence it follows from (11) that, except on an event whose probability tends to zero as $n \to \infty$,

$$\frac{d^2}{du^2}\ell(\alpha_n^* + u(g - \alpha_n^*)) \le -M_2\|g - \alpha_n^*\|^2$$

for $0 < u < 1$ and all $g \in \widetilde{\mathbb{G}}$ with $\|g - \alpha_n^*\| = a(N_n/n)^{1/2}$. Equivalently,

$$(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T\mathbf{D}(\boldsymbol{\beta}^* + u(\boldsymbol{\beta} - \boldsymbol{\beta}^*))(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \le -M_2|\boldsymbol{\beta} - \boldsymbol{\beta}^*|^2$$

for $0 < u < 1$ and all $\boldsymbol{\beta} \in \mathbb{R}^{N_n}$ with $|\boldsymbol{\beta} - \boldsymbol{\beta}^*| = a(N_n/n)^{1/2}$ or an event $\Omega_n(a)$ with $\lim_n P(\Omega_n(a)) = 1$. The desired result follows with $M = M_2\int_0^1(1-u)\,du = M_2/2$. □

## 6. Convergence in functional ANOVA model.

The proof of Theorem 3 is given in this section. Note that $N_n \sim \sum_{s\in\mathscr{S}} N_s$ and $\rho_n \le \bar{\rho}_n$. It follows from the Cauchy–Schwarz inequality that $A_n \lesssim (\sum_{s\in\mathscr{S}} A_s^2)^{1/2}$; see the proof of Theorem 2 of Huang (1998a). The first part of the theorem then follows from Theorem 2.

We have the ANOVA decompositions $\hat{\alpha}_n = \sum_{s\in\mathscr{S}} \hat{\alpha}_s$ and $\alpha^* = \sum_{s\in\mathscr{S}} \alpha_s^*$, where $\hat{\alpha}_s \in \mathbb{G}_s^0$ and $\alpha_s \in \mathbb{H}_s^0$ for $s \in \mathscr{S}$. To prove the second part of Theorem 3, we need the following result from Huang (1998b). (We have customized the notation.)

THEOREM 4. *Suppose that* (i) $\|h\| \sim \|h\|_{L_2}$ *uniformly in* $h \in \mathbb{H}_0$; (ii) $\sup_{g\in\mathbb{G}_0} |\|g\|_n/\|g\| - 1| = o_P(1)$; *and* (iii) *for every sequence* $f_n$, $n \ge 1$, *of uniformly bounded functions on* $\mathscr{X}$,

$$\sup_{g\in\mathbb{G}_s}\left|\frac{\langle f_n, g\rangle_n - \langle f_n, g\rangle}{\|g\|}\right| = O_P\left(\left(\frac{N_s}{n}\right)^{1/2}\right), \qquad s \in \mathscr{S}.$$

*Then* $\|\hat{\alpha}_s - \alpha_s^*\|^2 = O_P(\|\hat{\alpha} - \alpha^*\|^2 + \bar{\rho}_n^2 + N_n/n)$ *and* $\|\hat{\alpha}_s - \alpha_s^*\|_n^2 = O_P(\|\hat{\alpha} - \alpha^*\|^2 + \bar{\rho}_n^2 + N_n/n)$ *for* $s \in \mathscr{S}$.

Conditions (i) and (ii) hold because of Lemmas 1 and 3, respectively. Condition (iii) follows from Lemma 6 with $\mathbb{G}_0$ replaced by $\mathbb{G}_s$ for $s \in \mathscr{S}$. Consequently, Theorem 3 follows from Theorems 2 and 4.

## APPENDIX

**A.1. Existence of the best approximation.** This section contains the proof of Theorem 1. The proof is rather involved due to the identifiability constraint. We first give some ancillary results.

LEMMA 12. *Let $W$ be a random variable with $P(c_1 \le W \le c_2) = 1$, where $0 < c_1 \le c_2 < \infty$, and let $WZ$ have mean zero. Then*

$$\log E(We^Z) \ge 2\log\left(1 + \frac{1}{2}\left(\frac{c_1}{2c_2}\right)^{1/2} E|Z|\right) + \log\frac{c_1}{2}.$$

PROOF. Suppose first that $E(We^{|Z|}) < \infty$ and define the function $f$ on [0,1] by $f(s) = E(We^{sZ})$. Then $f(0) = EW$, $f'(0) = 0$ and

$$f''(s) = E(WZ^2e^{sZ}) \ge E(WZ_+^2) \ge \frac{[E(WZ_+)]^2}{EW} = \frac{[E(W|Z|)^2]}{4EW}, \qquad 0 < s < 1.$$

Consequently,

$$\begin{aligned}
E(We^Z) = f(1) &\ge EW + \frac{[E(W|Z|)]^2}{8EW} \\
&\ge c_1\left(1 + \frac{c_1(E|Z|)^2}{8c_2}\right) \\
&\ge \frac{c_1}{2}\left(1 + \frac{1}{2}\left(\frac{c_1}{2c_2}\right)^{1/2} E|Z|\right)^2,
\end{aligned}$$

which yields the desired result. The general result now follows from the monotone convergence theorem and an elementary truncation argument. □

COROLLARY 1. *Let $W$ be a random variable with $P(c_1 \le W \le c_2) = 1$, where $0 < c_1 \le c_2 < \infty$, and let $Z$ have finite mean. Then*

$$\log E(We^Z) - \frac{E(WZ)}{EW} \ge 2\log\left(1 + \frac{1}{4}\left(\frac{c_1}{2c_2}\right)^{1/2} E|Z - EZ|\right) + \log\frac{c_1}{2}.$$

PROOF. Let $c \in \mathbb{R}$. Then $|EZ - c| \le E|Z - c|$, so $E|Z - EZ| \le 2E|Z - c|$. The desired result now follows from Lemma 12 and the previous inequality with $c = E(WZ)/EW$. □

COROLLARY 2. *Let $W_1$ and $W_2$ be random variables with $P(c_1 \le W_1 \le c_2) = 1$ and $P(W_2 \ge c_1) = 1$, where $0 < c_1 \le c_2 < \infty$, and let $Z$ have finite mean. Then*

$$\log E(W_1 W_2 e^Z) - \frac{E(W_1 Z)}{EW_1} \ge 2 \log\left(1 + \frac{1}{4}\left(\frac{c_1}{2c_2}\right)^{1/2} E|Z - EZ|\right) + \log\frac{c_1^2}{2}.$$

Let $h$ be an integrable function on $\mathscr{X}$. Then

$$\begin{aligned} E\int_0^\tau h(\mathbf{X}(t))\,dN(t) &= E\left[E\left(\int_0^\tau h(\mathbf{X}(t))\,dN(t)\,\Big|\,\mathbf{X}\right)\right] \\ &= E\int_0^\tau h(\mathbf{X}(t))\,dE(N(t)|\mathbf{X}) \end{aligned}$$

and

$$E[Z(t)e^{h(\mathbf{X}(t))}] = E\Big(E[Z(t)e^{h(\mathbf{X}(t))}|\mathbf{X}]\Big) = E\Big(e^{h(\mathbf{X}(t))}E(Z(t)\mid\mathbf{X})\Big).$$

Note that $E(Z(t)|\mathbf{X}) = \rho(t\mid\mathbf{X})$ and $dE(N(t)\mid\mathbf{X}) = \gamma(t\mid\mathbf{X})\,dt$ where

$$\rho(t\mid\mathbf{X}) = \exp\left(-\int_0^t \lambda(u\mid\mathbf{X})\,du\right)P(C \ge t\mid\mathbf{X})$$

and $\gamma(t\mid\mathbf{X}) = \rho(t\mid\mathbf{X})\lambda(t\mid\mathbf{X})$. It follows from Condition 1 that $\rho(t\mid\mathbf{X})$ and $\gamma(t\mid\mathbf{X})$ are bounded away from zero and infinity uniformly over $t \in [0,\tau]$ and $\mathbf{X}$.

Set $W_1(t) = \gamma(t\mid\mathbf{X})$ and $W_2(t) = 1/\lambda(t\mid\mathbf{X})$, and let $c_1, c_2$ be such that $0 < c_1 \le c_2 < \infty$ and $P(c_1 \le W_1(t) \le c_2) = 1$ and $P(W_2(t) \ge c_1) = 1$ for $0 \le t \le \tau$.

Now we are ready to prove the first conclusion of Theorem 1. Choose $h \in \mathbb{H}$ and observe that

$$\begin{aligned} \Lambda(h) = \int_0^\tau \Big\{&E[h(\mathbf{X}(t))W_1(t)] \\ &- \log\Big(E\big[e^{h(\mathbf{X}(t))}W_1(t)W_2(t)\big]\Big)E[W_1(t)]\Big\}\,dt. \end{aligned} \tag{13}$$

It follows from Condition 1 that

$$\int_0^\tau \frac{E[h(\mathbf{X}(t)Z(t)]}{E[Z(t)]}\,dE[N(t)] = \int_{\mathscr{X}} h(\mathbf{x})\psi(\mathbf{x})\,d\mathbf{x} = 0, \qquad h \in \mathbb{H},$$

where the positive function $\psi$ is bounded away from zero and infinity on $\mathscr{X}$. Thus, by Corollary 2 and Condition 1 (see the proof of Corollary 1), there are positive constants $C_1$, $C_2$ and $C_3$ such that

$$\Lambda(h) \le -C_1 \log\Big(1 + C_2\int_{\mathscr{X}} |h(\mathbf{x})|\,d\mathbf{x}\Big) + C_3, \qquad h \in \mathbb{H}, \tag{14}$$

and hence $\Lambda(h) \le C_3$ for $h \in \mathbb{H}$. Consequently, the numbers $\Lambda(h)$, $h \in \mathbb{H}$, have a finite least upper bound $L$. Choose $h_n \in \mathbb{H}$ such that $\Lambda(h_n) > -\infty$ and

$\Lambda(h_n) \to L$ as $n \to \infty$. Observe that the numbers $\int_{\mathscr{X}} |h_n(\mathbf{x})|\, d\mathbf{x}$, $n \geq 1$, are bounded. It follows from (13) and Condition 1 that

$$\lim_{M\to\infty} \limsup_{n\to\infty} \int_{\{\mathbf{x}\in\mathscr{X}:\ h_n(\mathbf{x})\geq M\}} h_n(\mathbf{x})\, d\mathbf{x} = 0. \tag{15}$$

We will prove that an appropriately translated version of $h_n$ converges and that the limit is the desired best approximation.

Let $|A|$ denote the Lebesgue measure of a subset $A$ of $\mathscr{X}$. Let $c_0 > 0$ be sufficiently large that $|\mathscr{X}\setminus A_{mn}(c_0)| \leq |\mathscr{X}|/2$ [and hence $A_{mn}(c_0) \geq |\mathscr{X}|/2$] for $m, n \gg 1$, where

$$A_{mn}(c) = \{\mathbf{x} \in \mathscr{X}\colon |h_m(\mathbf{x})| \leq c \text{ and } |h_n(\mathbf{x})| \leq c\}, \qquad c > 0.$$

Set $\Psi_{mn}(u) = \Lambda((1-u)h_n + uh_m)$ for $0 \leq u \leq 1$. Then $\Psi_{mn}$ is bounded above by $L$, and it follows from Hölder's inequality that this function is concave. Set $h_{mn} = h_m - h_n$, let $c \geq c_0$, and set

$$\bar{h}_{mn}(c) = \frac{1}{|A_{mn}(c)|} \int_{A_{mn}(c)} h_{mn}(\mathbf{x})\, d\mathbf{x}.$$

We claim that there is a positive constant $M$ such that (uniformly in $u$)

$$\Psi''_{mn}(u) \leq -\frac{1}{M} \int_{A_{mn}(c)} [h_{mn}(\mathbf{x}) - \bar{h}_{mn}(c)]^2\, d\mathbf{x}, \qquad m, n \gg 1. \tag{16}$$

Observe that $\Psi_{mn}(\frac{1}{2}) \leq L$. Choose $\delta \geq 0$. Then $\Psi_{mn}(0) \geq L - \delta$ and $\Psi_{mn}(1) \geq L - \delta$ for $m$, $n \gg 1$. Also,

$$\Psi_{mn}\left(\tfrac{1}{2}\right) - \Psi_{mn}(0) \geq \tfrac{1}{2}\Psi'_{mn}\left(\tfrac{1}{2}\right) \tag{17}$$

since $\Psi'_{mn}(u)$ is a nonincreasing function of $u$. Moreover, we conclude from (16) that

$$\begin{aligned} \Psi_{mn}(1) \leq{}& \Psi_{mn}\left(\frac{1}{2}\right) + \frac{1}{2}\Psi'_{mn}\left(\frac{1}{2}\right) \\ &- \frac{1}{8M} \int_{A_{mn}(c)} [h_{mn}(\mathbf{x}) - \bar{h}_{mn}(c)]^2\, d\mathbf{x},\ m, n \gg 1. \end{aligned} \tag{18}$$

It follows from (17) and (18) that

$$\begin{aligned} &\frac{\Psi_{mn}(0) + \Psi_{mn}(1)}{2} \\ &\quad \leq \Psi_{mn}\left(\frac{1}{2}\right) - \frac{1}{16M} \int_{A_{mn}(c)} [h_{mn}(\mathbf{x}) - \bar{h}_{mn}(c)]^2\, d\mathbf{x}, \quad m, n \gg 1, \end{aligned}$$

and hence that

$$\int_{A_{mn}(c)} [h_{mn}(\mathbf{x}) - \bar{h}_{mn}(c)]^2\, d\mathbf{x} \leq 16\delta M, \qquad m, n \gg 1.$$

Since $\delta$ can be made arbitrarily small, we conclude that

$$\lim_{m,\, n\to\infty} \int_{A_{mn}(c)} [h_{mn}(\mathbf{x}) - \bar{h}_{mn}(c)]^2\, d\mathbf{x} = 0.$$

Consequently, for $\eta > 0$,

$$\lim_{m,n\to\infty} |\{\mathbf{x} \in A_{mn}(c)\colon |h_{mn}(\mathbf{x}) - \bar{h}_{mn}(c)| \geq \eta\}| = 0.$$

In particular,

$$\lim_{m,n\to\infty} |\{\mathbf{x} \in A_{mn}(c_0)\colon\; |h_{mn}(\mathbf{x}) - \bar{h}_{mn}(c_0)| \geq \eta\}| = 0.$$

Also,

$$\lim_{m,n\to\infty} |\{\mathbf{x} \in A_{mn}(c_0)\colon\; |h_{mn}(\mathbf{x}) - \bar{h}_{mn}(c)| \geq \eta\}| = 0$$

since $A_{mn}(c_0) \subset A_{mn}(c)$. Thus, for $m, n \gg 1$, there is an $\mathbf{x} \in A_{mn}(c_0)$ such that $|h_{mn}(\mathbf{x}) - \bar{h}_{mn}(c_0)| < \eta$ and $|h_{mn}(\mathbf{x}) - \bar{h}_{mn}(c)| < \eta$ and hence $|\bar{h}_{mn}(c) - \bar{h}_{mn}(c_0)| < 2\eta$. Since $\eta$ can be made arbitrarily small, $\lim_{m,n\to\infty} [\bar{h}_{mn}(c) - \bar{h}_{mn}(c_0)] = 0$, from which we conclude that

$$\lim_{m,n\to\infty} |\{\mathbf{x} \in A_{mn}(c)\colon |h_{mn}(\mathbf{x}) - \bar{h}_{mn}(c_0)| \geq \eta\}| = 0, \qquad \eta > 0.$$

Since $c$ can be made arbitrarily large and hence $\limsup_{m,n\to\infty} |\mathscr{X}\setminus A_{mn}(c)|$ can be made arbitrarily small, it follows that

$$\lim_{m,n\to\infty} |\{\mathbf{x} \in \mathscr{X}\colon |h_{mn}(\mathbf{x}) - h_n(\mathbf{x}) - \bar{h}_{mn}(c_0)| \geq \eta\}| = 0, \qquad \eta > 0.$$

Let $\mathbf{U}$ be uniformly distributed on $\mathscr{X}$ and set $U_n = h_n(\mathbf{U})$. Then $U_m - U_n - \bar{h}_{mn}(c_0)$ converges in probability to zero as $m, n \to \infty$. Moreover, $U_n$, $n \geq 1$, are bounded in probability. Thus, by a compactness argument, there is an increasing sequence $(n_j)$ of positive integers such that $U_{n_j}$ has a limiting distribution. Consequently, $U_{n_j} - U_{n_l}$ converges in probability to zero as $j, l \to \infty$. Therefore, there is an integrable function $\alpha^*$ such that $h_{n_j} \to \alpha^*$ in measure as $j \to \infty$. It follows from Condition 2 that $\alpha^* \in \mathbb{H}_0$. By adding a constant to $\alpha^*$ if necessary, we can assume that $\alpha^* \in \mathbb{H}$. It follows from (13), (15) and Fatou's lemma that $\Lambda(\alpha^*) \geq L$ and hence that $\Lambda(\alpha^*) = \max_{h\in\mathbb{H}} \Lambda(h)$. Similarly, if $h \in \mathbb{H}$ and $\Lambda(h) = \Lambda(\alpha^*)$, then $h = \alpha^*$ almost everywhere. Therefore, the first statement of the theorem is valid.

In order to verify (16), set

$$W_{mnt}(u) = \exp(h_n(\mathbf{X}(t)) + u h_{mn}(\mathbf{X}(t))) W_1(t)\, W_2(t).$$

Then $\Psi''_{mn}(u) = \int_0^\tau \Psi''_{mnt}(u) E[W_1(t)]\, dt$ where

$$\Psi''_{mnt}(u) = -\left[\frac{E[h^2_{mn}(\mathbf{X}(t)) W_{mnt}(u)]}{E[W_{mnt}(u)]} - \left(\frac{E[h_{mn}(\mathbf{X}(t)) W_{mnt}(u)]}{E[W_{mnt}(u)]}\right)^2\right].$$

For fixed $m, n$ and $t$, let $Y$ have the distribution given by

$$E[\phi(Y)] = \frac{E[\phi(h_{mn}(\mathbf{X}(t))) W_{mnt}(u)]}{E[W_{mnt}(u)]}.$$

Then $\Psi''_{mnt}(u) = -\mathrm{var}(Y)$. Let $V$ have the distribution given by

$$E[\phi(V)] = \frac{E[\phi(h_{mn}(\mathbf{X}(t)))W_{mnt}(u)\mathrm{ind}(\mathbf{X}(t) \in A_{mn}(c))]}{E[W_{mnt}(u)\mathrm{ind}(\mathbf{X}(t) \in A_{mn}(c))]},$$

and set $\mu = EY$ and $\nu = EV$. Then

$$\begin{aligned}\mathrm{var}(V) \le E[(V-\mu)^2] &\le \frac{\mathrm{var}(Y)E[W_{mnt}(u)]}{E[W_{mnt}(u)\mathrm{ind}(\mathbf{X}(t) \in A_{mn}(c))]} \\ &= -\frac{\Psi''_{mnt}(u)E[W_{mnt}(u)]}{E[W_{mnt}(u)\mathrm{ind}(\mathbf{X}(t) \in A_{mn}(c))]}.\end{aligned}$$

There is a positive constant $M_1$ such that

$$\frac{E[W_{mnt}(u)\mathrm{ind}(\mathbf{X}(t) \in A_{mn}(c))]}{E[W_{mnt}(u)]} \ge \frac{1}{M_1}, \qquad m, n \gg 1,$$

and hence such that

$$\Psi''_{mnt}(u) \le -\frac{1}{M_1}\mathrm{var}(V), \qquad m, n \gg 1. \tag{19}$$

There is a positive constant $M_2$ such that

$$\mathrm{var}(V) \ge \frac{1}{M_2}\int_{A_{mn}(c)} [h_{mn}(\mathbf{x}) - \bar{h}_{mn}(c)]^2\, d\mathbf{x}, \qquad m, n \gg 1. \tag{20}$$

Inequality (16) follows from (19) and (20).

To prove the second statement of the theorem, suppose that $\lambda(t\,|\,\mathbf{X}) = \lambda_0(t) e^{\alpha(\mathbf{X}(t))}$, where $\lambda_0 = \exp\alpha_0$. Then $dE(N(t)\,|\,\mathbf{X}) = \rho(t\,|\,\mathbf{X})\lambda_0(t)e^{\alpha(\mathbf{X}(t))}\,dt$, so

$$\begin{aligned}\Lambda(h) = E\int_o^\tau h(\mathbf{X}(t))e^{\alpha(\mathbf{X}(t))}\rho(t\,|\,\mathbf{X})\lambda_0(t)\,dt \\ - \int_0^\tau \log\Big(E[e^{h(\mathbf{X}(t))}\rho(t\,|\,\mathbf{X})]\Big)E[e^{\alpha(\mathbf{X}(t))}\rho(t\,|\,\mathbf{X})]\lambda_0(t)\,dt.\end{aligned}$$

Set $g = h - \alpha$ and

$$\psi(t\,|\,\mathbf{X}) = \frac{e^{\alpha(\mathbf{X}(t))}\rho(t\,|\,\mathbf{X})}{E[e^{\alpha(\mathbf{X}(t))}\rho(t\,|\,\mathbf{X})]}.$$

Then we can write

$$\Lambda(h) - \Lambda(\alpha) = \int_0^\tau \pi_t(g)E[e^{\alpha(\mathbf{X}(t))}\rho(t\,|\,\mathbf{X})]\lambda_0(t)\,dt,$$

where

$$\pi_t(g) = E[g(\mathbf{X}(t))\psi(t\,|\,\mathbf{X})] - \log\Big(E[e^{g(\mathbf{X}(t))}\psi(t\,|\,\mathbf{X})]\Big).$$

It follows from Jensen's inequality that $\pi_t(g) \le 0$ and that $\pi_t(g) < 0$ unless $g$ is essentially constant on $\mathscr{X}$. This yields the desired result.

### A.2. Proofs of Lemmas 3, 4 and 5.

PROOF OF LEMMA 3. For $f_1, f_2 \in \mathbb{G}_0$, write

$$\langle f_1, f_2\rangle_n - \langle f_1, f_2\rangle = I_1 + I_2 + I_3,$$

where

$$I_1 = \int_0^\tau E_n[f_1(\mathbf{X}(t))f_2(\mathbf{X}(t))Z(t)]\biggl(\frac{1}{E_n[Z(t)]} - \frac{1}{E[Z(t)]}\biggr)\,d\overline{N}(t),$$
$$I_2 = \int_0^\tau \frac{E_n[f_1(\mathbf{X}(t))f_2(\mathbf{X}(t))Z(t)] - E[f_1(\mathbf{X}(t))f_2(\mathbf{X}(t))Z(t)]}{E[Z(t)]}\,d\overline{N}(t)$$

and

$$I_3 = \int_0^\tau \frac{E_n[f_1(\mathbf{X}(t))f_2(\mathbf{X}(t))Z(t)]}{E[Z(t)]}\bigl(d\overline{N}(t) - dE[N(t)]\bigr).$$

By Lemma 10 of Huang (1998a),

$$\sup_{f_1, f_2\in\mathbb{G}_0} \frac{|I_3|}{\|f_1\|\|f_2\|} = o_P(1). \tag{21}$$

We can write $I_2 = I_{21} + I_{22}$, where

$$I_{21} = \frac{1}{n^2}\sum_j \int_0^\tau \frac{f_1(\mathbf{X}_j(t))f_2(\mathbf{X}_j(t))Z_j(t) - E[f_1(\mathbf{X}_j(t))f_2(\mathbf{X}_j(t))Z_j(t)]}{E[Z(t)]}\,dN_j(t)$$

and

$$\begin{aligned} I_{22} = \frac{n-1}{n^2}\sum_j \biggl[\frac{1}{n-1}\sum_{k\neq j}\biggl(&\int_0^\tau \frac{f_1(\mathbf{X}_k(t))f_2(\mathbf{X}_k(t))Z_k(t)}{E[Z(t)]}\,dN_j(t) \\ &- \int_0^\tau \frac{E[f_1(\mathbf{X}_k(t))f_2(\mathbf{X}_k(t))Z_k(t)]}{E[Z(t)]}\,dN_j(t)\biggr)\biggr]. \end{aligned}$$

By Lemma 1 and the definition of $A_n$,

$$\sup_{f_1, f_2\in\mathbb{G}_0} \frac{|I_{21}|}{\|f_1\|\|f_2\|} \lesssim \frac{1}{n^2}\sum_j \sup_{f_1, f_2\in\mathbb{G}_0} \frac{\|f_1\|_\infty\|f_2\|_\infty N_j(\tau)}{\|f_1\|_{L_2}\|f_2\|_{L_2}} \lesssim \frac{A_n^2}{n} = o(1).$$

By Lemma 10 of Huang (1998a), we can show that (proof will be given shortly)

$$\sup_{f_1, f_2\in\mathbb{G}_0} \frac{|I_{22}|}{\|f_1\|\|f_2\|} = o_P(1). \tag{22}$$

Thus we have that

$$\sup_{f_1, f_2\in\mathbb{G}_0} \frac{|I_2|}{\|f_1\|\|f_2\|} = o_P(1). \tag{23}$$

Observe that

$$\begin{aligned}|I_1| &= |I_1(f_1, f_2)| \\ &\lesssim \|f_1\|_\infty \bigg( \int_0^\tau \frac{E_n[f_2^2(\mathbf{X}(t))Z(t)]}{E[Z(t)]}\, d\overline{N}(t) \bigg)^{1/2} \\ &\qquad \times \bigg( \int_0^\tau \{E_n[Z(t)] - E[Z(t)]\}^2\, d\overline{N}(t) \bigg)^{1/2}.\end{aligned}$$

It follows from (21), (23) and the definition of $\|f_2\|^2$ that

$$\sup_{f_2 \in \mathbb{G}_0} \left| \frac{\int_0^\tau \{E_n[f_2^2(\mathbf{X}(t))Z(t)]/E[Z(t)]\}\, d\overline{N}(t)}{\|f_2\|^2} - 1 \right| = o_P(1).$$

Moreover, by checking moments and using the Markov inequality, we get that

$$\int_0^\tau \{E_n[Z(t)] - E[Z(t)]\}^2\, d\overline{N}(t) = O_P\bigg(\frac{1}{n}\bigg).$$

Thus, by the definition of $A_n$ and Lemma 1,

$$\sup_{f_1, f_2 \in \mathbb{G}_0} \frac{|I_1|}{\|f_1\|\|f_2\|} = O_P\bigg(\frac{1}{\sqrt{n}}\bigg) \sup_{f_1, f_2 \in \mathbb{G}_0} \frac{\|f_1\|_\infty \|f_2\|}{\|f_1\|_{L_2}\|f_2\|} = O_P\bigg(\frac{A_n}{\sqrt{n}}\bigg) = o_P(1).$$

This completes the proof. □

PROOF OF (22). Let $C$ denote an upper bound of $1/E[Z(t)]$ and fix $\varepsilon > 0$. For each $1 \le j \le n$, by conditioning on $N_j(\cdot)$ and applying Lemma 10 of Huang (1998a) [we use the exponential bound in its proof], we obtain that uniformly in $f_1, f_2 \in \mathbb{G}_0$,

$$\begin{aligned}&\bigg| \frac{1}{n-1} \sum_{k \ne j} \bigg( \int_0^\tau \frac{f_1(\mathbf{X}_k(t)) f_2(\mathbf{X}_k(t)) Z_k(t)}{E[Z_k(t)]}\, dN_j \\ &\qquad - \int_0^\tau \frac{E[f_1(\mathbf{X}_k(t)) f_2(\mathbf{X}_k(t)) Z_k(t)]}{E[Z_k(t)]}\, dN_j \bigg) \bigg| \\ &\quad \le \varepsilon \bigg( \int_0^\tau \frac{E[f_1^2(\mathbf{X}_k(t)) Z_k(t)]}{E[Z_k(t)]}\, dN_j \bigg)^{1/2} \bigg( \int_0^\tau \frac{E[f_2^2(\mathbf{X}_k(t)) Z_k(t)]}{E[Z_k(t)]}\, dN_j \bigg)^{1/2},\end{aligned}$$

except on an event with probability bounded by

$$\begin{aligned}P_n = 2 \sum_{l=1}^\infty \bigg\{ &\exp\bigg[ -\frac{\varepsilon^2}{16C} \bigg(\frac{n}{A_n^2}\bigg) \bigg(\frac{1}{18}\bigg) \bigg(\frac{3}{2}\bigg)^{2l} \bigg] \\ &+ \exp\bigg[ -\frac{3\varepsilon}{16C} \bigg(\frac{n}{A_n^2}\bigg) \bigg(\frac{1}{6}\bigg) \bigg(\frac{3}{2}\bigg)^{l} \bigg] \bigg\}.\end{aligned}$$

Thus, except on an event with probability bounded by $nP_n$, which tends to zero as $n \to \infty$ under the assumption that $\lim_n A_n^2(\log n)/n = 0$,

$$\begin{aligned}\Bigg|\frac{1}{n}\sum_j\bigg(\frac{1}{n-1}\sum_{k\neq j}\bigg(\int_0^\tau \frac{f_1(\mathbf{X}_k(t))f_2(\mathbf{X}_k(t))Z_k(t)}{E[Z_k(t)]}\,dN_j \\ - \int_0^\tau \frac{E[f_1(\mathbf{X}_k(t))f_2(\mathbf{X}_k(t))Z_k(t)]}{E[Z_k(t)]}\,dN_j\bigg)\bigg)\Bigg| \\ \le \varepsilon\bigg(\int_0^\tau \frac{E[f_1^2(\mathbf{X}_k(t))Z_k(t)]}{E[Z_k(t)]}\,d\overline{N}\bigg)^{1/2}\bigg(\int_0^\tau \frac{E[f_2^2(\mathbf{X}_k(t))Z_k(t)]}{E[Z_k(t)]}\,d\overline{N}\bigg)^{1/2},\end{aligned}$$

uniformly in $f_1, f_2 \in \mathbb{G}_0$. This together with (21) yields the desired result. □

Proof of Lemma 5. Note that

$$\begin{aligned}&(E_n^Z(\phi_i h_n))(t) - (E^Z(\phi_i h_n))(t) \\ &\quad = \frac{(E_n - E)(\phi_i(\mathbf{X}(t)h_n(\mathbf{X}(t))Z(t))}{E_n[Z(t)]} \\ &\qquad + E[\phi_i(\mathbf{X}(t))h_n(\mathbf{X}(t))Z(t)]\bigg(\frac{1}{E_n[Z(t)]} - \frac{1}{E[Z(t)]}\bigg).\end{aligned}$$

Hence

$$\begin{aligned}&\int_0^\tau [E_n^Z(\phi_i h_n) - E^Z(\phi_i h_n)]^2\,d\overline{N} \\ &\quad \le 2\int_0^\tau \frac{[(E_n - E)(\phi_i(\mathbf{X}(t))h_n(\mathbf{X}(t))Z(t))]^2}{\{E_n[Z(t)]\}^2}\,d\overline{N}(t) \\ &\qquad + 2\int_0^\tau E\{[\phi_i(\mathbf{X}(t))h_n(\mathbf{X}(t))Z(t)]^2\}\bigg(\frac{1}{E_n[Z(t)]} - \frac{1}{E[Z(t)]}\bigg)^2\,d\overline{N}(t).\end{aligned}$$

Observe that $1/E[Z(t)] \le 1/E[Z(\tau)] < \infty$ and $1/E_n[Z(t)] \le 1/E_n[Z(\tau)] = O_P(1)$ for $0 \le t \le \tau$. Moreover, the functions $h_n$, $n \ge 1$ are uniformly bounded. Consequently,

$$\sum_i \int_0^\tau [E_n^Z(\phi_i h_n) - E^Z(\phi_i h_n)]^2\,d\overline{N} = O_P(J_1 + J_2),$$

where

$$J_1 = \sum_i \int_0^\tau [(E_n - E)(\phi_i(\mathbf{X}(t))h_n(\mathbf{X}(t))Z(t))]^2\,d\overline{N}(t)$$

and

$$J_2 = \sum_i \int_0^\tau (E^Z\phi_i^2)(t)\{E_n[Z(t)] - E[Z(t)]\}^2\,d\overline{N}(t).$$

We have that $J_1 \le 2J_{11} + 2J_{12}$, where

$$J_{11} = \sum_i \frac{1}{n} \sum_j \int_0^\tau \bigg( \frac{1}{n} \sum_{k \ne j} \{(\phi_i(\mathbf{X}_k(t))h_n(\mathbf{X}_k(t))Z_k(t) \\ - E[\phi_i(\mathbf{X}_k(t))h_n(\mathbf{X}_k(t))Z_k(t)]\} \bigg)^2 dN_j(t)$$

and

$$J_{12} = \sum_i \frac{1}{n} \sum_j \int_0^\tau \bigg( \frac{1}{n} \{\phi_i(\mathbf{X}_j(t))h_n(\mathbf{X}_j(t))Z_j(t) \\ - E[\phi_i(\mathbf{X}_j(t))h_n(\mathbf{X}_j(t))Z_j(t)]\} \bigg)^2 dN_j(t)$$

By independence and the fact that $E[Z(t)]$ is bounded away from zero and infinity uniformly in $t \in [0, \tau]$,

$$\begin{aligned} E(J_{11}) &= \sum_i \frac{(n-1)}{n^2} \int_0^\tau E\Big( \{\phi_i(\mathbf{X}(t))h_n(\mathbf{X}(t))Z(t) \\ &\qquad - E[\phi_i(\mathbf{X}(t))h_n(\mathbf{X}(t))Z(t)]\}^2 \Big)\, dE[N(t)] \\ &\le \sum_i \frac{(n-1)}{n^2} \int_0^\tau E\{[\phi_i(\mathbf{X}(t)h_n(\mathbf{X}(t))Z(t)]^2\}\, dE[N(t)] \\ &\lesssim \sum_i \frac{1}{n} \int_0^\tau \frac{E\{[\phi_i(\mathbf{X}(t)Z(t)]^2\}}{E[Z(t)]}\, dE[N(t)] \\ &\sim \sum_i \frac{1}{n} \|\phi_i\|^2 = \frac{N_n}{n}. \end{aligned}$$

On the other hand,

$$\begin{aligned} E(J_{12}) &= \sum_i \frac{1}{n^2} E\Big( \int_0^\tau \{\phi_i(\mathbf{X}_j(t))h_n(\mathbf{X}_j(t))Z_j(t) \\ &\qquad - E[\phi_i(\mathbf{X}_j(t))h_n(\mathbf{X}_j(t))Z_j(t)]\}^2)\, dN_j(t) \Big) \\ &\lesssim \sum_i \frac{1}{n^2} \|\phi_i\|_\infty^2 E[N(\tau)] \\ &\lesssim \sum_i \frac{A_n^2}{n^2} \|\phi_i\|_{L_2}^2 E[N(\tau)]. \end{aligned}$$

By Lemma 1, this is bounded above by a constant multiple of

$$\sum_i \frac{A_n^2}{n^2} \|\phi_i\|^2 E[N(\tau)] = O\bigg( \frac{A_n^2 N_n}{n^2} \bigg) = O\bigg( \frac{N_n}{n} \bigg),$$

provided that $A_n^2/n = O(1)$. Consequently, $E(J_1) = O(N_n/n)$ and hence $J_1 = O_P(N_n/n)$. Similarly, we have that

$$E(J_2) \lesssim \sum_i \frac{1}{n}\|\phi_i\|^2 = \frac{N_n}{n}$$

and hence that $J_2 = O_P(N_n/n)$. This completes the proof of the lemma. □

PROOF OF LEMMA 4. Note that $\widetilde{\mathbb{G}} \subset \mathbb{H} = \{h \in \mathbb{H}_0, \langle h, 1\rangle = 0\}$. Since

$$E_n^Z[(f_1 - E_n^Z f_1)(f_2 - E_n^Z f_2)] - E_n^Z[(f_1 - E^Z f_1)(f_2 - E^Z f_2)]$$
$$= -(E_n^Z f_1 - E^Z f_1)(E_n^Z f_2 - E^Z f_2),$$

we have that

$$\langle f_1, f_2\rangle_{0n} - \langle f_1, f_2\rangle_0 = I_1 + I_2,$$

where

$$I_1 = -\int_0^\tau (E_n^Z f_1 - E^Z f_1)(E_n^Z f_2 - E^Z f_2)\, d\overline{N}$$

and

$$I_2 = \int_0^\tau E_n^Z[(f_1 - E^Z f_1)(f_2 - E^Z f_2)]\, d\overline{N}$$
$$- \int_0^\tau E^Z[(f_1 - E^Z f_1)(f_2 - E^Z f_2)]\, dEN.$$

The same argument as in the proof of Lemma 3 gives that

$$\sup_{f_1, f_2 \in \widetilde{\mathbb{G}}} \frac{|I_2|}{\|f_1\|_0\|f_2\|_0} = o_p(1).$$

It remains to show that

$$\sup_{f_1, f_2 \in \widetilde{\mathbb{G}}} \frac{|I_1|}{\|f_1\|_0\|f_2\|_0} = o_p(1). \tag{24}$$

Let $\{\phi_i\}$ be an orthonormal basis of $\widetilde{\mathbb{G}}$ relative to $\langle\cdot,\cdot\rangle$. For $f_1, f_2 \in \widetilde{\mathbb{G}}$, write $f_1 = \sum_i a_i\phi_i$ and $f_2 = \sum_i b_i\phi_i$. Thus $\|f_1\|^2 = \sum_i a_i^2$ and $\|f_2\|^2 = \sum_i b_i^2$. Applying the Cauchy–Schwarz inequality twice, we obtain that

$$|I_1| = \left|\sum_{i,j} a_i b_j \int_0^\tau (E_n^Z\phi_i - E^Z\phi_i)(E_n^Z\phi_j - E^Z\phi_j)\, d\overline{N}\right|$$
$$\le \left\{\sum_{i,j} a_i^2 b_j^2\right\}^{1/2} \left\{\sum_{i,j}\left(\int_0^\tau (E_n^Z\phi_i - E^Z\phi_i)(E_n^Z\phi_j - E^Z\phi_j)\, d\overline{N}\right)^2\right\}^{1/2}$$
$$\le \|f_1\|\|f_2\| \sum_i \int_0^\tau (E_n^Z\phi_i - E^Z\phi_i)^2\, d\overline{N}.$$

Thus, by Lemmas 1 and 2,

$$\sup_{f_1, f_2 \in \widetilde{\mathbb{G}}} \frac{|I_1|}{\|f_1\|_0 \|f_2\|_0} \lesssim \sum_i \int_0^\tau (E_n^Z \phi_i - E^Z \phi_i)^2 d\overline{N}$$

and hence (24) follows from Lemma 5. This completes the proof. □

## REFERENCES

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.

ANDERSEN, P. K. and GILL, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist*. **10** 1100–1120.

BRESLOW, N. E. (1972). Discussion of "Covariance analysis of censored survival data," by D. R. Cox. *J. Roy. Statist. Soc. Ser. B* **34** 216–217.

BRESLOW, N. E. (1974). Covariance analysis of censored survival data. *Biometrics* **30** 89–99.

COX, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34** 187–220.

COX, D. R. and OAKES, D. (1984). *Analysis of Survival Data*. Chapman and Hall, London.

DABROWSKA, D. M. (1997). Smoothed Cox regression. *Ann. Statist.* **25** 1510–1540.

DE BOOR, C. (1978). *A Practical Guide to Splines*. Springer, New York.

DEVORE, R. A. and LORENTZ, G. G. (1993). *Constructive Approximation*. Springer, Berlin.

EFRON, B. (1977). The efficiency of Cox's likelihood function for censored data. *J. Amer. Statist. Assoc.* **72** 557–565.

FLEMING, T. R. and HARRINGTON, D. P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

GENTLEMAN, R. and CROWLEY, J. (1991). Local full likelihood estimation for the proportional hazards model. *Biometrics* **47** 1283–1296.

GRAY, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *J. Amer. Statist. Assoc.* **87** 942–951.

GU, C. (1996). Penalized likelihood hazard estimation: a general procedure. *Statist. Sinica* **6** 861–876.

HUANG, J. Z. (1998a). Projection estimation for multiple regression with application to functional ANOVA models. *Ann. Statist.* **26** 242–272.

HUANG, J. Z. (1998b). Concave extended linear modeling: a theoretical synthesis. Unpublished manuscript.

HUANG, J. Z. and STONE, C. J. (1998). The $L_2$ rate of convergence for event history regression with time-dependent covariates. *Scand. J. Statist.* **25** 603–620.

KALBFLEISCH, J. D. and PRENTICE, R. L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

KOOPERBERG, C. and CLARKSON, D. B. (1997). Hazard regression with interval-censored data. *Biometrics* **53** 1485–1494.

KOOPERBERG C., STONE, C. J. and TRUONG, Y. K. (1995a). Hazard regression *J. Amer. Statist. Assoc.* **90** 78–94.

KOOPERBERG C., STONE, C. J. and TRUONG, Y. K. (1995b). The $L_2$ rate of convergence for hazard regression. *Scand. J. Statist.* **22** 143–157.

LEBLANC, M. and CROWLEY, J. (1999). Adaptive regression splines in the Cox model. *Biometrics* **55** 204–213.

O'SULLIVAN, F. (1988). Non-parametric estimation of relative risk using splines and cross-validation. *SIAM J. Sci. Statist. Comput.* **9** 531–542.

O'SULLIVAN, F. (1993). Nonparametric estimation in the Cox model. *Ann. Statist.* **21** 124–145.

SCHUMAKER, L. L. (1981). *Spline Functions: Basic Theory*. Wiley, New York.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

SLEEPER, L. A. and HARRINGTON, D. P. (1990). Regression splines in the Cox model with application to covariate effects in liver disease. *J. Amer. Statist. Assoc.* **85** 941–949.

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1348–1360.

STONE, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *Ann. Statist.* **22** 118–184.

STONE, C. J., HANSEN, M., KOOPERBERG, C. and TRUONG, Y. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *Ann. Statist.* **25** 1371–1470.

TIBSHIRANI, R. and HASTIE, T. (1987). Local likelihood estimation. *J. Amer. Statist. Assoc.* **82** 559–568.

WAHBA G., WANG, Y., GU, C., KLEIN, R. and KLEIN, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy. *Ann. Statist.* **23** 1865–1895.

J. Z. HUANG
DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104-6302
E-MAIL: jianhua@stat.wharton.upenn.edu

C. KOOPERBERG
DIVISION OF PUBLIC HEALTH SCIENCES
FRED HUTCHINSON CANCER RESEARCH CENTER
SEATTLE, WASHINGTON 98109-1024
E-MAIL: clk@fhcrc.org

C. J. STONE
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94729-3860
E-MAIL: stone@stat.berkeley.edu

Y. K. TRUONG
DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27599-7400
E-MAIL: truong@bios.unc.edu