ISSN 1471-0498

DEPARTMENT OF ECONOMICS

DISCUSSION PAPER SERIES

Asymptotics for Sieve Estimators of Hazard Rates: Estimating Hazard Functionals

James Lewis Wolter

Number 760
September 2015

Manor Road Building, Oxford OX1 3UQ

Asymptotics for Sieve Estimators of Hazard Rates: Estimating Hazard Functionals

James Lewis Wolter

Department of Economics, University of Oxford

Oxford-Man Institute

August 7, 2015

Abstract

This paper derives asymptotics for functionals of a hazard model with an exposure-time effect and time-varying covariates. A semi-nonparametric sieve maximum likelihood estimator of a competing risks model based on the Cox proportional hazard is considered. Consistency of the estimator and its rate of convergence in the Fisher norm are derived. These results are prerequisites for asymptotic normality of plug-in estimators of hazard functionals. This provides an inference procedure for a large class of functionals including the conditional probability of events and various asset pricing formulas for defaultable securities. Asset pricing formulas in this class include the value of mortgages, insurance contracts, bonds, swaps and other options.

1 Introduction

In many situations involving economic durations, hazards are a standard way of modeling events. In these cases, the conditional probability of an event (not simply the parameters of the hazard) is often of interest. For an observation with a hazard function $\lambda(t, Z(t))$ influenced by time-varying

covariates $Z(t)$, the probability of default in the time interval $[0, T]$ is:

$$1 - E\left[\exp\left(-\int_0^T \lambda\left(u, Z\left(u\right)\right) du\right) \middle| \mathcal{F}_0\right]. \tag{1}$$

Here, $\mathcal{F}_0$ is the information available at time zero. This conditional probability is a functional of the underlying hazard.

In many economic situations, there are several types of event which can end a duration. Mortgages can end with prepayment or default. Corporations can default, merge or end in other ways. Unemployment durations can end with employment or dropping out of the labor force. In these cases, more complicated conditional probabilities are of interest. For example, the probability of no prepayment or default within a certain time interval. These situations can be modeled with hazards using a competing risks framework. In competing risks, each type of potential event is modeled with a hazard function. This setup results in closed form expressions as in (1), but for more complicated conditional probabilities. These formulas are also functionals of the underlying hazards. Several examples of interest are given below.

While it has much broader applicability, estimation of hazard functionals such as (1) is of considerable interest for analyzing defaultable financial contracts. For example, investors are interested in the probability of mortgage holders prepaying when default is possible. This conditional probability has a closed form described in the sequel. Additionally, when events are modeled using hazard rates, asset pricing formulas for defaultable securities take a conditional expectation form which is similar to (1). This includes pricing formulas for mortgage contracts, insurance contracts, bonds, swaps or other options. Estimating these formulas provides a measure of risk premia. See Duffie and Singleton (2003), Bielecki and Rutkowski (2004) or Singleton (2006) for overviews of asset pricing formulas for defaultable securities.

The main contribution of this paper is to provide a methodology for inference on functionals such as (1) when hazard functions takes a semi-nonparametric form. This involves deriving asymptotic distributions of functional estimators. In the sequel, technical results are presented for the Cox proportional hazard form. However, the method can be straightforwardly extended to more general semi-nonparametric cases. Duffie, Saita and Wang (2007) apply similar parametric

estimation to corporate default data.

There exists an extensive literature on nonparametric and semi-nonparametric estimation of hazard functions using a kernel approach. See Nielsen and Linton (1995); Linton, Nielsen and van de Geer (2003) for the fully nonparametric case and Nielsen, Linton and Bickel (1998); van den Berg, Janys, Mammen and Nielsen (2014) for the semi-nonparametric case. While allowing for very flexible hazard specifications, these estimators cannot produce asymptotic normality of functionals such as (1). Kernel based results are local in nature, whereas functionals with a conditional expectation form depend on the entire hazard function. A different method is needed.

The goal of this paper is to derive an estimation approach which allows the underlying hazard functions to be flexible and can produce asymptotic distributions for the functionals of interest. The estimators ideally will have a straightforward asymptotic normal distribution. These goals are achieved by estimating hazard functions with sieve semi-nonparametric maximum likelihood. Nonparametric parts of the hazard are approximated with basis functions. Estimates are derived by maximizing a point process likelihood over an appropriate sieve space using these basis functions. See Chen (2007) for an overview of sieve methods.

There are several general results for deriving consistency and convergence rates in the sieve literature. See, for example, Shen and Wong (1994), Chen and Shen (1998) or Ai and Chen (2003) among others. However, these results are not well suited for the point process likelihood used below. Because of this, we modify less general methods designed for our specific case. In the sequel, consistency is derived for the competing risks model. This consistency result extends the approach in Karr (1987), who considers a more restricted case. Using consistency as an input, rates of convergence are then derived. Rates are given for convergence in the Fisher norm. The Fisher norm measures the distance between parameters using an information matrix based on the likelihood. This measure of distance is convenient because the likelihood is used in estimation. Our convergence rate result is based on Wong and Severini (1991). We extend their infeasible method to the feasible sieve case. The convergence rate in the Fisher norm is shown to be $o_p\left(n^{-1/4}\right)$. This is a requirement for asymptotic normality when using sieve estimation.

Once consistency and convergence rates are derived, Chen and Liao (2014) and Chen, Liao and

Sun (2014) can be used to derive asymptotic normality of plug-in estimators of functionals. Their approach allows us to interpret estimation as simple parametric maximum likelihood where the number of parameters is controlled by the amount of data. In order to apply these results, the $o_p\left(n^{-1/4}\right)$ convergence rate must be in a Hilbert space. We verify that the Fisher norm has this required structure. This justifies using the Fisher norm to describe convergence rates. The main purpose of our rate result is to facilitate asymptotic normality in the next step.

The most common estimation approach for a Cox hazard function uses partial likelihood (see Andersen, Borgan, Gill and Keiding (1993) or Martinussen and Scheike (2010)). These methods could potentially be used to derive asymptotic distributions of functionals using the functional delta method. In the competing risks case, the resulting distribution would depend on a combination of several Gaussian processes with complicated covariance structures. The covariances structures are unknown and would need to be estimated. Compared with this approach, asymptotic distributions derived with the sieve method are normal and the estimation has a parametric maximum likelihood interpretation. The sieve likelihood approach can also be extended to other semi-nonparametric models in a straightforward way. An eventual goal is to estimate functionals of fully nonparametric multiplicative hazards using the sieve method. See Linton et al. (2003) for kernel estimation of this case. Partial likelihood methods cannot be extended to this case because they depend on the Cox form.

Simulations are conducted to examine the performance of the proposed inference procedure in finite samples. Confidence intervals have finite sample distortion which depends on the sieve basis functions, the amount of data and the underlying hazard form. In theory, the flexibility of the sieve spaces must increase quickly with the sample size for asymptotic normality to follow. This is supported in the simulations. Coverage probability decreases markedly when the amount of data is increased beyond a certain point while fixing the sieve basis functions. Bootstrap confidence intervals are also considered. This refinement significantly improves performance in finite samples.

The remainder of the paper is organized as follows. Section 2 describes the hazard model used throughout the paper. Several functionals of interest are also presented. Section 3 presents the

estimation approach and derives consistency results. Section 4 derives rate-of-convergence results for the estimator in the Fisher norm. Section 5 describes how the consistency and convergence rate results facilitate asymptotic normality for functionals of interest. A simulation study examining the performance of the proposed estimators is presented. Section 6 concludes. All proofs are presented in the Appendix.

2 Models, Examples and Preliminaries

Here we describe how the hazard models considered in this paper are constructed. Observations are indexed by $i \in \mathbb{N}$. Each observation is at risk over a fixed time interval of length T. Let $\{ Z^i(t) | \, t \in [0,T] \}$ be d covariate stochastic processes specific to each observation. The support $\mathcal{Z}$ of $Z^i(t)$ is assumed to be the same for all $t \in [0,T]$. More assumptions will be made on these covariates below. Our construction of random times follows Bielecki and Rutkowski (2004) Example 9.1.5. In the sequel, $\alpha_0(\cdot)$ is the hazard function with covariates taken as arguments. We make the following assumption throughout:

(A1) *(i)* $\alpha_0 : [0,T] \times \mathcal{Z} \to \mathbb{R}$ *is a function such that*

$$\alpha_0\left(t, Z_t^i\right) = \exp\left(h_0(t) + \beta_0' Z^i(t)\right)^{1}$$

(ii)

$$\begin{aligned} \inf_{(t,z)\in[0,T]\times\mathcal{Z}} \alpha_0(t,z) &= \underline{C} > 0, \\ \sup_{(t,z)\in[0,T]\times\mathcal{Z}} \alpha_0(t,z) &= \overline{C} < \infty. \end{aligned}$$

Random times τ_i are defined as:

$$\begin{aligned} \Gamma_t^i &= \int_0^t \exp\left(h_0(u) + \beta_0' Z^i(u)\right) du, \qquad (2) \\ \tau_i &= \inf\left\{ t \in \mathbb{R}_+ | \, \Gamma_t^i \geq \eta_i \right\}.^{2} \end{aligned}$$

η_i is an independent, standard exponentially distributed random variable. As shown in BR Section 6.6, the choice of distribution for η_i is not arbitrary and must be standard exponential to produce a hazard model. We assume the covariates Z_t^i and the random times τ_i are observed. The covariates are observed up until $\tau_i \wedge T$. This setup produces a random time τ_i with hazard rate $\exp\left(h_0\left(u\right) + \beta_0' Z^i\left(u\right)\right)$ that is equivalent to standard hazard models in the literature.

In many economic situations, several types of event can end a duration. Because of the importance of these cases, we extend the basic hazard model defined above to a competing risk framework. This is done by constructing R random events as in (2). Each of these represents a different type of event which can end the duration. They then "compete" to see which type happens first, ending the duration. These R random times will be indexed as τ_i^j, $j \in \{1, \ldots, R\}$ for each observation i. Each type j has its own set of parameters $\left(h_0^j\left(t\right), \beta_0^j\right)$. Each τ_i^j is constructed with an independent η_i^j. All τ_i^j for an observation i share the same set of covariates Z_t^i. The first event to happen is defined as $\tau_i^{(1)} = \tau_i^1 \wedge \cdots \wedge \tau_i^R$. Which type corresponds to $\tau_i^{(1)}$ is observed. All other τ_i^j are censored by $\tau_i^{(1)}$. The covariates Z_t^i are observed up until $\tau_i^{(1)} \wedge T$. The notation $\tau_i = \left(\tau_i^1, \ldots, \tau_i^R\right)$ is used for the vector of event times. This setup reduces to the single event case when $R = 1$. It will turn out that competing risks can be analyzed in much the same way as the $R = 1$ case.

In the competing risks setup, the process,

$$\exp\left(h_0^j\left(t\right) + \beta_0^{j\prime} Z^i\left(t\right)\right) \mathbf{1}_{\left\{\tau_i^{(1)} \geq t\right\}},$$

is the hazard rate for event type j. The term $\mathbf{1}_{\left\{\tau_i^{(1)} \geq t\right\}}$ is included because the single spell case is considered. After $\tau_i^{(1)}$ there is no possibility of further events and the hazard drops to zero. See Fleming and Harrington (1991) or Andersen et al. (1993) for more details on the definition of a hazard model. We will write $N^{ij}\left(t\right) = \mathbf{1}\left\{\tau_i^{(1)} \leq t, \tau_i^{(1)} = \tau_i^j\right\}$ for the process that indicates $\tau_i^{(1)}$ has happened and its type is j. Define also,

$$\Lambda_t^{ij} = \int_0^t \exp\left(h_0^j\left(u\right) + \beta_0^{j\prime} Z^i\left(u\right)\right) \mathbf{1}_{\left\{\tau_i^{(1)} \geq u\right\}} du,$$

$$M_t^{ij} = N_t^{ij} - \Lambda_t^{ij}.$$

The asymptotic results that follow depend on M_t^{ij} being continuous-time martingales. This result is standard in the literature and the martingale structure will be used in the sequel. See the appendix for more details. See Fleming and Harrington (1991) and Bielecki and Rutkowski (2004) for excellent accounts of the martingale theory of counting processes.

2.1 Functional Examples

There are many functionals of interest when considering the hazard framework presented above. We now present several motivating examples.

Example 1 β_r, $r \in \{1, \ldots, d\}$.

Example 2 $\int_a^b \exp(h(s))\, ds$, $0 \leq a < b \leq T$.

These first two examples are simple starting points. Additionally, as shown in a related paper Wolter (2015), sieve methods used for these cases can be extended to more complicated situations of interest.

We now present functionals related to event probabilities and risk premia. In the following, the focus is on cases with one or two types of event. There is no difference in principal between the case with two types and an arbitrary finite number. All of the following functionals depend on random events being described by hazard models as in the setup above. For more details see Bielecki and Rutkowski (2004). In what follows, the index i is suppressed when referring to a generic observation. This convention is used through the paper.

When $R = 2$, the more compact notation $\lambda_t^1\left(h^1, \beta^1\right) = \exp\left(h^1(t) + \beta^{1\prime} Z^i(t)\right)$ and the corresponding $\lambda_t^2\left(h^2, \beta^2\right)$ are used. In cases with only one event type, this will be written $\lambda_t(h, \beta)$. The relevant information at time t is:

$$\mathcal{F}_t = \sigma\left\{ Z^i(s), \mathbf{1}_{\left\{\tau_i^{(1)} \geq s\right\}} \middle| s \in [0, t] \right\}.$$

Example 3

$$\begin{aligned} & P\left\{\tau > T \middle| \mathcal{F}_t\right\} \\ = \; & E\left[\exp\left(-\int_t^T \lambda_u\left(h,\beta\right) du\right) \middle| \mathcal{F}_t\right]. \end{aligned}$$

Example 4

$$\begin{aligned} & P\left\{\tau^1 > T, \tau^2 > T \middle| \mathcal{F}_t\right\} \\ = \; & E\left[\exp\left(-\int_t^T \lambda_u^1\left(h^1,\beta^1\right) + \lambda_u^2\left(h^2,\beta^2\right) du\right) \middle| \mathcal{F}_t\right]. \end{aligned}$$

Example 5

$$\begin{aligned} & P\left\{\tau^{(1)} = \tau^1, \tau^1 \leq T \middle| \mathcal{F}_t\right\} \\ = \; & E\left[\int_t^T \left\{\exp\left(-\int_t^u \left[\lambda_s^1\left(h^1,\beta^1\right) + \lambda_s^2\left(h^2,\beta^2\right)\right] ds\right) \lambda_u^1\left(h^1,\beta^1\right)\right\} du \middle| \mathcal{F}_t\right]. \end{aligned}$$

Example 3 gives the probability at time t of surviving past time T, conditional on $\mathcal{F}_t$. This and other related examples implicitly assume that the event $\tau_i^{(1)}$ has not happened at t. This is suppressed in the notation. Example 4 gives the probability that neither of the events happens before T, conditional on $\mathcal{F}_t$. Finally, example 5 gives the probability that event 1 happens first, and happens before T, conditional on $\mathcal{F}_t$. These expressions are of interest in many economic situations such as those described in the introduction.

We now present examples related to the pricing of defaultable securities. These pricing formulas are based on the no-arbitrage approach initiated by Black and Scholes (1973) and Merton (1973). There is an extensive literature in this area. A sampling of relevant citations are Duffie, Schroder and Skiadas (1996); Duffie and Singleton (1999), (2003); Collin-Dufresne, Goldstein and Hugonnier (2004); Bielecki and Rutkowski (2004); and Singleton (2006). See these citations for more details.

In the following formulas, r_t is the short interest rate and $g\left(\cdot\right)$ and $J\left(\cdot,\cdot\right)$ are functions representing the specifics of the asset contract.

Example 6

$$E_{\mathbb{P}}\left[\exp\left(-\int_t^T \left\{r_u + \lambda_u^1\left(h^1,\beta^1\right) + \lambda_u^2\left(h^2,\beta^2\right)\right\} du\right) g\left(Z_T\right)\middle|\mathcal{F}_t\right].$$

Example 7

$$E_{\mathbb{P}}\left[\int_t^T J\left(Z_u,u\right)\exp\left(-\int_t^u \left\{r_s + \lambda_s^1\left(h^1,\beta^1\right) + \lambda_s^2\left(h^2,\beta^2\right)\right\} ds\right)\lambda_u^1\left(h^1,\beta^1\right) du\middle|\mathcal{F}_t\right].$$

These examples are asset pricing formulas where the expectation is taken with respect to the physical measure $\mathbb{P}$ instead of the pricing measure $\mathbb{Q}$. This is emphasized in the notation above by including $\mathbb{P}$ subscripts on the expectations. These examples are of interest in examining a certain notion of risk premia. The difference between these formulas in the $\mathbb{P}$ and $\mathbb{Q}$ measure give a metric for the premia that specific contracts trade at with respect to the objective probability of future events. We observe the value of these contracts in the market and therefore we observe the formulas in $\mathbb{Q}$. If we subtract estimates of Example 6 or 7 from the corresponding traded value of the security, this gives a measure of risk premia.

Under the pricing measure $\mathbb{Q}$, Example 6 gives the value at t of a contingent claim paying $g\left(Z_T\right)$ at T, provided neither event happens before T. Example 7 gives the price of a contingent claim at t which pays out only if $\left\{\tau^{(1)} = \tau^1, \tau^1 \leq T\right\}$. If this is the case, the claim pays the amount $J\left(Z_{\tau^1},\tau^1\right)$ at the time of the event τ^1. In either of these formulas, the $R = 1$ case can be produced by simply removing $\lambda_s^2\left(h^2,\beta^2\right)$.

Once the pricing formulas presented above are derived, it is possible to combine them to represent more complex financial products. For example, the value of a mortgage, insurance contract, bond, swap or other options. See Bielecki and Rutkowski (2004) for more on pricing formulas in this area.

2.2 Asymptotic Distribution Approaches

In estimation, we often would like to consider the functionals presented above, not only the parameters of the underlying hazard. If this is the goal, to conduct inference we need asymptotic

distribution theory for estimators of the functionals. The form of this distribution theory will depend on how estimation is conducted.

A standard way of estimating the Cox model is with partial likelihood (See Martinussen and Scheike (2010) or Andersen et al. (1993) for a review). One approach to deriving inference for functionals is to use this asymptotic theory. Partial likelihood methods estimate the parameters $\Lambda_0(t) = \int_0^t \exp\left(h_0^j(s)\right) ds$ and β_0^j. For these estimators $\left(\widehat{\Lambda}^j(t), \widehat{\beta}^j\right)$, it has been shown that $\sqrt{n}\left(\widehat{\Lambda}^j(t) - \Lambda_0^j(t)\right)$ converges to a Gaussian processes and $\sqrt{n}\left(\widehat{\beta}^j - \beta_0\right)$ converges to a normal distribution. These asymptotic distributions are correlated. Each type of event has an estimator with this asymptotic form.

Initial partial likelihood results could potentially be used to characterize the distributions of functionals using the functional delta method. This would require finding Hadamard derivatives. Assuming the functionals are Hadamard differentiable, asymptotic distributions would be combinations of the asymptotic objects described in the previous paragraph. To the best of the author's knowledge, exact conditions for this program have not been derived in the literature for the functionals of interest. See Andersen et al. (1993) chapter II or van der Vaart (1998) chapter 20 for details on the functional delta method. See Andersen et al. (1993) chapter VII for applications using the Cox model and simpler cases than those presented here.

It is not the goal of this paper to pursue the functional delta method approach. Instead, we derive a sieve maximum likelihood estimator of $\left(h_0^j(t), \beta_0^j\right)$ using the full likelihood. The functional is then estimated by "plugging-in" the first stage estimators. In the sequel, this plug-in estimator is shown to have a straightforward asymptotically normal distribution. The first-stage estimators come from simple parametric maximum likelihood estimation. The parametrization is determined by the sieve basis functions which grow with the amount of data. Asymptotic distributions for plug-in estimators are derived with a delta method type argument adjusted for the growing number of parameters. This is more straightforward than the functional delta method and involves less complicated asymptotic distributions and approximations.

Nothing about Examples 3-7 depends on the Cox form. The hazard functions can take any other form and the functionals given above are still valid. An eventual goal is to extend the

sieve methods developed in this paper to fully nonparametric multiplicative hazards. This would involve a nontrivial technical extension of the present work. In principal, estimation in this case is a straightforward extension of what follows. The basic form of the point process likelihood used is the same for all hazard cases. Our results are a starting point for analyzing more complicated models. Partial likelihood methods cannot be applied more generally as they depend on the Cox form.

Many of the functionals described above take conditional expectations with respect to the underlying covariates Z_t. This requires knowledge of their distribution. A priori knowledge of Z_t is in most situations a strong assumption. Another approach is to additionally estimate Z_t's distribution. Estimation of this structure could then be combined with hazard estimators. This more complicated case raises a number of additional issues related to censoring. In particular, the fact that events censor our observation of Z_t creates complications involving survivorship bias. The analysis of this case is left to future research. We note that having to estimate this additional aspect of the model will reduce accuracy. It is possible that a simpler asymptotic distribution theory such as the one presented here could preform better that functional delta method approximations.

3 Point Process Likelihood Estimation

In this section, a point process likelihood approach is used to estimate the competing risks model using sieves. The model is estimated using the full likelihood instead of the partial likelihood. Our approach is similar to that of Karr (1987). Karr (1987) proves consistency in a similar sieve maximum likelihood case assuming that β_0 is known. We extend these results by deriving consistency when β_0 must be estimated as well. The model is also extended to allow for competing risks. To the best of the author's knowledge, Karr (1987) is the only other paper to use the full likelihood in the semi-nonparametric case when applying sieve estimation. Kernel methods have used the full likelihood (see Linton et al. (2003) and van den Berg et al. (2014)), but are not appropriate for our problem. See Brémaud (1981) or Andersen et al. (1993) for more on point process likelihoods.

Define $\Theta = \mathcal{H} \times B$, the space in which the parameters $\left(h_0^j, \beta_0^j\right)$ are assumed to be contained.

We also define a sequence of subspaces $\Theta_n \subset \Theta$ where $\Theta_n = \mathcal{H}_n \times B$. This sequence is referred to as a sieve. $\mathcal{H}_n$ is a sequence of function spaces such that $\mathcal{H}_n \subset \mathcal{H}_{n+1}$ and $\mathcal{H}_n \subset \mathcal{H}$ for all n. The point process likelihood defined below will be maximized over Θ_n where n corresponds to the number of observations. More specific requirements on the sieve spaces Θ_n are given in the sequel. In an abuse of terminology, we will refer to $\mathcal{H}_n$ as a sieve as well. See Chen (2007) for an overview of the sieve estimation approach.

In the sequel, it is assumed that the parameter space Θ and corresponding sieve spaces Θ_n are the same when estimating $\left(h_0^j, \beta_0^j\right)$ for different types of event τ^j. The same restrictions will be made on these spaces as well. As a result, we do not index Θ_n or Θ by j in what follows. It is possible to change this at the cost of increased notation.

The log of the point process likelihood that is the basis of our estimation is:

$$\log\left[\frac{d\widetilde{P}^j}{dP}\left(h^j, \beta^j\right)\right] = \int_0^T \left(h^j(u) + \beta^{j\prime} Z^i(u)\right) dN_u^{ij} + \int_0^T \left[1 - \exp\left(h^j(u) + \beta^{j\prime} Z^i(u)\right) \mathbf{1}_{\left\{\tau_i^{(1)} \geq u\right\}}\right] du.$$

There is one of these likelihoods for each event type j. This type of likelihood is well known in the literature. See Brémaud (1981) section VI.2 for more technical specifics. We define our estimators as,

$$Q_n^j\left(\widehat{h}^j, \widehat{\beta}^j\right) \geq \sup_{\beta^j, h^j \in \Theta_n} Q_n^j\left(h^j, \beta^j\right) + O_p\left(\varepsilon_n^2\right),$$

where the criterion function is,

$$Q_n^j\left(h^j, \beta^j\right) = \frac{1}{n}\sum_{i=1}^n \left[\int_0^T \left(h^j(u) + \beta^{j\prime} Z^i(u)\right) dN_u^{ij} + \int_0^T \left[1 - \exp\left(h^j(u) + \beta^{j\prime} Z^i(u)\right) \mathbf{1}_{\left\{\tau_i^{(1)} \geq u\right\}}\right] du\right],$$

and $\varepsilon_n \to 0$. Here, $O_p\left(\varepsilon_n^2\right)$ is chosen to be $o_p\left(n^{-1}\right)$ in order for asymptotic normality to hold. ε_n influences the final convergence rate. Estimators are produced for each of the R types of event. Several regulatory assumptions are needed for consistency:

(A2) *(i)* $Z^i(t)$ *are assumed to be i.i.d.* with d covariates *(ii)* $Z^i(t)$ *is a piecewise constant process updating at a finite number of deterministic times* t_l, $l = 0, 1, \ldots, m$.[3] *(iii) For all* $t \in [0, T]$, $Z^i(t)$ *has support* $\mathcal{Z}$ *equal to a compact rectangle normalized to be* $[0,1]^d$. *(iv) The*

distribution of $Z^i(t)$ *has a density function* f *on* $\overline{\mathcal{Z}} = \mathcal{Z}_0 \times \cdots \times \mathcal{Z}_m$ *with full support. (v)* $Z^i(t)$ *are assumed to be càglàd.*[4]

(A3) For each event type j: *(i)* β_0^j *is contained in the open rectangle* $B = (a_1, b_1) \times \cdots \times (a_d, b_d)$. *(ii)* $h_0^j \in \mathcal{H}$*, a set of cáglád functions on* $[0, T]$. *(iii) For all* $h \in \mathcal{H}$

$$C_{\min} < h < C_{\max},$$

for known fixed constants $C_{\min}$*,* $C_{\max}$.

(A4) *(i)* $\mathcal{H}_n$ is comprised of linear combinations of sieve basis functions. *(ii)* $\mathcal{H}_n \subset \mathcal{H}$ and $\cup_n \mathcal{H}_n$ *is dense in* $\mathcal{H}$ *in* L^1. *(iii) For* $h \in \mathcal{H}_n$*,* h *is differentiable and* $C_{\min} < h < C_{\max}$.

(A5) *For* $h \in \mathcal{H}_n$*:*

$$|h'| \leq K_n,$$

where $K_n = O\left(n^{1/2-\rho}\right)$*,* $\rho > 0$ *and* $K_n \to \infty$.

Conditions (A2)-(A3) are needed for technical reasons. Assuming *i.i.d.* is done for simplicity. The exact nature of dependence in a hazard model is case specific. For example, if macro variables are present, asymptotic approximations may require a sampling structure similar to panel data models. The *i.i.d.* case is a natural starting point for these more complicated extensions. Conditions (A2)(ii)-(iv) and (A3) simplify many technical arguments but can likely be relaxed. (A2)(v) and (A3)(ii) are needed for reasons related to the martingale nature of point processes.

Assumptions (A4)-(A5) restrict the functions which comprise the sieve spaces $\mathcal{H}_n$. (A4)(ii) simplifies implementation by allowing for convex optimization of the criterion function, but can be relaxed. (A4)(ii) requires that we can approximate any function in $\mathcal{H}$ in an L^1 sense. The consistency result that follows will be in L^1. (A4)(iii) requires derivatives and boundedness for functions in $\mathcal{H}_n$. (A5) controls the first derivative of functions in the spaces $\mathcal{H}_n$. The first derivative is allowed to grow as we gather more data. In applications, a set of basis functions which determine $\mathcal{H}_n$ must be chosen. There are many possibilities such as polynomials, splines or wavelets. See Chen (2007) for a lengthy discussion of potential sieve choices.

Theorem 8 *Make Assumptions (A1)–(A5). Then for all* $j \in \{1, \ldots, R\}$*:*

$$\widehat{\beta}^j \to \beta_0^j,$$

$$\widehat{h}^j \to^{L^1} h_0^j,$$

almost surely.

The proof of the result shows that, if a sequence $\left(h_n^j, \beta_n^j\right) \in \Theta_n$ satisfies:

$$E\left\{\log\left[\frac{d\widetilde{P}^j}{dP}\left(h_0^j, \beta_0^j\right)\right]\right\} - E\left\{\log\left[\frac{d\widetilde{P}^j}{dP}\left(h_n^j, \beta_n^j\right)\right]\right\} \to 0,$$

then $\left(h_n^j, \beta_n^j\right)$ converges (i.e. $h_n^j \to^{L^1} h_0^j$ and $\beta_n^j \to \beta_0^j$). Therefore, the situation is well-posed. Convergence of $\left(\widehat{h}^j, \widehat{\beta}^j\right)$ then follows by showing

$$E\left\{\log\left[\frac{d\widetilde{P}^j}{dP}\left(h_0^j, \beta_0^j\right)\right]\right\} - E\left\{\log\left[\frac{d\widetilde{P}^j}{dP}\left(\widehat{h}^j, \widehat{\beta}^j\right)\right]\right\} \to 0,$$

almost surely. Identification is proven under our assumptions as part of this program.

Condition (A5) allows the first derivative of the sieve spaces to grow as we gather more data (i.e. $K_n \to \infty$). This is done because we may be unwilling to bound the first derivative on the underlying path $h_0^j(t)$. Arbitrarily large derivatives are allowed in the limit. Another possibility is that $h_0^j(t)$ has discontinuities. In this case, the estimators still converges in L^1. $\widehat{h}^j$ will become increasingly steep around the jumps, becoming discontinuous in the limit.

In the sequel, conditions on $\mathcal{H}$ are strengthened to derive asymptotics for functionals. This will require additional smoothness conditions. In particular, we must uniformly bound the first derivative of functions in $\mathcal{H}$. This uniform bound removes the need for the condition (A5). However, doing this causes us to lose some of our understanding of consistency. K_n can increase so fast that consistency fails. K_n effectively bounds how complex the spaces $\mathcal{H}_n$ are allowed to be depending on how much data there is. If we uniformly bound the derivatives, this technically removes the problem. Eventually K_n will be greater than this bound and will no longer enter the

asymptotic requirements for consistency. But it is still possible for the function space $\mathcal{H}_n$ to be so large compared with n that our estimator is poor. Allowing $\mathcal{H}_n$ to contain functions with very large, but uniformly bounded derivatives leads to an unfavorable bias-variance trade-off when n is relatively small. This problem will go away in the limit, but is still relevant in understanding finite sample properties of our estimators.

4 Fisher Norm and Rate of Convergence

In the previous section, a consistency result for sieve semi-nonparametric estimation of the Cox model was presented. In this section, we use our consistency result as an input in deriving rates of convergence. There are several papers in the literature deriving convergence rates for sieve estimates. See, for example, Shen and Wong (1994), Chen and Shen (1998) or Ai and Chen (2003) among others. However, the types of assumptions necessary for these results are not well suited to our point process likelihood case. In particular, the fact that the likelihood does not have a simple linear regression form complicates many of the required conditions. The corresponding asymptotic normality results from the above cited papers create similar problems.

Because of these complications, we will instead modify the results of Wong and Severini (1991) (hereafter WS) for the sieve case. WS consider the infeasible case of maximum likelihood over an infinite dimensional parameter space. As in that work, we use the likelihood structure of our estimator to measure the rate of convergence. This is done by measuring distance between parameters using a metric based on the information matrix of the likelihood. The metric is called the Fisher norm. An appropriate rate of convergence in the Fisher norm is a prerequisite for deriving asymptotic normality of plug-in estimators. See Shen (1997); Ai and Chen (2003), (2007); Chen (2007); Chen and Pouzo (2009); (2012); and Chen, Tamer and Torgovitsky (2011) for more on the use of related norms in econometric applications of sieve asymptotics. The program that follows derives rates of convergence in this norm, then uses these rates to connect our estimators with asymptotic normality results.

4.1 Fisher Norm

Chen and Liao (2014) and Chen, Liao and Sun (2014) (hereafter CLS) derive conditions under which plug-in sieve estimators of functionals have asymptotically normal distributions. In order to use these results, $\left(\widehat{h}^j, \widehat{\beta}^j\right)$ must converge at the rate $o_p\left(n^{-1/4}\right)$. This is a general requirement for asymptotic normality of functionals using sieve methods (see Chen (2007)). It is also required that the metric under which $\left(\widehat{h}^j, \widehat{\beta}^j\right)$ converges has a Hilbert space structure. Once these requirements are satisfied, the results in CLS can be used to derive asymptotic normality for a large class of functionals.

In deriving convergence rates, we are not concerned with choosing well known norms such at L^2 to measure the distance between parameters. The main utility of the following rate results is facilitating connection with asymptotic normality results. For this we can choose any metric which has the required Hilbert space structure. Because it suits these purposes, we chose the Fisher norm described below.

The Fisher norm describes the distance between parameters in Θ using an information matrix derived from the likelihood. This allows for a straightforward connection with our sieve maximum likelihood estimation approach. In order to describe the norm, we must derive score functions of our likelihood using pathwise derivatives in the parameter space. These types of derivatives are required in the rate-of-converge proof as well. This follows the approach in WS and we adopt their notation.

In the following description of the Fisher norm, we suppress the index j on the parameters (h, β). This is done for notational simplicity. It is understood that the following will be applied to each event type. For any $\alpha_1, \alpha_2 \in \Theta$, define $\alpha_1 - \alpha_2 = (h_1 - h_2, \beta_1 - \beta_2)$. Scalar multiplication of elements in Θ is defined in the obvious way. For any $\alpha_m \in \Theta$, define $\phi_m = (\alpha_m - \alpha_0) = (h_m - h_0, \beta_m - \beta_0)$. The following notation is used in the sequel:

$$\begin{aligned} \phi_0\left(\mathbf{t}, \phi_1, \phi_2, \phi_3\right) &\equiv \alpha_0 + t_1\phi_1 + t_2\phi_2 + t_3\phi_3, \\ \phi_0\left(\mathbf{t}, \phi_1, \phi_2, \phi_3, Z_t\right) &\equiv \begin{array}{c} h_0 + t_1\left(h_1 - h_0\right) + t_2\left(h_2 - h_0\right) + t_3\left(h_3 - h_0\right) + \\ \left[\beta_0 + t_1\left(\beta_1 - \beta_0\right) + t_2\left(\beta_2 - \beta_0\right) + t_3\left(\beta_3 - \beta_0\right)\right]' Z_t \end{array}, \end{aligned}$$

and

$$\begin{aligned}\phi(Z_t) &= (\alpha-\alpha_0)(Z_t) \equiv h - h_0 + (\beta-\beta_0)' Z_t, \\ \alpha(Z_t) &\equiv h + \beta' Z_t.\end{aligned}$$

In situations where no confusion can arise, the terms (ϕ_1, ϕ_2, ϕ_3) are suppressed in the notation given above. The log-likelihood with parameters $\phi_0(\mathbf{t},\phi_1, \phi_2, \phi_3)$ can be written:

$$l_{\phi_0(\mathbf{t})}\left(Z_t, \tau^j\right) \equiv \int_0^T \left\{\phi_0(\mathbf{t}, Z_u)\right\} dN_u^j + \int_0^T \left[1 - \exp\left\{\phi_0(\mathbf{t},Z_u)\right\} \mathbf{1}_{\left\{\tau^{(1)} \geq u\right\}}\right] du.$$

The log-likelihood with parameters α will be generically written as $l_\alpha(Z_t, \tau^j)$. A partial derivative w.r.t. t_1 gives us the pathwise derivative of the log-likelihood in the direction of ϕ_1:

$$\begin{aligned}l'_{\phi_0(\mathbf{s})}[\phi_1] &\equiv \frac{\partial}{\partial t_1} l_{\phi_0(\mathbf{t})}\Big|_{\mathbf{t}=\mathbf{s}}, \\ &= \int_0^T \left\{\phi_1(Z_u)\right\} dN_u^j - \int_0^T \exp\left[\phi_0(\mathbf{s}, Z_u)\right] \cdot \phi_1(Z_u) \mathbf{1}_{\left\{\tau^{(1)} \geq u\right\}} du.\end{aligned}$$

Another derivative can be taken in the direction of ϕ_2:

$$\begin{aligned}l''_{\phi_0(\mathbf{s})}[\phi_1, \phi_2] &\equiv \frac{\partial^2}{\partial t_2 \partial t_1} l_{\phi_0(\mathbf{t})}\Big|_{\mathbf{t}=\mathbf{s}}, \\ &= -\int_0^T \exp\left\{\phi_0(\mathbf{s}, Z_u)\right\} \cdot \left\{\phi_2(Z_u)\right\} \left\{\phi_1(Z_u)\right\} \mathbf{1}_{\left\{\tau^{(1)} \geq u\right\}} du.\end{aligned}$$

If $\phi_1 = \phi_2$ this will be written as $l''_{\phi_0(\mathbf{s})}[\phi_1]$. Setting $\mathbf{s} = \mathbf{0}$ in $l'_{\phi_0(\mathbf{s})}[\phi_1]$, define:

$$\Delta(Z_t, \tau, \alpha_0)[\alpha_1 - \alpha_0] \equiv l'_{\alpha_0}[\phi_1] = \int_0^T \left\{\phi_1(Z_u)\right\} dM_u^j.$$

$\Delta(Z_t, \tau, \alpha_0)[\alpha_1 - \alpha_0]$ is the semi-nonparametric score function for our point process log-likelihood. The derivative is taken pathwise in the direction of ϕ_1. For any choice of $\alpha_1 \in \Theta$, the process $\int_0^t \left\{\phi_1(Z_u)\right\} dM_u^j$ is a continuous-time martingale in t under our assumptions and an appropriate filtration (see the appendix). This martingale structure is needed for arguments in the sequel.

The Fisher norm is based on the following inner product on the recentered space $\Theta - \alpha_0$:[5]

$$\begin{aligned} \langle \alpha_1 - \alpha_0, \alpha_2 - \alpha_0 \rangle &\equiv E\left\{\Delta\left(Z_t, \tau, \alpha_0\right)\left[\alpha_1 - \alpha_0\right] \times \Delta\left(Z_t, \tau, \alpha_0\right)\left[\alpha_2 - \alpha_0\right]\right\}, \\ &= E\left\{\int_0^T \left\{\phi_1\left(Z_u\right)\right\} dM_u^j \times \int_0^T \left\{\phi_2\left(Z_u\right)\right\} dM_u^j\right\}. \end{aligned} \tag{3}$$

Define also:

$$\|\alpha_1 - \alpha_0\| \equiv \sqrt{\langle \alpha_1 - \alpha_0, \alpha_1 - \alpha_0 \rangle}.$$

The inner product $\langle \alpha_1 - \alpha_0, \alpha_1 - \alpha_0 \rangle$ is the information matrix for our likelihood for the pathwise derivative in the direction $\alpha_1 - \alpha_0$. We use $\|\widehat{\alpha} - \alpha_0\|$ to describe the rate of convergence of our estimator $\widehat{\alpha}$. The inner product structure of $\|\widehat{\alpha} - \alpha_0\|$ allows us to consider the parameter space $clsp\left(\Theta\right) - \alpha_0$ a Hilbert space and derive asymptotic normality. $clsp\left(\Theta\right)$ is the closed linear span of Θ under $\|\cdot\|$. This Hilbert space structure will be verified below.

The martingale structure on $\Delta\left(Z_u, \tau, \alpha_0\right)\left[\alpha_1 - \alpha_0\right]$ allows us to characterize (3) in a more convenient way. Using Fleming and Harrington (1991) Theorem 2.4.4, it follows that

$$E\left\{\int_0^T \left\{\phi_1\left(Z_u\right)\right\} dM_u^j \times \int_0^T \left\{\phi_2\left(Z_u\right)\right\} dM_u^j\right\} = E\left\{\int_0^T \left\{\phi_1\left(Z_u\right)\right\}\left\{\phi_2\left(Z_u\right)\right\} \lambda_0\left(u\right) du\right\}.$$

In this expression, $\lambda_0\left(t\right) = \left\{\exp\left(h_0\left(t\right) + \beta_0' Z_t\right)\right\} \mathbf{1}_{\left\{\tau^{(1)} \geq t\right\}}$ and this notation is used throughout the sequel. This equality simplifies the inner product and helps in verifying various technical assumptions. Notice that this is an information equality as in the standard maximum likelihood case.

The Fisher norm defined above is only a measure of distance for parameters corresponding to one type of event τ^j. This is all that is needed if $R = 1$. In the competing risks case, we must define a measure of distance for R types of event. In order to derive asymptotic normality for functionals such as Examples 6 and 7 given above, this metric must also have a Hilbert space structure. This can be achieved by using the direct sum of Hilbert spaces for each event type.[6] Specifically, consider the product of parameter spaces for all event types $\Theta^1 - \alpha_0^1 \times \cdots \times \Theta^R - \alpha_0^R$. Elements of this space are written as $\alpha^{1:R}$. An inner product for this parameter space can be

defined as:

$$\left\langle \alpha_1^{1:R}, \alpha_2^{1:R} \right\rangle_{1:R} \equiv \left\langle \alpha_1^1 - \alpha_0^1, \alpha_2^1 - \alpha_0^1 \right\rangle_1 + \cdots + \left\langle \alpha_1^R - \alpha_0^R, \alpha_2^R - \alpha_0^R \right\rangle_R .$$

This just adds up the individual inner products from each event type. It is easily seen to inherit the inner product structure from its components. We can now measure distance between $\alpha^{1:R}$ and $\alpha_0^{1:R}$ using:

$$\left\| \alpha^{1:R} - \alpha_0^{1:R} \right\|_{1:R} = \sqrt{\langle \alpha_1^{1:R}, \alpha_2^{1:R} \rangle_{1:R}}.$$

The closed linear span of $\Theta^1 \times \cdots \times \Theta^R$ is written as $clsp\left(\Theta^1 \times \cdots \times \Theta^R\right)$ and the corresponding recentered space as $clsp\left(\Theta^1 \times \cdots \times \Theta^R\right) - \alpha_0^{1:R}$. This setup has a Hilbert space structure analogously to the description above.

4.2 Rate of Convergence

Deriving rates of convergence for estimators in the previous section requires smoothness properties on the Fisher norm described above. Achieving the required $o_p\left(n^{-1/4}\right)$ rate also required restricting the complexity of the function space $\mathcal{H}$. Both these conditions require smoothness assumptions on the underlying specification. Hölder balls are chosen to characterize the required smoothness, although other choices are possible.

Definition 9 (Hölder Ball) *For any $\varpi \in \mathbb{N}$, $0 < \gamma \leq 1$ and $0 < K < \infty$, the set of functions $f : [0,1]^d \to \mathbb{R}$ such that f has all partial derivatives of order $\leq \varpi + \gamma$ and*

$$\left(\max_{|p| \leq \varpi} \sup_{x \in [0,1]^d} \left| D^p f(x) \right| + \max_{|p| = \varpi} \sup_{x \neq y \in [0,1]^d} \frac{\left| D^p f(x) - D^p f(y) \right|}{|x - y|^\gamma} \right) < K,$$

where $p \in \mathbb{N}^d$ and $|p| = p_1 + \cdots + p_d$, is called a Hölder ball with smoothness $\varpi + \gamma$ and radius K. We write this as $\mathcal{L}_{\varpi,\gamma,K}\left([0,1]^d\right)$.

The following smoothness assumptions are made in the sequel.

(B1) $\mathcal{H} \subset \mathcal{L}_{\varpi,1,K_1}([0,1])$ *for some* $\varpi \geq 1$ and $0 < K_1 < \infty$.

(B2) *(i) Z_t follows assumption (A2) and its density $f : \overline{\mathcal{Z}} \to \mathbb{R}$ is bounded and Lipschitz. (ii)* $0 < C_f = \inf_{z \in \overline{\mathcal{Z}}} f(z)$.

Condition (B1) puts restrictions on the smoothness of the baseline hazards $h_0^j(t)$. This uniformly bounds the first derivative of functions in $\mathcal{H}$. As discussed in Section 3, this removes the necessity of assumption (A5). Condition (B1) also implies that L^1 convergence derived in Theorem 8 is strengthened to sup norm convergence. It then follows that our estimators converge in the Fisher norm. Condition (B2) requires the covariates have a density f and restricts its smoothness. f is also assumed to have a strictly positive lower bound.

In order to state convergence rates, a few more definitions and restrictions are needed. The rate at which convergence happens depends on the lever of smoothness ϖ. Additionally, the convergence rate depends on how fast the sieve spaces $\mathcal{H}_n$ can approximate the true parameters h_0^j as $n \to \infty$. The following assumptions describe restrictions on these two aspects of the setup which are needed for our rate result. We first define $\widetilde{\alpha}_0^j = \left(\widetilde{h}_n^j, \beta_0^j\right)$ where $\widetilde{h}_n^j$ is an element in $\mathcal{H}_n$ closest to h_0^j in L^1.[7]

(B3) *(i)* $1 > (d + R)/\varpi$. *(ii) k and δ are constants such that* $k \leq \delta$; $0 < k \leq 1/2$;

$$\delta \geq \left(\frac{\varpi}{\varpi - d - R}\right) k,$$

and

$$\frac{(\varpi - d - R)}{2(\varpi + d + R)} \geq k.$$

(*B4*) for each event type j: (i)

$$\int_0^T \left|\left(h_0^j - \widetilde{h}_n^j\right)(u)\right| du = O(g_n). \tag{4}$$

(ii) *For some* $\varkappa_1, \varkappa_2 > 0$:

$$\sup_{t \in [0,T]} \left|\left(h_0^j - \widetilde{h}_n^j\right)(t)\right| = O\left(n^{-\varkappa_1 - \varkappa_2}\right).$$

Theorem 10 *Make the assumptions (A1)-(A4), (B1)-(B4). Then clsp* $\left(\Theta^1 \times \cdots \times \Theta^R\right) - \alpha_0^{1:R}$ *is a Hilbert space w.r.t. the norm* $\|\cdot\|_{1:R}$ *and the rate of convergence is:*

$$\left\|\widehat{\alpha}^{1:R} - \alpha_0^{1:R}\right\|_{1:R} = O_p\left(\max\left\{n^{-\delta}, n^{\delta}\varepsilon_n^2, n^{\delta}g_n, n^{\delta-1/2-\varkappa_1/2}, n^{-1/2+\delta-k}\right\}\right). \tag{5}$$

The infeasible case presented in WS characterizes a rate result similar to (5). Terms analogous to $n^{-\delta}$ and $n^{-1/2+\delta-k}$ appear in that rate. Their result requires high-level assumptions on both the smoothness of the Fisher norm and the metric entropy of the score functions. In our point process likelihood case, we characterize smoothness of the Fisher norm and metric entropy bounds on the score functions under our assumptions. The rates $n^{-\delta}$ and $n^{-1/2+\delta-k}$ in (5) are then proved using these results. The infeasible case also contains a rate analogous to $n^{\delta}\varepsilon_n^2$. $n^{\delta}\varepsilon_n^2$ depends on the error in optimizing the criterion function. With a fast enough rate for ε_n^2 this term can be ignored.

In our feasible sieve case, the additional terms $n^{\delta}g_n$ and $n^{\delta-1/2-\varkappa_1/2}$ appear. These are determined by the chosen sieve spaces $\mathcal{H}_n$. The terms g_n and $\varkappa_1$ characterize how the best sieve approximation to the underlying true functions h_0^j influences the convergence rate. If $\left|h_0^j - \widetilde{h}_n^j\right|$ converges to zero arbitrarily fast in the sup norm, then both of these terms disappear as the remaining terms are slower. This arbitrarily fast sup norm approximation essentially gives the infeasible case. In applications, the infeasible case would preform poorly because of high variance in finite samples. We may also wish to restrict the parameter space as in the sieve case in order to derive asymptotic normality results after an appropriate rate of convergence is determined. This is the approach taken below.

With enough smoothness (i.e. ϖ large enough) (5) is $o_p\left(n^{-1/4}\right)$. Achieving this rate requires an appropriately chosen sequence of sieve spaces $\mathcal{H}_n$. Specifically, the form of (5) requires $n^{-\delta} = n^{-1/4-a}$ for some small $a > 0$. Under this restriction on δ, for $n^{\delta}g_n$ to be faster than $n^{-1/4}$ we need,

$$\int_0^T \left|\left(h_0^j - \widetilde{h}_n^j\right)(u)\right| du = o\left(n^{-1/2}\right).$$

The sieve spaces must approximate the true functions h_0^j at an $o\left(n^{-1/2}\right)$ rate in L^1. $\mathcal{H}_n$ must meet this requirement for asymptotic normality to hold. Simulation evidence presented in Section 5 shows this requirement is important for finite sample performance. The L^1 nature of the approximation needed in (B4) is a result of the likelihood form, not any choice we make. The $O\left(n^{-\varkappa_1-\varkappa_2}\right)$ rate required in (B4) is a very weak restriction as both $\varkappa_1, \varkappa_2 > 0$ can be chosen arbitrarily small.

5 Asymptotic Normality of Functionals

In the previous section, we derived $o_p\left(n^{-1/4}\right)$ convergence rates for our estimator in the Fisher norm. This rate result and the norm's Hilbert space structure are required prerequisites for deriving asymptotic normality of plug-in estimators of the functionals presented in Section 2.2. CLS derive conditions under which the above findings result in asymptotic normality for a large class of plug-in estimators. Normality depends on the specific functional, sieve choice and other considerations. We outline their estimation procedure and provide simulation evidence of the method's performance.

CLS not that, in finite samples, sieve MLE is equivalent to maximizing a parametric maximum likelihood where the number of parameters is controlled by the amount of data. We can write the likelihood function as $l_\gamma\left(Z_t^i, \tau_i^j\right)$ where γ represents the parameters required when using the sieve space Θ_n in estimation. The number of parameters in γ corresponds to the amount of data n, which is suppressed in the notation. This perspective suggests that we estimate the covariance of the parameters γ as we would in the parametric case. A standard sandwich covariance matrix is used:

$$\begin{aligned}
\widehat{R}_n &= -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 l_{\widehat{\gamma}}\left(Z_t^i, \tau_i\right)}{\partial\gamma\partial\gamma'}, \\
\widehat{B}_n &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial l_{\widehat{\gamma}}\left(Z_t^i, \tau_i\right)}{\partial\gamma}\frac{\partial l_{\widehat{\gamma}}\left(Z_t^i, \tau_i\right)}{\partial\gamma'}\right], \\
\widehat{\Sigma}_n &= \widehat{R}_n^{-1}\widehat{B}_n\widehat{R}_n^{-1}
\end{aligned}$$

Similarly, functionals can be written as functions of the parameters γ: $f(\gamma)$. The plug-in estimator of $f(\gamma)$ is formed by plugging in the estimator of γ appropriately. This plug-in estimator $f(\widehat{\gamma})$ is exactly the plug-in functional estimator $f(\widehat{\alpha})$. Motivated by the delta method, CLS estimate the standard deviation of $f(\widehat{\gamma})$ using:

$$\begin{aligned}\widehat{F}_n &= \frac{\partial f(\widehat{\gamma})}{\partial \gamma}, \\ \widehat{\sigma} &= \left[\widehat{F}_n' \widehat{R}_n^{-1} \widehat{B}_n \widehat{R}_n^{-1} \widehat{F}_n\right]^{1/2}.\end{aligned}$$

This estimator is used in our sieve asymptotics. The difference between this and standard MLE or QMLE is that the number of parameters γ is growing with n. CLS derive condition under which:

$$\sqrt{n}\frac{f(\widehat{\alpha}) - f(\alpha_0)}{\widehat{\sigma}} \Rightarrow N(0,1).$$

This allows us to conduct inference on the functionals in Section 2.2. The above results are stated for the $R = 1$ case. This is easily extended to the competing risks case by adding the criterion functions for each j into a single criterion function.

5.1 Simulations

In this section, the performance of the inference procedure outlined above is examined with a simulation study. Consider a single event type ($R = 1$) with hazard function having the Cox form. We estimate the following functional of the underlying hazard rate:

$$\begin{aligned}& P\{\tau_i > T | \mathcal{F}_0\}, \\ = \; & E\left[\exp\left(-\int_0^T \exp\left(h_0(u) + \beta_{01} Z^{i1}(u) + \beta_{02} Z^{i2}(u)\right) du\right) \middle| Z^i(0) = z\right], \end{aligned} \tag{6}$$

which is the conditional probability of survival given $Z^i(0)$. The processes and defaults are observed over the interval $[0, T]$ scaled to be $[0, 5]$. The covariates are assumed to follow a

$\gamma = 1.0$, $\beta_{02} = 0.2$	$\mathbb{P}\{\tau_i > T \mid \mathcal{F}_0\} = 0.28215$								
	Mean Estimate			Cov Prob			Mean CI Length		
	BStrap	$j=0$	$j=1$	BStrap	$j=0$	$j=1$	BStrap	$j=0$	$j=1$
$n = 400$	0.2570	0.2587	0.2555	0.7807	0.565	0.578	0.0964	0.0675	0.0664
$n = 800$		0.2569	0.2565		0.438	0.442		0.0474	0.0470
$n = 1,200$		0.2571	0.2566		0.336	0.282		0.0387	0.0383
$\gamma = 0.75$, $\beta_{02} = 0.2$	$\mathbb{P}\{\tau_i > T \mid \mathcal{F}_0\} = 0.34279$								
	Mean Estimate			Cov Prob			Mean CI Length		
	BStrap	$j=0$	$j=1$	BStrap	$j=0$	$j=1$	BStrap	$j=0$	$j=1$
$n = 400$	0.3193	0.3198	0.3181	0.8097	0.564	0.666	0.1017	0.0705	0.0699
$n = 800$		0.3191	0.3186		0.476	0.522		0.0498	0.0491
$n = 1,200$		0.3189	0.3196		0.377	0.420		0.0406	0.0402
$\gamma = 0.50$, $\beta_{02} = 0.2$	$\mathbb{P}\{\tau_i > T \mid \mathcal{F}_0\} = 0.41580$								
	Mean Estimate			Cov Prob			Mean CI Length		
	BStrap	$j=0$	$j=1$	BStrap	$j=0$	$j=1$	BStrap	$j=0$	$j=1$
$n = 400$	0.3952	0.3965	0.3925	0.8690	0.597	0.656	0.1043	0.0724	0.0719
$n = 800$		0.3966	0.3965		0.524	0.626		0.0512	0.0507
$n = 1,200$		0.3952	0.3946		0.439	0.500		0.0416	0.0412

Table 1: $exp[h_0(t)] = \gamma \times (1/(12.5)) \times (t - 2.5)^2 + 0.1$

Gaussian VAR(1):

$$\begin{bmatrix} Z_t^{i1} \\ Z_t^{i2} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} Z_{t-1}^{i1} \\ Z_{t-1}^{i2} \end{bmatrix} + \begin{bmatrix} \epsilon_t^{i1} \\ \epsilon_t^{i2} \end{bmatrix},$$

$$\begin{bmatrix} \epsilon_t^{i1} \\ \epsilon_t^{i2} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \right).$$

$\sigma_1 = \sigma_2 = 0.5$, $A_{11} = A_{22} = 0.8$ and $A_{12} = A_{21} = 0$ are used throughout. These covariates are assumed to update their value every $1/12$ units of time. Therefore, the covariates change their value 60 times in the interval $[0, 5]$. The functional is considered at a fixed value $z = (-2, -2)$. Note that this is an unusual value considering that the VAR process is mean zero. In the simulations, the observed covariates are started at their unconditional distribution.

The function $h_0(t)$ is estimated using Cardinal B-Spline basis functions. See Chui (1992) for a comprehensive account of these objects. $\mathcal{H}_n$ is based on the following function:

$$B_3(x) = \frac{1}{2} \sum_{j=1}^{3} (-1)^j \binom{3}{j} [\max(0, x - j)]^2.$$

$\gamma = 1.0$, $\beta_{02} = 0.1$	$\mathbb{P}\{\tau_i > T \mid \mathcal{F}_0\} = 0.27510$								
	Mean Estimate			Cov Prob			Mean CI Length		
	BStrap	$j=0$	$j=1$	Bstrap	$j=0$	$j=1$	Bstrap	$j=0$	$j=1$
$n = 400$	0.2493	0.2508	0.2512	0.7695	0.579	0.646	0.0960	0.0688	0.0680
$n = 800$		0.2499	0.2516		0.453	0.502		0.0485	0.0481
$n = 1,200$		0.2514	0.2517		0.417	0.396		0.0396	0.0393
$\gamma = 0.75$, $\beta_{02} = 0.1$	$\mathbb{P}\{\tau_i > T \mid \mathcal{F}_0\} =$0.33577								
	Mean Estimate			Cov Prob			Mean CI Length		
	BStrap	$j=0$	$j=1$	Bstrap	$j=0$	$j=1$	Bstrap	$j=0$	$j=1$
$n = 400$	0.3141	0.3132	0.3148	0.8620	0.633	0.692	0.1020	0.0727	0.0719
$n = 800$		0.3127	0.3147		0.518	0.546		0.0513	0.0508
$n = 1,200$		0.3144	0.3136		0.487	0.478		0.0419	0.0414
$\gamma = 0.50$, $\beta_{02} = 0.1$	$\mathbb{P}\{\tau_i > T \mid \mathcal{F}_0\} =$0.40906								
	Mean Estimate			Cov Prob			Mean CI Length		
	Bstrap	$j=0$	$j=1$	Bstrap	$j=0$	$j=1$	Bstrap	$j=0$	$j=1$
$n = 400$	0.3907	0.3892	0.3913	0.8628	0.665	0.728	0.1049	0.0750	0.0739
$n = 800$		0.3917	0.3912		0.578	0.668		0.0529	0.0524
$n = 1,200$		0.3905	0.3914		0.513	0.580		0.0430	0.0426

Table 2: $exp[h_0(t)] = \gamma \times (1/(12.5)) \times (t - 2.5)^2 + 0.1$

$\gamma = 1.0$, $\beta_{02} = 0.0$	$\mathbb{P}\{\tau_i > T \mid \mathcal{F}_0\} = 0.26468$								
	Mean Estimate			Cov Prob			Mean CI Length		
	BStrap	$j=0$	$j=1$	BStrap	$j=0$	$j=1$	BStrap	$j=0$	$j=1$
$n = 400$	0.2422	0.2400	0.2411	0.7988	0.615	0.684	0.0956	0.0710	0.0698
$n = 800$		0.2411	0.2426		0.528	0.574		0.0501	0.0494
$n = 1,200$		0.2415	0.2427		0.431	0.438		0.0408	0.0402
$\gamma = 0.75$, $\beta_{02} = 0.0$	$\mathbb{P}\{\tau_i > T \mid \mathcal{F}_0\} = 0.32546$								
	Mean Estimate			Cov Prob			Mean CI Length		
	BStrap	$j=0$	$j=1$	BStrap	$j=0$	$j=1$	BStrap	$j=0$	$j=1$
$n = 400$	0.3048	0.3053	0.3035	0.8397	0.695	0.722	0.1019	0.0755	0.0743
$n = 800$		0.3052	0.3043		0.623	0.612		0.0532	0.0525
$n = 1,200$		0.3053	0.3052		0.524	0.511		0.0435	0.0427
$\gamma = 0.50$, $\beta_{02} = 0.0$	$\mathbb{P}\{\tau_i > T \mid \mathcal{F}_0\} = 0.39988$								
	Mean Estimate			Cov Prob			Mean CI Length		
	BStrap	$j=0$	$j=1$	BStrap	$j=0$	$j=1$	BStrap	$j=0$	$j=1$
$n = 400$	0.3810	0.3828	0.3814	0.8755	0.731	0.752	0.1057	0.0780	0.0771
$n = 800$		0.3831	0.3834		0.649	0.698		0.0551	0.0542
$n = 1,200$		0.3832	0.3840		0.591	0.640		0.0450	0.0442

Table 3: $exp[h_0(t)] = \gamma \times (1/(12.5)) \times (t - 2.5)^2 + 0.1$

$B_3(x)$ has support $[0,3]$ and is hill shaped. $B_3(x)$ forms a sequence of increasing function spaces V^j indexed by $j \in \mathbb{N}$. V^j are defined as the linear span of the basis functions:

$$\left\{ 2^{j/2} B_3\left(2^j x - k\right) \middle| \, k \in \mathbb{Z} \right\}. \tag{7}$$

As j increases, $V^j \subset V^{j+1}$ and therefore approximations based on V^j become more accurate. See Chui (1992) chapter 4 for more details.

The basis functions defining V^j are used to determine $\mathcal{H}_n$. As n grows, j is increased as well. Only a finite number of functions from (7) are required because $h_0(t)$ has support $[0,5]$. Two sets of basis functions are used in the simulations. These are $B_3(x-k)$ with $k = -2,-1,0,1,2,3,4$ and $2^{1/2} B_3(2x-k)$ with $k = -2,-1,0,1,2,3,4,5,6,7,8,9$ corresponding to $j=0$ and $j=1$. Theoretically, as the amount of data increases, the number of basis functions should increase. Our asymptotic theory does not give much practical advice for choosing j given n. These two cases are considered to compare their relative performance using different amounts of data. The expectation is that $j=1$ will preform better when n is larger.

The simulations produced consider $n = 400$, 800 and 1200. The value for β_{01} was fixed at 0.1 and β_{02} was varied as $\beta_{02} = 0.2, 0.1$ and 0.0. The following $h_0(t)$ were used:

$$\exp\left[h_0(t)\right] = \gamma \times \frac{1}{12.5} \times (t - 2.5)^2 + 0.1,$$

where simulations were conducted for $\gamma = 1, 0.75$ and 0.5. Notice that these $h_0(t)$ are not in V^j for $j = 0, 1$. This is in keeping with the idea that sieve estimation should holds for a large class of function, including those that are difficult to approximate in finite samples using a given sieve space. As γ increases, $h_0(t)$ becomes more spiked at the edges, making it more difficult for $\mathcal{H}_n$ to approximate. We expect performance to decrease as γ increases.

1000 simulations were conducted at each of the model specifications described above. More extensive simulations studies were conducted and produced similar results. Tables 1-3 report the average estimate of the functional, the coverage probability of 95% confidence intervals and the average length of confidence intervals from 1000 simulations for each specification. The true

conditional probability functional (6) computed from 100,000 simulations is also reported.

There are several aspects of the simulations worth noting. First, regardless of the specification, the estimators have a relatively small bias. This increases as γ increases, but is never very large. However, the coverage probability of confidence intervals is significantly distorted in finite samples. As expected, performance decreases as γ increases in almost all specifications. Somewhat surprisingly, as n increases, the estimator's coverage probability decreases in all cases. For specifications with $n = 400$, the $j = 1$ estimator outperforms $j = 0$ in all cases. For $n = 800$, the $j = 1$ estimator still wins most of the time. When n increases to $1,200$, the $j = 1$ estimator is outperformed in about half of the cases. In particular, it seems to have trouble when $\gamma = 1$, exactly where the increased flexibility is expected to be most useful.

How should we interpret these simulation results in light of the theory derived in this paper? They can partially be explained by a parametric MLE interpretation of the estimators. Sieve estimation can be thought of as misspecified parametric MLE where the misspecification is controlled and disappears in the limit. Because the functions $h_0(t)$ are not contained in $\mathcal{H}_n$, each of the simulations above has a QMLE interpretation. This explains some of the distortion. More generally, there is little reason to expect an arbitrarily chosen parametrization will contain the true underlying function in a given application. As a result, there is little reason to expect an arbitrarily chosen parametric specification will preform better than the results presented here.

Another way of understanding these simulations is with the rate result derived in Theorem 10. As noted above, these results require:

$$\int_0^T \left| \left(h_0 - \widetilde{h}_n \right)(u) \right| du = o\left(n^{-1/2} \right).$$

This means the approximation from the sieve spaces must quickly becomes very accurate as n increases. For the $n = 400$ cases, we can clearly see the $j = 1$ estimator has superior coverage. When n is increase to 800 or $1,200$, the $j = 1$ estimator deteriorates. It is likely that the basis functions from $j = 1$ are not approximating $h_0(t)$ fast enough when $n = 800$ or $1,200$. A better estimator would consider larger j in these cases. The simulations presented took several weeks on a 60 core computing cluster. Increasing j would require optimization over a much larger parameter

space. This simulation is not conducted to conserve on scarce computing resources.

In addition to the simulations conducted above, a bootstrap estimator was considered. These results are presented in Table 1-3 under the heading "BStrap". Using our initial estimation procedure, the $\widehat{J}_2$ confidence interval from Hall (1992) chapter 2 is computed. Critical values are estimated by resampling the data 100 times. $1,000$ simulations were conducted for all $n = 400$ and $j = 0$ specifications using this procedure. The same summary statistics as before are reported. While the resulting confidence intervals do not have 95% coverage, they substantially outperform both the $j = 1$ and $j = 0$ estimators. This suggests that bootstrapping can significantly alleviate finite sample distortions in our asymptotic theory. The bootstrap simulations presented took approximately two weeks on a 60 core computing cluster. This substantial computational burden is why additional bootstrap simulations were not considered.

6 Conclusion

In this paper, asymptotic results are derived for a semi-nonparametric sieve likelihood estimator of the Cox proportional hazard model in a competing risks framework. Consistency of the estimator and its rate of convergence in the Fisher norm are derived. These initial results are prerequisites for deriving asymptotic normality of plug-in estimators of functionals. This program can be used to conduct inference on a large number of functionals including formulas for conditional probabilities and asset pricing. Simulations are conducted showing the performance of the inference procedure. Coverage probability has finite sample distortion which depends on the choice of sieve basis functions. Bootstrapping is shown to significantly increase finite sample performance.

There are several directions that future research in this area could take. One possibility is to derive consistent estimators of the covariates' distribution. The exact form of this estimator is not clear as survivorship bias due to censoring will distort standard estimation. Once this theory is developed, these estimators can be combined with hazard estimation to characterize the asymptotic distribution of functionals when the covariate structure is unknown.

A second possibility is to analyze bootstrapping or other higher order asymptotic refinements. It is clear from the simulations above that inference is distorted in finite samples. Bootstrapping

estimators greatly increases the coverage probability of confidence intervals. A closer analysis of this would be helpful. A related problem is how to choose the number of basis functions in a data driven way. The theory and simulations in this paper make clear that the complexity of the sieve spaces needs to grow quickly in order for a satisfactory normal approximation to hold. Exact choices for the number of sieve basis functions derived from data would be useful in practice.

A third possibility is to consider dependence structures between observations. This would involve modifying proofs which dependent on *i.i.d.* empirical process theory. This could require considering more complex sampling structures such as panel data in order to include macro variables.

The results presented in this paper are derived for the Cox proportional hazard model. A fourth topic for future research is how to extend these results to more general cases. One such extension is to fully nonparametric multiplicative hazard functions as in Linton et al. (2003). In kernel estimation, the pointwise rate of convergence in the multidimensional case is the same as the one-dimensional case when a multiplicative structure is imposed. How this more general model affects the rate of convergence in the sieve case is of interest. If similar $o_p\left(n^{-1/4}\right)$ convergence rates can be achieved, then the same inference procedure for hazard functionals is possible.

Notes

[1]Note that the form $\exp\left(h\left(t\right)+\beta' Z^i\left(t\right)\right)$ is equivalent to the traditional Cox proportional hazard form $h\left(t\right)\exp\left(\beta' Z^i\left(t\right)\right)$. The form used here provides simplifications in later sections.

[2]We will arbitrarily define the hazard rate at a fixed constant C on (T,∞). This does nothing to change the observed data. This simplifies statements of certain proofs in the appendix.

[3]The conventions $t_0=0$ and $t_{m+1}=T$ are adopted.

[4]A process is càglàd when its paths are left continuous with right-hand-limits. Càglàd paths imply that the processes $Z^i\left(t\right)$ used below are predictable, an important technical property for our results. This follows from, for example, Protter (2005) pg. 102.

[5]This is not a true inner product yet. We still have to take the linear span of the space Θ and prove that $\langle\alpha-\alpha_0,\alpha-\alpha_0\rangle=0$ only if $\alpha=\alpha_0$.

[6]See Conway (1990) Section I.6 for a definition of the direct sum of Hilbert spaces.

[7]A closest element need not exist because we are considering distance in L^1. We may choose an arbitrary

element within an $\epsilon_n > 0$ of

$$\inf_{h \in \mathcal{H}_n} \|h - h_0\|_{L^1},$$

with $\epsilon_n \to 0$.

A Appendix

Throughout this appendix, many proofs are given for an arbitrary event type j. The stated results then follow by repeating the proofs for each j. In order to reduce the notation and make the proofs more readable, we remove the j index in most of the proofs that follow. It is to be understood that the proofs hold for an arbitrary j. The statements of all Lemmas, Corollaries and Propositions below will contain j indexes for completeness.

All expectations in what follow are with respect to the true underlying parameters in the model α_0. The notation C, C', C'' etc. will be used for an arbitrary constants which can change from line to line. To more compactly present results, the notation $\widehat{\widehat{h}}(s) = \exp\left(\widehat{h}(s)\right)$ and $\widehat{\widehat{h}}_0(s) = \exp(h_0(s))$ is adopted.

A.1 Martingale Properties

We briefly sketch the needed martingale structure of our competing risks model. See Bielecki and Rutkowski (2004) or Fleming and Harrington (1991) for more details. Define the σ-fields $\mathcal{F} = \sigma\left\{Z^i(s) \middle| s \in [0,T]\right\}$ and $\mathcal{H}_t^j = \sigma\left\{\mathbf{1}\left\{\tau_i^j \leq s\right\} \middle| s \in [0,t]\right\}$. Bielecki and Rutkowski (2004) Lemma 9.1.1 can be used to show that,

$$\mathbf{1}_{\left\{\tau_i^j \leq t\right\}} - \int_0^t \exp\left(h_0^j(u) + \beta_0^{j\prime} Z^i(u)\right) \mathbf{1}_{\left\{\tau_i^j \geq u\right\}} du,$$

is a mean-zero martingale with respect to the filtration $\mathcal{F} \vee \mathcal{H}_t^1 \vee \cdots \vee \mathcal{H}_t^R$. $\tau_i^{(1)}$ is a stopping time with respect to this filtration. Therefore, stopping the process above at $\tau_i^{(1)}$ gives $M_t^{ij} = N_t^{ij} - \Lambda_t^{ij}$ and preserved the martingale structure (see Protter (2005) Chapter I, Theorem 18). By Fleming

and Harrington (1991) Theorem 2.4.4, processes with the form:

$$\int_0^t H(u)\, dM_u^{ij},$$

are mean-zero martingales with respect to the filtration give above assuming weak regulatory conditions on $H(u)$. These conditions will always be satisfied under our assumptions.

A.2 Consistency

We prove several preliminary lemmas before proving the result. The expected log-likelihood will be written as:

$$H^j(h,\beta) = E\left\{\log\left[\frac{d\widetilde{P}^j}{dP}(h,\beta)\right]\right\}.$$

Lemma 11 *Under assumptions (A1)-(A3), for each event type j:*

$$H^j\left(h_0^j,\beta_0^j\right) > H^j(h,\beta) \qquad \left(h_0^j,\beta_0^j\right) \neq (h,\beta) \in \Theta.$$

Proof. $\frac{d\widetilde{P}^j}{dP}(h,\beta)$ is defined as:

$$\frac{d\widetilde{P}^j}{dP}(h,\beta) = \begin{cases} \exp\left(\int_0^T \left[1-\exp\left(h(u)+\beta' Z^i(u)\right)\mathbf{1}_{\left\{\tau^{(1)}\geq u\right\}}\right] du\right) & \text{if } \tau^1 > T \\ \exp\left(h(\tau^1)+\beta' Z^i(\tau^1)\right)\mathbf{1}_{\left\{\tau^{(1)}\geq \tau^1\right\}} & \text{if } \tau^1 \leq T \\ \times \exp\left(\int_0^T \left[1-\exp\left(h(u)+\beta' Z^i(u)\right)\right]\mathbf{1}_{\left\{\tau^{(1)}\geq u\right\}} du\right) & \end{cases}.$$

Let P define a probability measure on an underlying space supporting an independent Poisson process with arrival rate $\lambda = 1$, an independent random variable on $\overline{\mathcal{Z}}$ with the distribution of $Z^i(t)$ and $j-1$ independent standard exponential distributions with support $\overline{\mathcal{X}}$. The notation τ^1 is used for the first event in the Poisson process. From this space we can construct random times for the $j-1$ event types other than j. This is done using exactly the same construction presented in Section 2. Brémaud (1981) section VI.2 shows that $\frac{d\widetilde{P}^j}{dP}(h,\beta)$ defines a change of probability measure so that the random time τ^1 has a hazard rate given by $\exp\left(h(t)+\beta' Z^i(t)\right)\mathbf{1}_{\left\{\tau^{(1)}\geq t\right\}}$ over the interval $[0,T]$. This change of measure gives the observed portions of $(\tau, Z(t))$ the same

distribution as in our construction. This can be shown with Brémaud (1981) section VI.2 and a conditioning argument. This change or probability measure defines a change of distribution on $[0,\infty)\times\overline{\mathcal{Z}}\times\overline{\mathcal{X}}$ where $[0,\infty)$ represents the duration until the first event in the Poisson process. Because of assumptions (A2)-(A3), each $\frac{d\widetilde{P}^j}{dP}(h,\beta)$ is a different function of $[0,\infty)\times\overline{\mathcal{Z}}\times\overline{\mathcal{X}}$ for different values of (h,β). Therefore, the distribution on $[0,\infty)\times\overline{\mathcal{Z}}\times\overline{\mathcal{X}}$ which the measure is changed to is different for each value (h,β). By van der Vaart (1998) Lemma 5.35, the expected log-likelihood has a unique maxima at $\left(h_0^j,\beta_0^j\right)$, the true parameters of the model. The result follows. This argument can be repeated for each type j. ■

Lemma 12 *Make assumptions (A1)-(A5). For each j: if $(h_n,\beta_n)\in\Theta_n$ and*

$$H^j\left(h_0^j,\beta_0^j\right)-H^j\left(h_n,\beta_n\right)\to 0,$$

then

$$h_n\to^{L^1}h_0^j,$$

$$\beta_n\to\beta_0^j.$$

Proof. Throughout this proof we write $\widehat{\widehat{h}}_n(s)=\exp\left(h_n(s)\right)$. By a similar manipulation as in Karr's (1987) proof of Theorem 3.3,

$$\begin{aligned}
& H\left(h_0,\beta_0\right)-H\left(h_n,\beta_n\right)\\
= \;& E\left\{\int_0^T\left[\widehat{\widehat{h}}_n(u)\exp\left(\beta_n'Z^i(u)\right)-\widehat{\widehat{h}}_0(u)\exp\left(\beta_0'Z^i(u)\right)\right]\mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}}du\right\}\\
& -E\left\{\int_0^T\log\left[\frac{\widehat{\widehat{h}}_n(u)\exp\left(\beta_n'Z^i(u)\right)}{\widehat{\widehat{h}}_0(u)\exp\left(\beta_0'Z^i(u)\right)}\right]dN_u^i\right\}\\
= \;& E\left\{\begin{array}{c}\int_0^T\left[\frac{\widehat{\widehat{h}}_n(u)\exp\left(\beta_n'Z^i(u)\right)}{\widehat{\widehat{h}}_0(u)\exp\left(\beta_0'Z^i(u)\right)}-1-\log\left(\frac{\widehat{\widehat{h}}_n(u)\exp\left(\beta_n'Z^i(u)\right)}{\widehat{\widehat{h}}_0(u)\exp\left(\beta_0'Z^i(u)\right)}\right)\right]\\ \times\widehat{\widehat{h}}_0(u)\exp\left(\beta_0'Z^i(u)\right)E\left[\mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}}\middle|\,Z^i(t),\ t\in[0,T]\right]du\end{array}\right\}\\
\geq \;& E\left\{\begin{array}{c}\int_0^T\left[\frac{\widehat{\widehat{h}}_n(u)\exp\left(\beta_n'Z^i(u)\right)}{\widehat{\widehat{h}}_0(u)\exp\left(\beta_0'Z^i(u)\right)}-1-\log\left(\frac{\widehat{\widehat{h}}_n(u)\exp\left(\beta_n'Z^i(u)\right)}{\widehat{\widehat{h}}_0(u)\exp\left(\beta_0'Z^i(u)\right)}\right)\right]\\ \times\widehat{\widehat{h}}_0(u)\exp\left(\beta_0'Z^i(u)\right)Cdu\end{array}\right\}
\end{aligned}\tag{8}$$

The second equality follows from the mean-zero martingale properties described in Section A.1. The constant $C > 0$ in the final inequality follows from boundedness of the hazard. By assumption, (8) converges to zero. For a fixed path $z(t)$, if

$$\int_0^T \left[\begin{array}{c} \frac{\widehat{\widehat{h}}_n(u)\exp(\beta_n' z(u))}{\widehat{\widehat{h}}_0(u)\exp(\beta_0' z(u))} - 1 \\ -\log\left(\frac{\widehat{\widehat{h}}_n(u)\exp(\beta_n' z(u))}{\widehat{\widehat{h}}_0(u)\exp(\beta_0' z(u))}\right) \end{array} \right] \widehat{\widehat{h}}_0(u)\exp(\beta_0' z(u)) C du \to 0, \tag{9}$$

then

$$\widehat{\widehat{h}}_n(t)\exp(\beta_n' z(t)) \to^{L^1} \widehat{\widehat{h}}_0(t)\exp(\beta_0' z(t)). \tag{10}$$

This holds because of boundedness and the form of the function $g(x) = x - 1 - \log(x)$. $g(x)$ on the interval $(0, \infty)$ is uniquely minimized at 1 where its value is 0.

Let S be the set of paths $z(t)$ possible give our assumption (A2). Assume (10) does not hold for all paths $z(t)$ in an open ball in the sup norm. This ball is restricted to S. The realization of $Z^i(t)$ is in this ball with positive probability by (A2)(iv). Because of the positive probability of $Z^i(t)$ having a realization in this set, (8) would fail to converge to zero if the assumption is true. This is a contradiction because we assume (8) converges to zero. Therefore (10) cannot fail on an open ball in the sup norm in S. It must hold for $z(t) \in D$, where D is a dense set of paths in S with the sup norm topology.

Assume there exists a path $z_0 \in S$ such that (10) fails. By the arguments given above, for any $\gamma > 0$ there exists a $z' \in S$, $z' \neq z_0$ which satisfied (10) such that $\xi(t) = z'(t) - z_0(t)$ and

$$\sup_{t\in[0,T]} \|\xi(t)\| < \gamma.$$

Now consider:

$$\begin{aligned}
& \int_0^T \left| \widehat{\widehat{h}}_n(u) \exp\left(\beta_n' z_0(u)\right) - \widehat{\widehat{h}}_0(u) \exp\left(\beta_0' z_0(u)\right) \right| du, \\
\leq & \int_0^T \left| \widehat{\widehat{h}}_n(u) \exp\left(\beta_n' z_0(u)\right) - \widehat{\widehat{h}}_n(u) \exp\left(\beta_n' z_0(u)\right) \exp\left(\beta_n' \xi(u)\right) \right| du, \\
& + \int_0^T \left| \widehat{\widehat{h}}_n(u) \exp\left(\beta_n' z_0(u)\right) \exp\left(\beta_n' \xi(u)\right) - \widehat{\widehat{h}}_0(u) \exp\left(\beta_0' z_0(u)\right) \exp\left(\beta_0' \xi(u)\right) \right| du, \\
& + \int_0^T \left| \widehat{\widehat{h}}_0(u) \exp\left(\beta_0' z_0(u)\right) \exp\left(\beta_0' \xi(u)\right) - \widehat{\widehat{h}}_0(u) \exp\left(\beta_0' z_0(u)\right) \right| du, \\
= & \int_0^T \left| \widehat{\widehat{h}}_n(u) \exp\left(\beta_n' z_0(u)\right) \right| \left| 1 - \exp\left(\beta_n' \xi(u)\right) \right| du + o(1), \\
& + \int_0^T \left| \widehat{\widehat{h}}_0(u) \exp\left(\beta_0' z_0(u)\right) \right| \left| \exp\left(\beta_0' \xi(u)\right) - 1 \right| du, \\
\leq & C \sup_{\beta, s} \left| 1 - \exp\left(\beta' \xi(u)\right) \right| + o(1), \\
& + C \sup_{s} \left| \exp\left(\beta_0' \xi(u)\right) - 1 \right|.
\end{aligned}$$

We get the $o(1)$ term because $z_0(t) + \xi(t)$ satisfies (10). The first and third terms have a constant C because of boundedness. Because we may choose $\xi(t)$ such that (10) holds for any $\gamma > 0$; for any $\eta > 0$ we can choose $\xi(t)$ such that these exists an N such that

$$\int_0^T \left| \widehat{\widehat{h}}_n(u) \exp\left(\beta_n' z_0(u)\right) - \widehat{\widehat{h}}_0(u) \exp\left(\beta_0' z_0(u)\right) \right| du < \eta$$

for all $n \geq N$. Therefore, (10) holds for $z_0(t)$. It follows that (10) holds for all $z \in S$.

From (B2), for any path $z_0 \in S$ and $\gamma > 0$, there exists a corresponding path $z_0(t) + \epsilon(t) \in S$ such that for one covariate r: $\inf |\epsilon_r(t)| > \gamma$. For the remaining covariates $\epsilon(t)$ is zero. Above we have proven the following holds:

$$\int_0^T \left| \widehat{\widehat{h}}_n(u) \exp\left(\beta_n' z_0(u)\right) - \widehat{\widehat{h}}_0(u) \exp\left(\beta_0' z_0(u)\right) \right| du \to 0, \tag{11}$$

and

$$\int_0^T \left| \widehat{\widehat{h}}_n(u) \exp\left(\beta_n' z_0(u)\right) \exp\left(\beta_{rn} \epsilon_r(u)\right) - \widehat{\widehat{h}}_0(u) \exp\left(\beta_0' z_0(u)\right) \exp\left(\beta_{r0} \epsilon_r(u)\right) \right| du \to 0. \tag{12}$$

We now Taylor expand of the term $\exp\left(\beta_{rn}'\epsilon_r\left(u\right)\right)$ around β_{r0} (the value in β_0 corresponding to the perturbation $\epsilon_r\left(t\right)$). This is done in (12):

$$\begin{aligned}
& \int_0^T \left| \begin{array}{c} \widehat{\widehat{h}}_n\left(u\right)\exp\left(\beta_n'z_0\left(u\right)\right)\left[\exp\left(\beta_{r0}\epsilon_r\left(u\right)\right)+\epsilon_r\left(u\right)\exp\left(c^*\left(u\right)\epsilon_r\left(u\right)\right)\left(\beta_{rn}-\beta_{r0}\right)\right] \\ -\widehat{\widehat{h}}_0\left(u\right)\exp\left(\beta_0'z_0\left(u\right)\right)\exp\left(\beta_{r0}\epsilon_r\left(u\right)\right) \end{array} \right| du, \\
= & \int_0^T \left| \begin{array}{c} \left[\widehat{\widehat{h}}_n\left(u\right)\exp\left(\beta_n'z_0\left(u\right)\right)-\widehat{\widehat{h}}_0\left(u\right)\exp\left(\beta_0'z_0\left(u\right)\right)\right]\exp\left(\beta_{r0}\epsilon_r\left(u\right)\right) \\ +\widehat{\widehat{h}}_n\left(u\right)\exp\left(\beta_n'z_0\left(u\right)\right)\epsilon_r\left(u\right)\exp\left(c^*\left(u\right)\epsilon_r\left(u\right)\right)\left(\beta_{rn}-\beta_{r0}\right) \end{array} \right| du, \\
\geq & \int_0^T \left| \begin{array}{c} \left|\left[\widehat{\widehat{h}}_n\left(u\right)\exp\left(\beta_n'z_0\left(u\right)\right)-\widehat{\widehat{h}}_0\left(u\right)\exp\left(\beta_0'z_0\left(u\right)\right)\right]\exp\left(\beta_{r0}\epsilon_r\left(u\right)\right)\right| \\ -\left|\widehat{\widehat{h}}_n\left(u\right)\exp\left(\beta_n'z_0\left(u\right)\right)\epsilon_r\left(u\right)\exp\left(c^*\left(u\right)\epsilon_r\left(u\right)\right)\left(\beta_{rn}-\beta_{r0}\right)\right| \end{array} \right| du, \\
\geq & \left| \begin{array}{c} \int_0^T\left|\left[\widehat{\widehat{h}}_n\left(u\right)\exp\left(\beta_n'z_0\left(u\right)\right)-\widehat{\widehat{h}}_0\left(u\right)\exp\left(\beta_0'z_0\left(u\right)\right)\right]\exp\left(\beta_{r0}\epsilon_r\left(u\right)\right)\right| du \\ -\int_0^T\left|\widehat{\widehat{h}}_n\left(u\right)\exp\left(\beta_n'z_0\left(u\right)\right)\epsilon_r\left(u\right)\exp\left(c^*\left(u\right)\epsilon_r\left(u\right)\right)\left(\beta_{rn}-\beta_{r0}\right)\right| du \end{array} \right|. \qquad (13)
\end{aligned}$$

(11) shows that the first term in (13) converges to zero. The term

$$\left|\widehat{\widehat{h}}_n\left(u\right)\exp\left(\beta_n'z_0\left(u\right)\right)\epsilon_r\left(u\right)\exp\left(c^*\left(u\right)\epsilon_r\left(u\right)\right)\right|,$$

is uniformly bounded away from zero over $u\in[0,T]$. Therefore, $\beta_{rn}\to\beta_{r0}$ must hold because (13) must converge to zero. Because we can define an appropriate $\epsilon_r\left(t\right)$ for each covariate by (A2)(vi), we have $\beta_n\to\beta_0$.

Because (10) holds for all $z \in S$, $\beta_n \to \beta_0$ implies $\widehat{\widehat{h}}_n(t) \to^{L^1} \widehat{\widehat{h}}_0(t)$. We can see this from the following Taylor expansion around β_0:

$$
\begin{aligned}
& \int_0^T \left| \widehat{\widehat{h}}_n(u) \exp\left(\beta_n' z_0(u)\right) - \widehat{\widehat{h}}_0(u) \exp\left(\beta_0' z_0(u)\right) \right| du \\
= & \int_0^T \left| \begin{array}{c} \widehat{\widehat{h}}_n(u) \left[\exp\left(\beta_0' z_0(u)\right) + \sum_{r=1}^d z_{r0}(u) \exp\left(c^{*\prime}(u) z_{r0}(u)\right) \left(\beta_{rn} - \beta_{r0}\right) \right] \\ -\widehat{\widehat{h}}_0(u) \exp\left(\beta_0' z_0(u)\right) \end{array} \right| du \\
\geq & \int_0^T \left| \begin{array}{c} \widehat{\widehat{h}}_n(u) \exp\left(\beta_0' z_0(u)\right) - \widehat{\widehat{h}}_0(u) \exp\left(\beta_0' z_0(u)\right) \\ +\widehat{\widehat{h}}_n(u) \sum_{r=1}^d z_{r0}(u) \exp\left(c^{*\prime}(u) z_0(u)\right) \left(\beta_{rn} - \beta_{r0}\right) \end{array} \right| du \\
\geq & \int_0^T \left| \begin{array}{c} \left| \left[\widehat{\widehat{h}}_n(u) - \widehat{\widehat{h}}_0(u) \right] \exp\left(\beta_0' z_0(u)\right) \right| \\ - \left| \widehat{\widehat{h}}_n(u) \sum_{r=1}^d z_{r0}(u) \exp\left(c^{*\prime}(u) z_{r0}(u)\right) \left(\beta_{rn} - \beta_{r0}\right) \right| \end{array} \right| du \\
\geq & \left| \begin{array}{c} \int_0^T \left| \left[\widehat{\widehat{h}}_n(u) - \widehat{\widehat{h}}_0(u) \right] \exp\left(\beta_0' z_0(u)\right) \right| du \\ - \int_0^T \left| \widehat{\widehat{h}}_n(u) \sum_{r=1}^d z_{r0}(u) \exp\left(c^{*\prime}(u) z_0(u)\right) \left(\beta_{rn} - \beta_{r0}\right) \right| du \end{array} \right|
\end{aligned} \tag{14}
$$

We have proven that (14) converges to zero. If $\beta_n \to \beta_0$ and $\widehat{\widehat{h}}_n(s) \not\to^{L^1} \widehat{\widehat{h}}_0(s)$ we have a contradiction because

$$
\left| \widehat{\widehat{h}}_n(u) z_{r0}(u) \exp\left(c^{*\prime}(u) z_0(u)\right) \right|,
$$

is uniformly bounded above for all r. Therefore, $\widehat{\widehat{h}}_n(s) \to^{L^1} \widehat{\widehat{h}}_0(s)$ and $\beta_n \to \beta_0$. $\widehat{\widehat{h}}_n(s) \to^{L^1} \widehat{\widehat{h}}_0(s)$ implies $\widehat{h}_n(s) \to^{L^1} h_0(s)$. This argument can be repeated for each j. The result follows. ■

See van der Vaart and Wellner (1996) for the definition of subgraphs and VC-index.

Lemma 13 *Define the set of functions indexed by $s \in [0,T]$:*

$$
f_s : [0,\infty)_1 \times \cdots \times [0,\infty)_R \to \{0,1\},
$$

where

$$
f_s(t) = \mathbf{1}\{t_1 \leq s \leq T\} \mathbf{1}\{t_1 \leq t_2 \leq T\} \cdots \mathbf{1}\{t_1 \leq t_R \leq T\}.
$$

This set of functions has subgraphs with VC-index of degree 2.

Proof. No two points in the set $[0,\infty)_1 \times \cdots \times [0,\infty)_R \times [0,1]$ can be shattered by the subgraphs of f_s. Consider two points x^1 and x^2 in the space $[0,\infty)_1 \times \cdots \times [0,\infty)_R \times [0,1]$. The point's value in $[0,1]$ is irrelevant for the argument that follows (the edge cases follow trivially). If any of the values of a point x corresponding to $t_2, \ldots, t_R$ are less than the value corresponding to t_1, that point can never be chosen by the subgraphs of the functions f_s. This is because the function would always be zero at such a point. It follows that no set of points can be shattered if one of the points has this property. We now rule out the possibility of shattering two points where $x_2, \ldots, x_R$ are all greater than or equal to x_1. Assume x^1 and x^2 are such points. Without loss of generality, assume $x_1^1 \leq x_1^2$, where these values correspond to the argument t_1. Now we want to find a function f_s which has x^2 in its subgraph, but not x^1. Any function which has x^2 in its subgraph satisfies $x_1^2 \leq s \leq T$ and therefore $x_1^1 \leq s \leq T$. Because we are considering only points where $x_1^1 \leq x_l^1$ for $l = 2, \ldots, d$, any function that selects x^2 must select x^1 as well. We cannot shatter two points in this space. The result follows. ■

Lemma 14 *Make Assumptions (A1)-(A5). For each j: The following holds almost surely:*

$$K_n \sup_{s \in [0,T]} \left| \frac{1}{n} \sum_{i=1}^{n} N^{ij}(s) - E\left[N^{ij}(s)\right] \right| \to 0. \tag{15}$$

Proof. (15) can be cast as a empirical process problem as in Pollard (1984) chapter II. The notation and terminology from Pollard (1984) will be used throughout this proof. Write the set of functions from Lemma 13 as $\mathcal{F}$. This set is polynomial by Lemma 13. We argue for $\tau^j = \tau^1$ and the rest follow by symmetry. Notice that for $f_s \in \mathcal{F}$:

$$\begin{aligned} N^i(s) &= f_s\left(\tau^1, \ldots, \tau^R\right), \\ &= \mathbf{1}\left\{\tau^1 \leq s \leq T\right\} \mathbf{1}\left\{\tau^1 \leq \tau^2\right\} \cdots \mathbf{1}\left\{\tau^1 \leq \tau^R\right\}. \end{aligned}$$

The function $F = 1$ is an obvious envelope function. For any probability measure Q on $[0,\infty)$, $0 < QF = 1 < \infty$. Therefore, for any Q,

$$N_1(\epsilon, Q, \mathcal{F}) \leq A\epsilon^{-W}.$$

This follows from Pollard (1984) Lemma 25. We now apply the proof of Theorem 37 from Pollard (1984). Set $\delta_n = 1$, which always satisfies the conditions of the theorem. The sequence α_n must satisfy:

$$\frac{n\alpha_n^2}{\log(n)} \to \infty.$$

This holds if:

$$\alpha_n = \frac{1}{n^{1/2-\rho}},$$

and $\rho > 0$. By the argument in the proof of Theorem 37 from Pollard (1984), for $\epsilon_n = \epsilon\alpha_n$ and a large enough n:

$$\begin{aligned} P\left\{\sup_{\mathcal{F}} |P_n f - Pf| > 8\epsilon\alpha_n\right\} \leq\ & 8A\epsilon^W \exp\left[W\log(1/\alpha_n) - n\epsilon^2\alpha_n^2/128\right] \\ & +4A(\epsilon\alpha_n)^{-W}\exp(-n) \end{aligned} \tag{16}$$

From our choice of α_n, we see that the right hand side of (16) converges to zero at an exponential rate. This implies, for any $\epsilon > 0$, the sum of probabilities (16) is a convergent sequence in n. By a Borel-Cantelli argument:

$$n^{1/2-\rho}\sup_{\mathcal{F}} |P_n f - Pf| \to 0,$$

almost surely. This argument can be repeated for each j. ■

Lemma 15 *Make Assumptions (A1)-(A3). For each j: the following holds almost surely:*

$$\sup_{\beta}\left|\frac{1}{n}\sum_{i=1}^{n}\int_0^T \left[\beta' Z^i(u)\right] dN_u^{ij} - E\left(\int_0^T \left[\beta' Z^i(u)\right] dN_u^{ij}\right)\right| \to 0 \tag{17}$$

Proof. The functions above have the Lipschitz property defined in van der Vaart and Wellner (1996) Section 2.7.4 over the parameter space B. Boundedness means we can define envelope functions. van der Vaart and Wellner (1996) Theorem 2.7.11 and 2.4.1 then give the desired results. The argument is standard and the details are omitted. The argument can be repeated for each j. ■

Lemma 16 *Make Assumptions (A1)-(A5). The following holds almost surely:*

$$\sup_{\beta}\left(\int_0^T\left|\frac{1}{n}\sum_{i=1}^n\left\{\exp\left(\beta' Z^i(u)\right)\mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}}-E\left[\exp\left(\beta' Z^i(u)\right)\mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}}\right]\right\}\right|du\right)\to 0$$

Proof. We can bound this term as follows:

$$\begin{aligned}
&\sup_{\beta}\left(\int_0^T\left|\frac{1}{n}\sum_{i=1}^n\left\{\exp\left(\beta' Z^i(u)\right)\mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}}-E\left[\exp\left(\beta' Z^i(u)\right)\mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}}\right]\right\}\right|du\right),\\
\leq\ &\sup_{\beta,u\in[0,T]}\left|\frac{1}{n}\sum_{i=1}^n\left\{\exp\left(\beta' Z^i(u)\right)\mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}}-E\left[\exp\left(\beta' Z^i(u)\right)\mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}}\right]\right\}\right|,\\
\leq\ &\sum_{l=0}^m\sup_{\beta,u\in[t_l,t_{l+1}]}\left|\frac{1}{n}\sum_{i=1}^n\left\{\exp\left(\beta' Z_{t_l}^i\right)\mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}}-E\left[\exp\left(\beta' Z_{t_l}^i\right)\mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}}\right]\right\}\right|.
\end{aligned}$$

Above, the interval $[0,T]$ was divided into subintervals $[t_l,t_{l+1}]$ corresponding to updates of the process $Z^i(t)$. The class of functions $\exp(\beta' z)$ has a finite L^1 bracketing number by van der Vaart and Wellner (1996) Theorem 2.7.11. Take an arbitrary pair of bracket functions f_u and f_l from this L^1 bracketing. Divide the subintervals $[t_l,t_{l+1}]$ into a grid of equally spaced points s_k for some finite number of k. New brackets can be defined as $f_u\mathbf{1}_{\{t\geq s_k-\epsilon\}}$ and $f_l\mathbf{1}_{\{t\geq s_k+\epsilon\}}$ to account for the indicator function. This forms a new finite number of brackets. The L^1 distance between these brackets can be controlled as the bracket functions are bounded and the points s_k can be made arbitrarily close to each other. Therefore, the L^1 bracketing number is finite for all $\epsilon>0$. This implies a Glivenko-Cantelli theorem holds by van der Vaart and Wellner (1996) Theorem 2.4.1. The result follows. ■

Proof (Theorem 8). We show that $H(h_0,\beta_0)-H\left(\widehat{h},\widehat{\beta}\right)\to 0$ almost surely and the result follows from Lemma 12. This holds for arbitrary j. First, define $\widetilde{\alpha}_0=\left(\widetilde{h}_n,\beta_0\right)$ where $\widetilde{h}_n$ is an element in $\mathcal{H}_n$ closest to h_0 in L^1. This is the same definition as described in Section 4.2.

Following Karr (1987) Theorem 3.3:

$$\begin{aligned} H(\alpha_0) - H(\widehat{\alpha}) &= H(\alpha_0) - H(\widetilde{\alpha}_0), \\ &\quad + H(\widetilde{\alpha}_0) - Q_n(\widetilde{\alpha}_0), \\ &\quad + Q_n(\widetilde{\alpha}_0) - Q_n(\widehat{\alpha}), \\ &\quad + Q_n(\widehat{\alpha}) - H(\widehat{\alpha}), \\ &\leq o(1), \\ &\quad + H(\widetilde{\alpha}_0) - Q_n(\widetilde{\alpha}_0), \\ &\quad + o_{a.s.}(1), \\ &\quad + Q_n(\widehat{\alpha}) - H(\widehat{\alpha}). \end{aligned} \tag{18}$$

If the second and fourth terms in (18) converge to zero a.s., then $H(\alpha_0) - H(\widehat{\alpha}) \to 0$ a.s. The first line in (18) is $o(1)$ by assumption (A4) and the form of the expected likelihood. A Taylor expansion argument gives this result. Note that the third line in (18) is $o(1)$ provided the other lines are $o(1)$. This is because $\widehat{\alpha}$ is chosen to maximize $Q_n(\alpha)$ and $0 \leq H(\alpha_0) - H(\widehat{\alpha})$. Consider the fourth term:

$$\begin{aligned} & Q_n(\widehat{\alpha}) - H(\widehat{\alpha}) \\ = \; & E\left[\int_0^T \widehat{\widehat{h}}(u) \exp\left(\widehat{\beta}' Z^i(u)\right) \mathbf{1}_{\left\{\tau_i^{(1)} \geq u\right\}} du\right] - \frac{1}{n}\sum_{i=1}^n \int_0^T \widehat{\widehat{h}}(u) \exp\left(\widehat{\beta}' Z^i(u)\right) \mathbf{1}_{\left\{\tau_i^{(1)} \geq u\right\}} du, \\ & + \frac{1}{n}\sum_{i=1}^n \int_0^T \widehat{h}(u)\, dN_u^i - E\left[\int_0^T \widehat{h}(u)\, dN_u^i\right], \\ & + \frac{1}{n}\sum_{i=1}^n \int_0^T \left(\widehat{\beta}' Z^i(u)\right) dN_u^i - E\left[\int_0^T \left[\left(\widehat{\beta}' Z^i(u)\right)\right] dN_u^i\right]. \end{aligned}$$

Using integration by parts, this becomes

$$\begin{aligned}
= \; & E\left[\int_0^T \widehat{\widehat{h}}(u)\exp\left(\widehat{\beta}'Z^i(u)\right)\mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}}du\right] - \frac{1}{n}\sum_{i=1}^n\int_0^T \widehat{\widehat{h}}(u)\exp\left(\widehat{\beta}'Z^i(u)\right)\mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}}du, \\
& -\frac{1}{n}\sum_{i=1}^n\int_0^T N^i(u)\widehat{h}'(u)\,du + E\left[\int_0^T N^i(u)\widehat{h}'(u)\,du\right], \\
& +\frac{1}{n}\sum_{i=1}^n \widehat{h}(T)N^i(T) - E\left[\widehat{h}(T)N^i(T)\right], \\
& +\frac{1}{n}\sum_{i=1}^n\int_0^T\left(\widehat{\beta}'Z^i(u)\right)dN_u^i - E\left[\int_0^T\left[\left(\widehat{\beta}'Z^i(u)\right)\right]dN_u^i\right].
\end{aligned}$$

The forth line converges to zero almost surely by Lemma 15. We can bound the third line as follows:

$$\begin{aligned}
& \left|\frac{1}{n}\sum_{i=1}^n \widehat{h}(T)N^i(T) - E\left[\widehat{h}(T)N^i(T)\right]\right|, \\
\leq \; & C\left|\frac{1}{n}\sum_{i=1}^n N^i(T) - E\left[N^i(T)\right]\right|.
\end{aligned}$$

This converges to zero almost surely by the strong law of large numbers. The second line can be bounded as follows:

$$\begin{aligned}
& \left|\frac{1}{n}\sum_{i=1}^n\int_0^T N^i(u)\widehat{h}'(u)\,du - E\left[\int_0^T N^i(u)\widehat{h}'(u)\,du\right]\right|, \\
= \; & \left|\int_0^T\left\{\widehat{h}'(u)\left[\frac{1}{n}\sum_{i=1}^n N^i(u) - E\left[N^i(u)\right]\right]\right\}du\right|, \\
\leq \; & K_n\int_0^T\left|\left\{\frac{1}{n}\sum_{i=1}^n N^i(u) - E\left[N^i(u)\right]\right\}du\right|, \\
\leq \; & K_nT\sup_{u\in[0,T]}\left|\frac{1}{n}\sum_{i=1}^n N^i(u) - E\left[N^i(u)\right]\right|.
\end{aligned}$$

This converges to zero almost surely by Lemma 14. Finally, consider the first line:

$$\begin{aligned}
& \left| \frac{1}{n} \sum_{i=1}^{n} \left(\int_0^T \widehat{\widehat{h}}(u) \left\{ E\left[\exp\left(\widehat{\beta}' Z^i(u) \right) \mathbf{1}_{\left\{ \tau_i^{(1)} \geq u \right\}} \right] - \exp\left(\widehat{\beta}' Z^i(u) \right) \mathbf{1}_{\left\{ \tau_i^{(1)} \geq u \right\}} \right\} du \right) \right|, \\
\leq & \left| \int_0^T \left(\widehat{\widehat{h}}(u) \frac{1}{n} \sum_{i=1}^{n} \left\{ E\left[\exp\left(\widehat{\beta}' Z^i(u) \right) \mathbf{1}_{\left\{ \tau_i^{(1)} \geq u \right\}} \right] - \exp\left(\widehat{\beta}' Z^i(u) \right) \mathbf{1}_{\left\{ \tau_i^{(1)} \geq u \right\}} \right\} \right) du \right|, \\
\leq & C_{\max} \int_0^T \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ E\left[\exp\left(\widehat{\beta}' Z^i(u) \right) \mathbf{1}_{\left\{ \tau_i^{(1)} \geq u \right\}} \right] - \exp\left(\widehat{\beta}' Z^i(u) \right) \mathbf{1}_{\left\{ \tau_i^{(1)} \geq u \right\}} \right\} \right| du, \\
\leq & C_{\max} \sup_{\beta} \left(\int_0^T \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ E\left[\exp\left(\beta' Z^i(u) \right) \mathbf{1}_{\left\{ \tau_i^{(1)} \geq u \right\}} \right] - \exp\left(\beta' Z^i(u) \right) \mathbf{1}_{\left\{ \tau_i^{(1)} \geq u \right\}} \right\} \right| du \right).
\end{aligned}$$

This converges to zero almost surely by Lemma 16.

Therefore, by the assumptions of the theorem, $|Q_n(\widehat{\alpha}) - H(\widehat{\alpha})| \to 0$ a.s. Notice that throughout the proof the exact value of $\widehat{\alpha}$ was irrelevant and the results hold for an arbitrary sequence $\alpha_n \in \Theta_n$ under the assumptions on Θ_n. Therefore, $|H(\widetilde{\alpha}_0) - Q_n(\widetilde{\alpha}_0)| \to 0$ a.s. and by (18), $H(\alpha_0) - H(\widehat{\alpha}) \to 0$ a.s. By Lemma 12,

$$\widehat{\beta} \to \beta_0,$$

$$\widehat{h} \to^{L^1} h_0,$$

almost surely. This argument can be repeated for each j. ■

Corollary 17 *Make Assumptions (A1)-(A4). If $\mathcal{H}$ is restricted to an open Hölder ball, then for all j:*

$$\widehat{h}^j \to^{L^\infty} h_0^j,$$

almost surely.

Proof. This follows because derivatives of functions in Hölder balls are uniformly bounded. ■

A.3 Rate of Convergence

In keeping with our previous convention, all definitions that follow will be made without indexing for j. All proofs will continue to be made without the j index. It is understood that the proofs

that follows hold for an arbitrary j. These proofs can be repeated for each $j \in \{1, \ldots, R\}$ and all stated results will follow. This is done to keep already cumbersome notation more manageable.

A.3.1 Preliminary Definitions and Lemmas

Make the following definitions for arbitrary j:

$$\begin{aligned} u_n &\equiv \widehat{\alpha} - \alpha_0, \\ \widetilde{u}_n &\equiv \widehat{\alpha} - \widetilde{\alpha}_0, \\ \widetilde{\phi}_n &= \widetilde{\alpha}_0 + s_n \widetilde{u}_n. \end{aligned}$$

s_n is a sequence of real numbers defined in Lemma 20 below. Define the following derivatives of the criterion function Q_n, again for arbitrary j:

$$\begin{aligned} & Q_n' \left(\alpha + s_1 \left(\alpha_1 - \alpha_0\right) + s_2 \left(\alpha_2 - \alpha_0\right)\right) \left[\alpha_1 - \alpha_0\right], \\ \equiv \; & \frac{1}{n} \sum_{i=1}^{n} l'_{\alpha + s_1(\alpha_1 - \alpha_0) + s_2(\alpha_2 - \alpha_0)} \left[\alpha_1 - \alpha_0\right], \end{aligned}$$

$$\begin{aligned} & Q_n'' \left(\alpha + s_1 \left(\alpha_1 - \alpha_0\right) + s_2 \left(\alpha_2 - \alpha_0\right)\right) \left[\alpha_1 - \alpha_0, \alpha_2 - \alpha_0\right], \\ \equiv \; & \frac{1}{n} \sum_{i=1}^{n} l''_{\alpha + s_1(\alpha_1 - \alpha_0) + s_2(\alpha_2 - \alpha_0)} \left[\alpha_1 - \alpha_0, \alpha_2 - \alpha_0\right]. \end{aligned}$$

In all cases that follow, either s_1 or s_2 is set to zero. The result will be written as, for example:

$$\begin{aligned} & Q_n' \left(\alpha + s \left(\alpha_1 - \alpha_0\right)\right) \left[\alpha_1 - \alpha_0\right], \\ & Q_n'' \left(\alpha + s \left(\alpha_1 - \alpha_0\right)\right) \left[\alpha_1 - \alpha_0, \alpha_2 - \alpha_0\right]. \end{aligned}$$

$Q_n\left(\alpha\right)$ is written for the criterion function with argument α. When $\alpha_1 - \alpha_0 = \alpha_2 - \alpha_0$, the notation,

$$Q_n'' \left(\alpha + s \left(\alpha_1 - \alpha_0\right)\right) \left[\alpha_1 - \alpha_0\right],$$

is adopted.

Lemma 18 *Assume (A1)-(A4), (B1)-(B4). For each j: for all $\alpha, \alpha_1, \alpha_2 \in \Theta$, the following derivatives exist:*

$$\begin{aligned} Q_0'\left(\alpha + s\left(\alpha_1 - \alpha_0\right)\right)\left[\alpha_1 - \alpha_0\right] &\equiv \frac{\partial}{\partial t} E\left[l_{\alpha + t(\alpha_1 - \alpha_0)}\left(\tau_i, Z_u^i\right)\right]\Big|_{t=s}, \\ Q_0''\left(\alpha + s\left(\alpha_1 - \alpha_0\right)\right)\left[\alpha_1 - \alpha_0, \alpha_2 - \alpha_0\right] &\equiv \frac{\partial^2}{\partial t_2 \partial t_1} E\left[l_{\alpha + t_1(\alpha_1 - \alpha_0) + t_2(\alpha_2 - \alpha_0)}\left(\tau_i, Z_u^i\right)\right]\Big|_{t_1 = s, t_2 = 0}, \end{aligned}$$

and we can differentiate under the expectation:

$$\begin{aligned} Q_0'\left(\alpha + s\left(\alpha_1 - \alpha_0\right)\right)\left[\alpha_1 - \alpha_0\right] &= E\left[\frac{\partial}{\partial t} l_{\alpha + t(\alpha_1 - \alpha_0)}\left(\tau_i, Z_u^i\right)\Big|_{t=s}\right], \\ Q_0''\left(\alpha + s\left(\alpha_1 - \alpha_0\right)\right)\left[\alpha_1 - \alpha_0, \alpha_2 - \alpha_0\right] &= E\left[\frac{\partial^2}{\partial t_2 \partial t_1} l_{\alpha + t_1(\alpha_1 - \alpha_0) + t_2(\alpha_2 - \alpha_0)}\left(\tau_i, Z_u^i\right)\Big|_{t_1 = s, t_2 = 0}\right]. \end{aligned}$$

Proof. This follows from Folland (1999) theorem 2.27 and our boundedness assumptions. ■

Lemma 19 *Assume (A1)-(A4), (B1)-(B4). For each j: for all $\alpha \in \Theta$, $Q_0'\left(\alpha_0\right)\left[\alpha - \alpha_0\right] \leq 0$. This implies $Q_0'\left(\alpha_0\right)\left[u_n\right] \leq 0$.*

Proof. For all t_1 small enough, $\alpha + t_1\left(\alpha_1 - \alpha_0\right) \in \Theta$. This holds because we chose an open Hölder ball for the functions. Therefore, $E\left[l_{\alpha_0 + t_1(\alpha_1 - \alpha_0)}\left(\tau_i, Z_u^i\right)\right]$ is maximized at $t_1 = 0$ by Lemma 11. It follows that $Q_0'\left(\alpha_0\right)\left[\alpha - \alpha_0\right] = 0$. ■

Lemma 20 *Assume (A1)-(A4), (B1)-(B4). For each j: for any $\alpha_1, \alpha_2 \in \Theta$ and any n, the following mean value theorem holds: there exists $s \in (1/2, 1)$ such that*

$$Q_n\left(\alpha_1 + \frac{1}{2}\left(\alpha_2 - \alpha_1\right)\right) - Q_n\left(\alpha_2\right) = -\frac{1}{2} Q_n'\left(\alpha_1 + s\left(\alpha_2 - \alpha_1\right)\right)\left[\alpha_2 - \alpha_1\right].$$

Using this expression, the sequence $s_n \in (1/2, 1)$ can be implicitly defined to satisfy:

$$Q_n\left(\widetilde{\alpha}_0 + \frac{1}{2}\widetilde{u}_n\right) - Q_n\left(\widehat{\alpha}\right) = -\frac{1}{2} Q_n'\left(\widetilde{\alpha}_0 + s_n \widetilde{u}_n\right)\left[\widetilde{u}_n\right].$$

Define:

$$\widetilde{\phi}_n = \widetilde{\alpha}_0 + s_n \widetilde{u}_n.$$

Proof. $Q_n(\alpha_1 + s(\alpha_2 - \alpha_1))$ is differentiable on $s \in [1/2, 1]$ for every realization of the random variables. The mean value theorem then gives the results. ■

Lemma 21 *Assume (A1)-(A4), (B1)-(B4). For each j: For any $\alpha, \alpha_0 \in \Theta$ and any $s \in (1/2, 1)$, there exists $t \in [0, s]$ such that the following mean value theorem holds:*

$$Q_0'(\alpha_0 + s(\alpha - \alpha_0))[\alpha - \alpha_0] = Q_0'(\alpha_0)[\alpha - \alpha_0] + sQ_0''(\alpha_0 + t(\alpha - \alpha_0))[\alpha - \alpha_0].$$

It follows that:

$$Q_0'(\alpha_0 + su_n)[u_n] = Q_0'(\alpha_0)[u_n] + sQ_0''(\alpha_0 + tu_n)[u_n].$$

Proof. This follows from the mean value theorem and Folland (1999) theorem 2.27. ■

Lemma 22 *Make assumptions (A1)-(A4), (B1)-(B4). For each j: $\left\{\omega | \widetilde{\alpha}_0 + \frac{1}{2}\widetilde{u}_n \in \Theta_n\right\}$ is a sequence of sets with probability approaching one. Therefore, $Q_n\left(\widetilde{\alpha}_0 + \frac{1}{2}\widetilde{u}_n\right) - Q_n(\widehat{\alpha}) \leq O_p(\varepsilon_n^2)$.*

Proof. $\beta_0 + \frac{1}{2}\left(\widehat{\beta} - \beta_0\right)$ converges almost surely to β_0 by the consistency proof. Therefore, $\beta_0 + \frac{1}{2}\left(\widehat{\beta} - \beta_0\right)$ is inside the parameter set eventually almost surely.

$$\widetilde{h}_n + \frac{1}{2}\left(\widehat{h} - \widetilde{h}_n\right) = \widetilde{h}_n + \frac{1}{2}\left(\widehat{h} - h_0 + h_0 - \widetilde{h}_n\right).$$

Because h_0, $\widehat{h}$ and $\widetilde{h}_n$ are in the same open Hölder ball, $\widehat{h} - h_0 \to 0$ and $h_0 - \widetilde{h}_n \to 0$ almost surely in the sup norm. As a result, $\widetilde{h}_n + \frac{1}{2}\left(\widehat{h} - \widetilde{h}_n\right)$ is inside $\mathcal{H}_n$ eventually almost surely. Therefore, the first statement holds. This implies that with probability approach one, $Q_n\left(\widetilde{\alpha}_0 + \frac{1}{2}\widetilde{u}_n\right) - Q_n(\widehat{\alpha}) \leq O_p(\varepsilon_n^2)$ because $\widehat{\alpha}$ is defined as maximizing $Q_n(\alpha)$ over Θ_n. ■

Lemma 23 *Assume (A1)-(A4), (B1)-(B4). For each j: For any sequence $t_n \in [0, 1]$, $\{\omega | \alpha_0 + t_n u_n \in \Theta\}$ is a sequence of sets with probability approaching one.*

Proof. $\widehat{\beta} - \beta_0 \to 0$ almost surely. This and the fact that the parameters are in an open set implies the result for β. For $h_0 + t_n\left(\widehat{h} - h_0\right)$, $\left(\widehat{h} - h_0\right) \to 0$ in the sup norm almost surely. Because h_0 and $\widehat{h}$ are in the same open Hölder ball, the result follows. ■

Lemma 24 *Make assumptions (A1)-(A4), (B1)-(B4). For each j: there exists $0 < \varrho < 1$ and a uniform constant C^* such that for any $\alpha \in \Theta$, $z_t \in \overline{\mathcal{Z}}$:*

$$\sup_{u\in(0,T],z_t^i\in\overline{\mathcal{Z}}} \left|(h-h_0)(u) + (\beta-\beta_0)' z_u\right| \leq C^* \left\|\alpha_3 - \alpha_0\right\|^{\varrho}. \tag{19}$$

Proof. Because the density f is bounded below by (B2), we have the following:

$$\begin{aligned} & \sup_{u\in(0,T],z_t^i\in\overline{\mathcal{Z}}} \left|(h-h_0)(u) + (\beta-\beta_0)' z_u\right|, \\ \leq\ & \frac{1}{C_f^{1/2}} \sup_{u\in(0,T],z_t^i\in\overline{\mathcal{Z}}} \left|\left\{(h-h_0)(u) + (\beta-\beta_0)' z_u\right\}\left[f\left(z_t^i\right)\right]^{1/2}\right|. \end{aligned}$$

We now bound this upper bound. The proof uses Lemma 2 of Chen and Shen (1998) (hereafter CS).

$$\begin{aligned} & \sup_{u\in(0,T],z_t^i\in\overline{\mathcal{Z}}} \left|\left\{(h-h_0)(u) + (\beta-\beta_0)' z_u\right\}\left[f\left(z_t^i\right)\right]^{1/2}\right|, \\ =\ & \max_l \sup_{u\in(t_l,t_{l+1}],z_t^i\in\overline{\mathcal{Z}}} \left|\left\{(h-h_0)(u) + (\beta-\beta_0)' z_u\right\}\left[f\left(z_t^i\right)\right]^{1/2}\right|, \\ \leq\ & \max_l C\left\{E\left[\int_{t_l}^{t_{l+1}} \left\{(h-h_0)(u) + (\beta-\beta_0)' Z_u\right\}^2 du\right]\right\}^{c_1/2}, \\ \leq\ & C'\left\{E\left[\int_0^T \left\{(h-h_0)(u) + (\beta-\beta_0)' Z_u\right\}^2 du\right]\right\}^{c_1/2}, \\ \leq\ & C^*\left\{E\left[\int_0^T \left\{(h-h_0)(u) + (\beta-\beta_0)' Z_u\right\}^2 \exp\left(\alpha_0\left(Z_u\right)\right) \mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}} du\right]\right\}^{c_1/2}, \\ =\ & C^*\left\|\alpha_3 - \alpha_0\right\|^{c_1}. \end{aligned}$$

Line three follows from CS Lemma 2, where $0 < c_1 < 1$ as described in that lemma. We have assumed that $h, h_0 \in \mathcal{L}_{\varpi,1,K_1}([0,T])$. We also assume (B2) to control the smoothness and boundedness of $[f(z_t^i)]^{1/2}$. (B2) implies that $[f(z_t^i)]^{1/2}$ is inside $\mathcal{L}_{0,1,K_2}\left(\overline{\mathcal{Z}}\right)$ for some $K_2 > 0$. Therefore, over any subinterval $[t_l, t_{l+1}]$, $\left\{(h-h_0)(u) + (\beta-\beta_0)' z_{t_l}\right\}[f^1(z_t^i)]^{1/2}$ is contained in $\mathcal{L}_{0,1,K}\left([t_l, t_{l+1}] \times \overline{\mathcal{Z}}\right)$ for any h, h_0, β, β_0 and some constant $K > 0$. From CS Lemma 2, the constant C can be chosen uniformly for all $(h,\beta), (h_0,\beta_0) \in \Theta$. The final inequality above follows

from boundedness of the hazard function and a conditional expectation argument on $\mathbf{1}_{\left\{\tau_i^{(1)} \geq u\right\}}$. Specifically, $E\left[\mathbf{1}_{\left\{\tau_i^{(1)} \geq u\right\}} \middle| Z^i(t),\ t \in [0,T]\right]$ has a uniform lower bound because the hazard has a uniform lower bound. ■

Proposition 25 *Make assumptions (A1)-(A4), (B1)-(B4). For each j: there exists universal constants $C_1, C_2 > 0$ and $0 < \varrho < 1$ such that, for any $\alpha_i \in \Theta$, $i = 1,2,3$ and $s \in [0,1]$:*

$$\left| E\left(l'_{\alpha_0+s\phi_3}[\phi_1] \cdot l'_{\alpha_0+s\phi_3}[\phi_2]\right) - E\left(l'_{\alpha_0}[\phi_1] \cdot l'_{\alpha_0}[\phi_2]\right)\right| \leq C_1 \|\phi_3\|^{\varrho} \times \|\phi_2\| \times \|\phi_1\|, \tag{20}$$

and

$$\left| E\left(l''_{\alpha_0+s\phi_3}[\phi_1,\phi_2]\right) - E\left(l''_{\alpha_0}[\phi_1,\phi_2]\right)\right| \leq C_2 \|\phi_3\|^{\varrho} \times \|\phi_2\| \times \|\phi_1\|. \tag{21}$$

Proof. We take the Taylor expansion of the term $l'_{\alpha_0+s\phi_3}[\phi_1]$ around $s = 0$. The same can be done for ϕ_2. This gives us:

$$\begin{aligned} l'_{\alpha_0+s\phi_3}[\phi_1] &= \int_0^T \left\{\phi_1\left(Z_u^i\right)\right\} dM_u^i, \\ &\quad - \int_0^T \exp\left(\phi_u^*\right)\left[\phi_3\left(Z_u^i\right)\right]\left[\phi_1\left(Z_u^i\right)\right]\mathbf{1}_{\left\{\tau_i^{(1)} \geq u\right\}} du \times (s-0). \end{aligned} \tag{22}$$

In the above, $\phi_u^* = \left(h_0 + s^*(h_3 - h_0) + \left[\beta_0 + s^*(\beta_3 - \beta_0)\right]' Z_u^i\right)$ where $s^* \in (0,s)$ and is determined by the Taylor expansion. Because of our boundedness assumptions, $\exp\left(\phi_t^*\right) s$ is uniformly bounded over all $t \in [0,T]$, $\alpha_3, \alpha_0 \in \Theta$, $s^*, s \in [0,1]$ and all paths of Z_t^i.

Putting this Taylor expansion into the term (20) gives:

$$\begin{aligned} &\left| E\left(l'_{\alpha_0+s\phi_3}[\phi_1] \cdot l'_{\alpha_0+s\phi_3}[\phi_2]\right) - E\left(l'_{\alpha_0}[\phi_1] \cdot l'_{\alpha_0}[\phi_2]\right)\right|, \\ \leq\ &\left| E\left[\int_0^T \left\{\phi_1\left(Z_u^i\right)\right\} dM_u^i \int_0^T \exp\left(\phi_u^{*,2}\right) s\left[\phi_3\left(Z_u^i\right)\right]\left[\phi_2\left(Z_u^i\right)\right]\mathbf{1}_{\left\{\tau_i^{(1)} \geq u\right\}} du\right]\right|, \qquad (23) \\ &+ \left| E\left[\int_0^T \left\{\phi_2\left(Z_u^i\right)\right\} dM_u^i \int_0^T \exp\left(\phi_u^{*,1}\right) s\left[\phi_3\left(Z_u^i\right)\right]\left[\phi_1\left(Z_u^i\right)\right]\mathbf{1}_{\left\{\tau_i^{(1)} \geq u\right\}} du\right]\right|, \qquad (24) \\ &+ \left| E\left[\begin{array}{l} +\int_0^T \exp\left(\phi_u^{*,1}\right) s\left[\phi_3\left(Z_u^i\right)\right]\left[\phi_1\left(Z_u^i\right)\right]\mathbf{1}_{\left\{\tau_i^{(1)} \geq u\right\}} du \\ \times \int_0^T \exp\left(\phi_u^{*,2}\right) s\left[\phi_3\left(Z_u^i\right)\right]\left[\phi_2\left(Z_u^i\right)\right]\mathbf{1}_{\left\{\tau_i^{(1)} \geq u\right\}} du \end{array}\right]\right|. \qquad (25) \end{aligned}$$

We will show that all the terms (23)-(25) have a bound with the form $C\left\|\phi_3\right\|^{\varrho}\left\|\phi_2\right\|\left\|\phi_1\right\|$ and the result follows. The term (25) has the bound:

$$\begin{aligned}
& \left| E\left\{ \begin{array}{c} \int_0^T \exp\left(\phi_u^{*,1}\right) s\left[\phi_3\left(Z_u^i\right)\right]\left[\phi_1\left(Z_u^i\right)\right] \mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}} du \\ \times \int_0^T \exp\left(\phi_u^{*,2}\right) s\left[\phi_3\left(Z_u^i\right)\right]\left[\phi_2\left(Z_u^i\right)\right] \mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}} du \end{array} \right\} \right|, \\
\leq\ & CE\left\{ \begin{array}{c} \int_0^T \left|\left[\phi_3\left(Z_u^i\right)\right]\right|\left|\left[\phi_1\left(Z_u^i\right)\right]\right| \mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}} du \\ \times \int_0^T \left|\left[\phi_3\left(Z_u^i\right)\right]\right|\left|\left[\phi_2\left(Z_u^i\right)\right]\right| \mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}} du \end{array} \right\}, \\
\leq\ & C \sup_{u\in(0,T], z_t^i\in\overline{\mathcal{Z}}} \left\{ \left| \left\{h_3 - h_0\right\}(u) + \left[\left\{\beta_3 - \beta_0\right\}\right] z_u^i \right|^2 \right\} \qquad (26) \\
& \times E\left\{ \int_0^T \left|\phi_1\left(Z_u^i\right)\right| \mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}} du \times \int_0^T \left|\phi_2\left(Z_u^i\right)\right| \mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}} du \right\}.
\end{aligned}$$

The first inequality follows because of the boundedness discussed above. The constant C holds uniformly as previously mentioned. Notice that s is now gone from the expression. The second inequality uses a supremum over $u \in (0,T]$ because integration makes the value at 0 irrelevant. Lemma 24 and boundedness gives the bound $C'\left\|\alpha_3 - \alpha_0\right\|^{\varrho}$ on the first term in (26). Now we use Hölder's inequality to control the second term in (26).

$$\begin{aligned}
& E\left\{ \int_0^T \left|\phi_1\left(Z_u^i\right)\right| \mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}} du \times \int_0^T \left|\phi_2\left(Z_u^i\right)\right| \mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}} du \right\}, \\
\leq\ & \begin{array}{l} \left[E\left\{ \left[\int_0^T \left|\phi_1\left(Z_u^i\right)\right| \mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}} du \right]^2 \right\} \right]^{1/2} \\ \times \left[E\left\{ \left[\int_0^T \left|\phi_2\left(Z_u^i\right)\right| \mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}} du \right]^2 \right\} \right]^{1/2} \end{array}, \\
\leq\ & \begin{array}{l} \left[E\left\{ \left[\int_0^T \left|\phi_1\left(Z_u^i\right)\right|^2 \mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}} du \right] \right\} \right]^{1/2} \\ \times \left[E\left\{ \left[\int_0^T \left|\phi_2\left(Z_u^i\right)\right|^2 \mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}} du \right] \right\} \right]^{1/2} \end{array}, \\
\leq\ & \begin{array}{l} \left[CE\left\{ \left[\int_0^T \left|\phi_1\left(Z_u^i\right)\right|^2 \exp\left(\alpha_0\left(Z_u^i\right)\right) \mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}} du \right] \right\} \right]^{1/2} \\ \times \left[CE\left\{ \left[\int_0^T \left|\phi_2\left(Z_u^i\right)\right|^2 \exp\left(\alpha_0\left(Z_u^i\right)\right) \mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}} du \right] \right\} \right]^{1/2} \end{array}, \\
=\ & C\left\|\phi_2\right\| \times \left\|\phi_1\right\|.
\end{aligned}$$

The last inequality holds because of the boundedness of $\exp\left(\alpha_0\left(Z_t^i\right)\right)$ following from our assumptions. This gives us the final bound $C''\left\|\phi_3\right\|^{\varrho}\left\|\phi_2\right\|\left\|\phi_1\right\|$.

We now handle the term (23). (24) can be controlled with the same argument:

$$\begin{aligned}
& \left|E\left[\int_0^T\left\{\phi_1\left(Z_u^i\right)\right\}dM_u^i\int_0^T\exp\left(\phi_u^{*,2}\right)s\left[\phi_3\left(Z_u^i\right)\right]\left[\phi_2\left(Z_u^i\right)\right]\mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}}du\right]\right|, \\
\leq\ & E\left[\left|\int_0^T\left\{\phi_1\left(Z_u^i\right)\right\}dM_u^i\int_0^T\exp\left(\phi_u^{*,2}\right)s\left[\phi_3\left(Z_u^i\right)\right]\left[\phi_2\left(Z_u^i\right)\right]\mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}}du\right|\right], \\
\leq\ & \left(E\left[\int_0^T\left\{\phi_1\left(Z_u^i\right)\right\}dM_u^i\right]^2\right)^{1/2}, \\
& \times\left(E\left[\int_0^T\exp\left(\phi_u^{*,2}\right)s\left[\phi_3\left(Z_u^i\right)\right]\left[\phi_2\left(Z_u^i\right)\right]\mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}}du\right]^2\right)^{1/2}, \\
\leq\ & \left\|\phi_1\right\|\times C\left(E\left[\int_0^T\left[\phi_3\left(Z_u^i\right)\right]^2\left[\phi_2\left(Z_u^i\right)\right]^2\mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}}du\right]\right)^{1/2}, \\
\leq\ & \left\|\phi_1\right\|\times C'\sup_{u\in(0,T],z_t^i\in\mathcal{Z}}\left\{\left|\left\{h_3-h_0\right\}(u)+\left[\left\{\beta_3-\beta_0\right\}\right]z_u^i\right|\right\}, \\
& \times\left(E\left[\int_0^T\left[\phi_2\left(Z_u^i\right)\right]^2\mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}}du\right]\right)^{1/2}, \\
\leq\ & C''\left\|\phi_3\right\|^{\varrho}\left\|\phi_2\right\|\left\|\phi_1\right\|.
\end{aligned}$$

The second inequality above is by Hölder's inequality. The rest of the arguments follow what was done for the term (25) above. The final bound is $C''\left\|\phi_3\right\|^{\varrho}\left\|\phi_2\right\|\left\|\phi_1\right\|$. Combining the bounds from the terms (23)-(25) gives the final result.

Now we verify (21). First, Taylor expand the term around $s=0$ as follows:

$$\begin{aligned}
& E\left(l''_{\alpha_0+s\phi_3}\left[\phi_1,\phi_2\right]\right)-E\left(l''_{\alpha_0}\left[\phi_1,\phi_2\right]\right), \\
=\ & -E\left[\int_0^T\exp\left\{\begin{array}{c}h_0+s\left\{h_3-h_0\right\}+\\ \left[\beta_0+s\left\{\beta_3-\beta_0\right\}\right]Z_u^i\end{array}\right\}\left\{\phi_2\left(Z_u^i\right)\right\}\left\{\phi_1\left(Z_u^i\right)\right\}\mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}}du\right], \\
& +E\left[\int_0^T\exp\left\{h_0+\beta_0'Z_u^i\right\}\left\{\phi_2\left(Z_u^i\right)\right\}\left\{\phi_1\left(Z_u^i\right)\right\}\mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}}du\right], \\
=\ & -E\left[\int_0^T\exp\left\{\phi_u^{*,3}\right\}\left\{\phi_3\left(Z_u^i\right)\right\}\left\{\phi_2\left(Z_u^i\right)\right\}\left\{\phi_1\left(Z_u^i\right)\right\}\mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}}du\right]\times(s-0).
\end{aligned}$$

This expansion exists by an argument as in the proof of Lemma 18. We can bound this by using arguments given above:

$$\begin{aligned} & \left| -E\left[\int_0^T \exp\left\{ \phi_u^{*,3} \right\} s\left\{ \phi_3\left(Z_u^i \right) \right\} \left\{ \phi_2\left(Z_u^i \right) \right\} \left\{ \phi_1\left(Z_u^i \right) \right\} \mathbf{1}_{\left\{ \tau_i^{(1)} \geq u \right\}} du \right] \right|, \\ \leq \; & CE\left[\int_0^T \left| \left\{ \phi_3\left(Z_u^i \right) \right\} \left\{ \phi_2\left(Z_u^i \right) \right\} \left\{ \phi_1\left(Z_u^i \right) \right\} \mathbf{1}_{\left\{ \tau_i^{(1)} \geq u \right\}} \right| du \right]. \end{aligned} \tag{27}$$

Notice that s is now gone from this expression. Next, manipulate the term (27)

$$\begin{aligned} & CE\left[\int_0^T \left| \left\{ \phi_3\left(Z_u^i \right) \right\} \left\{ \phi_2\left(Z_u^i \right) \right\} \left\{ \phi_1\left(Z_u^i \right) \right\} \mathbf{1}_{\left\{ \tau_i^{(1)} \geq u \right\}} \right| du \right], \\ \leq \; & C \sup_{u \in (0,T], z_t^i \in \overline{\mathcal{Z}}} \left| \left\{ \left(h_3 - h_0 \right)(u) + \left(\beta_3 - \beta_0 \right)' z_u \right\} \right|, \\ & \times E\left[\int_0^T \left| \phi_2\left(Z_u^i \right) \phi_1\left(Z_u^i \right) \mathbf{1}_{\left\{ \tau_i^{(1)} \geq u \right\}} \right| du \right]. \end{aligned} \tag{28}$$

The first term has the bound $C \left\| \phi_3 \right\|^{\varrho}$ as argued above. The second term in (28) can be handled using Hölder's inequality:

$$\begin{aligned} & E\left[\int_0^T \left| \phi_2\left(Z_u^i \right) \phi_1\left(Z_u^i \right) \mathbf{1}_{\left\{ \tau_i^{(1)} \geq u \right\}} \right| du \right], \\ \leq \; & E\left[\left(\int_0^T \left\{ \phi_2\left(Z_u^i \right) \right\}^2 \mathbf{1}_{\left\{ \tau_i^{(1)} \geq u \right\}} du \right)^{1/2} \times \left(\int_0^T \left\{ \phi_1\left(Z_u^i \right) \right\}^2 \mathbf{1}_{\left\{ \tau_i^{(1)} \geq u \right\}} du \right)^{1/2} \right], \\ \leq \; & \left(E\left[\int_0^T \left\{ \phi_2\left(Z_u^i \right) \right\}^2 \mathbf{1}_{\left\{ \tau_i^{(1)} \geq u \right\}} du \right] \right)^{1/2} \times \left(E\left[\int_0^T \left\{ \phi_1\left(Z_u^i \right) \right\}^2 \mathbf{1}_{\left\{ \tau_i^{(1)} \geq u \right\}} du \right] \right)^{1/2}, \\ \leq \; & C' \left\| \phi_2 \right\| \times \left\| \phi_1 \right\|. \end{aligned} \tag{29}$$

The first and second inequality follow from Hölder's inequality. The last line follows from boundedness of the term $\exp\left(\alpha_0\left(Z_u^i \right) \right)$. This gives the desired bound $C'' \left\| \phi_3 \right\|^{\varrho} \left\| \phi_2 \right\| \left\| \phi_1 \right\|$. This argument can be repeated for each j. ■

Corollary 26 *Make assumptions (A1)-(A4), (B1)-(B4). For each j: with probability approaching one:*

$$\frac{1}{2} \left| \left\| u_n \right\|^2 - \left\{ -Q_0''\left(\alpha_0 + t_n u_n \right)\left[u_n \right] \right\} \right| \leq \frac{1}{2} C_2 \left\| u_n \right\|^{2+\varrho}.$$

Proof. This follows from Lemma 23 and Proposition 25. The argument can be repeated for each j. ■

A.3.2 Rate of Convergence Proof

Because of assumption (A2)(ii) we will need the following decomposition in the sequel:

$$\begin{aligned}
& l'_{\alpha_0+s(\alpha-\alpha_0)}\left[\alpha-\alpha_0\right](Z_t,\tau), \\
= & \int_0^T \left\{\phi(Z_u)\right\} dN_u^j - \int_0^T \exp\left[\left(\alpha_0+s(\alpha-\alpha_0)\right)(Z_u)\right]\cdot\phi(Z_u)\,\mathbf{1}_{\left\{\tau^{(1)}\geq u\right\}}du, \\
= & \sum_{l=0}^{m}\int_{[t_l,t_{l+1}]}\left\{\phi(Z_u)\right\} dN_u^j - \int_{t_l}^{t_{l+1}} \exp\left[\left(\alpha_0+s(\alpha-\alpha_0)\right)(Z_u)\right]\cdot\phi(Z_u)\,\mathbf{1}_{\left\{\tau^{(1)}\geq u\right\}}du, \\
\equiv & \sum_{l=0}^{m} l'_{\alpha_0+s(\alpha-\alpha_0)}\left[\alpha-\alpha_0\right](Z_t,\tau,t_l).
\end{aligned}$$

We ignore overlap in the first term's support because events happen at $\{t_l\}_{l=0}^m$ with probability zero. This decomposition defines the terms $l'_{\alpha_0+s(\alpha-\alpha_0)}\left[\alpha-\alpha_0\right](Z_t,\tau,t_l)$ for each $\{t_l\}_{l=0}^m$.

We define the following set of functions for each t_l and n:

$$\mathcal{G}_{t_l,n} = \left\{ n^{-\delta}\left\|\alpha-\alpha_0\right\|^{-1} l'_{\alpha_0+s_n(\alpha-\alpha_0)}\left[\alpha-\alpha_0\right](\cdot,\cdot,t_l)\middle|\, \epsilon_0 > \left\|\alpha-\alpha_0\right\| \geq n^{-\delta}, \alpha\in\Theta_n \right\}.$$

These are functions of Z_{t_l} and τ and therefore are functions from $[0,\infty)^R\times\mathcal{Z}$ to $\mathbb{R}$. s_n is defined previously in Lemma 20. Under our assumptions, all sets of functions $\mathcal{G}_{t_l,n}$ are uniformly bounded. This holds because $n^{-\delta}\left\|\alpha-\alpha_0\right\|^{-1}$ is uniformly bounded and so is the term $l'_{\alpha_0+s_n(\alpha-\alpha_0)}\left[\alpha-\alpha_0\right](Z_t,\tau,t_l)$. This boundedness allows for an application of Alexander (1984) Corollary 2.2. For this, we will need the following lemmas.

Lemma 27 *Make the assumptions (A1)-(A4), (B1)-(B4). For any j: Define*

$$\theta_n^* = \sup_{\mathcal{G}_{t_l,n}} var\left[n^{-\delta}\left\|\alpha-\alpha_0\right\|^{-1} l'_{\alpha_0+s_n(\alpha-\alpha_0)}\left[\alpha-\alpha_0\right](Z_t,\tau,t_l)\right].$$

For some universal constant $C > 0$, for all n:

$$\theta_n^* \leq Cn^{-2\delta}. \tag{30}$$

This holds for any t_l.

Proof. The smoothness result (20) given in Proposition 25 holds for the terms $l'_{\alpha_0+s(\alpha-\alpha_0)}[\alpha-\alpha_0](Z_t,\tau,t_l)$, which simply restrict the terms in (20) to the subintervals $[t_l, t_{l+1}]$ where the covariates do not update. This holds with exactly the same proof as that of Proposition 25. In what follows, we add to terms an argument or subscript t_l to represent that we are considering the integrals in those terms restricted to the subinterval $[t_l, t_{l+1}]$. Therefore, using the result (20):

$$\left|E\left(l'_{\alpha_0+s\phi_3}[\phi_1](t_l)\cdot l'_{\alpha_0+s\phi_3}[\phi_2](t_l)\right) - E\left(l'_{\alpha_0}[\phi_1](t_l)\cdot l'_{\alpha_0}[\phi_2](t_l)\right)\right| \leq C_1\left\|\phi_3\right\|_{t_l}^{\varrho}\times\left\|\phi_2\right\|_{t_l}\times\left\|\phi_1\right\|_{t_l}. \tag{31}$$

Choosing $\phi_1=\phi_2=\phi_3=\alpha-\alpha_0$, we get:

$$\begin{aligned} & E\left(l'_{\alpha_0+s(\alpha-\alpha_0)}[\alpha-\alpha_0](t_l)\cdot l'_{\alpha_0+s(\alpha-\alpha_0)}[\alpha-\alpha_0](t_l)\right), \\ \leq\ & E\left(l'_{\alpha_0}[\alpha-\alpha_0](t_l)\cdot l'_{\alpha_0}[\alpha-\alpha_0](t_l)\right) + C^*\left\|\alpha-\alpha_0\right\|_{t_l}^{2+\varrho}. \end{aligned}$$

Note that

$$\begin{aligned} E\left(l'_{\alpha_0}[\alpha-\alpha_0](t_l)\cdot l'_{\alpha_0}[\alpha-\alpha_0](t_l)\right) &= \left\|\alpha-\alpha_0\right\|_{t_l}^2, \\ &= E\left[\int_{t_l}^{t_{l+1}}\left\{(\alpha-\alpha_0)\left(Z_u^i\right)\right\}^2\exp\left(\alpha_0\left(Z_u^i\right)\right)\mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}}du\right], \\ &\leq E\left[\int_0^{T}\left\{(\alpha-\alpha_0)\left(Z_u^i\right)\right\}^2\exp\left(\alpha_0\left(Z_u^i\right)\right)\mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}}du\right], \\ &= \left\|\alpha-\alpha_0\right\|^2. \end{aligned}$$

Therefore, we have

$$\begin{aligned} E\left(l'_{\alpha_0+s(\alpha-\alpha_0)}\left[\alpha-\alpha_0\right](t_l)\cdot l'_{\alpha_0+s(\alpha-\alpha_0)}\left[\alpha-\alpha_0\right](t_l)\right) &\leq \left\|\alpha-\alpha_0\right\|_{t_l}^2 + C^*\left\|\alpha-\alpha_0\right\|_{t_l}^{2+\varrho}, \\ &= \left\|\alpha-\alpha_0\right\|_{t_l}^2\left(1+C^*\left\|\alpha-\alpha_0\right\|_{t_l}^{\varrho}\right), \\ &\leq \left\|\alpha-\alpha_0\right\|^2\left(1+C^*\left\|\alpha-\alpha_0\right\|^{\varrho}\right). \end{aligned}$$

However, the variance of functions in $\mathcal{G}_{t_l,n}$ is bounded by

$$n^{-2\delta}\left\|\alpha-\alpha_0\right\|^{-2} E\left(l'_{\alpha_0+s(\alpha-\alpha_0)}\left[\alpha-\alpha_0\right](t_l)\cdot l'_{\alpha_0+s(\alpha-\alpha_0)}\left[\alpha-\alpha_0\right](t_l)\right).$$

Combining this with the above, we get

$$\begin{aligned} var\left[n^{-\delta}\left\|\alpha-\alpha_0\right\|^{-1} l'_{\alpha_0+s_n(\alpha-\alpha_0)}\left[\alpha-\alpha_0\right](Z_t,\tau,t_l)\right] &\leq n^{-2\delta}\left(1+C^*\left\|\alpha-\alpha_0\right\|^{\varrho}\right), \\ &\leq n^{-2\delta}\left(1+C^*\epsilon_0^{\varrho}\right). \end{aligned}$$

This implies (30) as desired. ■

See Alexander (1984) for a definition of L^∞ metric entropy.

Lemma 28 *Make the assumptions (A1)-(A4), (B1)-(B4). For each j: there exists a universal constant $A>0$ such that; for all $s\in(1/2,1)$, $n\in\mathbb{N}$ and $l\in\{0,1,\ldots,m\}$, the L^∞ metric entropy of $\mathcal{G}_{t_l,n}$ has the following bound:*

$$H_\infty\left(\epsilon,\mathcal{G}_{t_l,n}\right)\leq A\epsilon^{-2(d+R)/\varpi}. \tag{32}$$

Proof. Functions in $\mathcal{G}_{t_l,n}$ can be shown to have this entropy bound by using entropy bounds for Hölder balls. Consider the likelihood parts of the functions in $\mathcal{G}_{t_l,n}$:

$$\begin{aligned} & l'_{\alpha_0+s(\alpha-\alpha_0)}\left[\alpha-\alpha_0\right](Z_t,\tau,t_l), \qquad (33) \\ = & \int_{[t_l,t_{l+1}]}\left\{\phi\left(Z_u\right)\right\}dN_u^j - \int_{t_l}^{t_{l+1}}\exp\left[\left(\alpha_0+s\left(\alpha-\alpha_0\right)\right)\left(Z_u\right)\right]\cdot\phi\left(Z_u\right)\mathbf{1}_{\left\{\tau^{(1)}\geq u\right\}}du. \end{aligned}$$

These are functions with support $[0,\infty)^R \times \mathcal{Z}$ where $[0,\infty)^R$ represents the support for the R random events (one for each type) and $\mathcal{Z}$ the support of the covariates. As we vary over the parameters $h, h_0 \in \mathcal{H}, \beta, \beta_0 \in B$ this defines a class of functions. If we multiply these functions by $n^{-\delta} \|\alpha - \alpha_0\|^{-1}$, this class of functions contains all $\mathcal{G}_{t_l,n}$. If we derive the entropy bound for this larger class of functions, then this will bound the entropy for all $\mathcal{G}_{t_l,n}$. In what follows, the entropy bound $A\epsilon^{-2(d+R)/\varpi}$ is derived for this class of functions. We argue that each of the two pieces of (33) defines a class of functions and each class has L^∞ entropy bounded by $A'\epsilon^{-(d+R)/\varpi}$. Combining the functions defining these entropy numbers gives the final entropy bound $A\epsilon^{-2(d+R)/\varpi}$.

Consider the class of functions $\int_{[t_l,t_{l+1}]} \{\phi(Z_u)\} dN_u^j$ first. Temporarily assume the other $j-1$ event types do not censor N_u^j. Then this class of functions restricted to the support $[t_l, t_{l+1}]^R \times \mathcal{Z}$ are all in the same Hölder ball $\mathcal{L}_{\varpi,1,K}\left([t_l, t_{l+1}]^R \times \mathcal{Z}\right)$ for some $K > 0$. This holds because of our assumption (A1)-(A3), (B1) and the form of the functions. Multiplying by $n^{-\delta} \|\alpha - \alpha_0\|^{-1} \leq 1$ keeps these functions in the same Hölder ball. Hölder balls with these smoothness conditions have the L^∞ entropy bound $A'\epsilon^{-(d+R)/\varpi}$ where A' is a universal constant. See Alexander (1984) for a statement of this entropy result, first derived in Kolmogorov and Tihomirov (1959). Therefore, the L^∞ entropy bound $A'\epsilon^{-(d+R)/\varpi}$ holds for this class of functions over the support $[t_l, t_{l+1}]^R \times \mathcal{Z}$.

The remaining parts of the support do not increase the entropy number. This is because the functions defining the covering on $[t_l, t_{l+1}]^R \times \mathcal{Z}$ can be extended to the whole support while preserving their sup norm distance between functions with the form $\int_{[t_l,t_{l+1}]} \{\phi(Z_u)\} dN_u^j$. This can be done by setting the functions defining the covering to zero when τ^j is outside of $[t_l, t_{l+1}]$. The other $j-1$ event types have no effect on this class of functions when N_u^j is not censored, so extending these to their full support trivially preserves the covering number as well.

Adding back in censoring by the other $j-1$ event types for N_u^j does not increase the entropy bound. This is because the previously derived covering functions can be modified in regions of $[0,\infty)^R \times \mathcal{Z}$ where the censoring matters in such a way that preserves their sup norm distance from function in the class considered. We simply set the covering functions derived before to zero when the censoring matters. Therefore, we have the entropy bound $A'\epsilon^{-(d+R)/\varpi}$ over this class of functions.

A similar argument shows that the class of functions defined by the second term in (33) has the same form of entropy bound $A''\epsilon^{-(d+R)/\varpi}$. We give an outline of the proof here. In this case, initially replace $\tau^{(1)}$ with τ^j, then argue that the entropy number has the correct bound over the restricted support $[t_l, t_{l+1}]^R \times \mathcal{Z}$. This is done using a Hölder ball argument as above. Second, extend the covering functions to the full support. Finally, replace τ^j with $\tau^{(1)}$ and modify the covering functions to preserve the entropy number. Combining covering functions for these two terms gives the final bound $A\epsilon^{-2(d+R)/\varpi}$. This bounds the entropy for all $\mathcal{G}_{t_l,n}$ as discussed above. We can bound the entropy over all l by repeating the proof and choosing the largest constant A. this entire program can be repeated for each j. ■

Lemma 29 *Make the assumptions (A1)-(A4), (B1)-(B4). For each j:*

$$\sup_{\alpha\in\Theta_n, n^{-\delta}\leq\|\alpha-\alpha_0\|\leq\epsilon_0} \left| \|\alpha-\alpha_0\|^{-1} \left\{ \begin{array}{c} Q_n'\left(\alpha_0+s_n\left(\alpha-\alpha_0\right)\right)\left[\alpha-\alpha_0\right] \\ -Q_0'\left(\alpha_0+s_n\left(\alpha-\alpha_0\right)\right)\left[\alpha-\alpha_0\right] \end{array} \right\} \right| = O_p\left(n^{-1/2+\delta-k}\right), \quad (34)$$

where $\epsilon_0 > 0$.

Proof. This rate is derived using results from Alexander (1984) (Hereafter AL). The supremum (34) is bounded by the sum of suprema:

$$\sum_{l=0}^{m} \sup_{\alpha\in\Theta_n, n^{-\delta}\leq\|\alpha-\alpha_0\|\leq\epsilon_0} \left| \|\alpha-\alpha_0\|^{-1} \left\{ \begin{array}{c} Q_n'\left(\alpha_0+s_n\left(\alpha-\alpha_0\right)\right)\left[\alpha-\alpha_0\right](t_l) \\ -Q_0'\left(\alpha_0+s_n\left(\alpha-\alpha_0\right)\right)\left[\alpha-\alpha_0\right](t_l) \end{array} \right\} \right|. \quad (35)$$

Here there is one term in the sum for each subinterval where the data does not update $(t_l, t_{l+1}]$. We prove that each of the suprema in the sum are $O_p\left(n^{-1/2+\delta-k}\right)$ and therefore the initial supremum (34) is $O_p\left(n^{-1/2+\delta-k}\right)$. The edge terms can be ignored because they are zero with probability one. Notice also that these terms differ from the functions in $\mathcal{G}_{t_i,n}$ because they do not have the $n^{-\delta}$ term multiplying them. This means that whatever rate is derived using AL must be reduced by a n^{δ} term.

We now derive the rate $O_p\left(n^{-1/2+\delta-k}\right)$ for the terms in the sum of (35). The proof is similar to WS Theorem 2. An outline of how Corollary 2.2 from AL will be used to prove the required rate

of convergence is first presented. More conditions will be needed for the result to hold. These will be given after the outline. Corollary 2.2 from AL proves that:

$$P\left\{\sup_{\mathcal{G}}\left|\nu_n\left(f\left(X_i\right)\right)\right|>M\right\}\leq 5\exp\left(-C\psi\left(M,n,\theta\right)\right), \tag{36}$$

where $\mathcal{G}$ is a class of uniformly bounded functions which take X_i as arguments. X_i are $i.i.d.$ random variables, $\nu_n\left(f\left(X_i\right)\right)$ gives the centered empirical process multiplied by $n^{-1/2}$ and θ is a uniform bound on the variance of $f\left(X_i\right)$ over $\mathcal{G}$. A sequence of these probability bounds are used to prove a rate of convergence for (35). The following function is used when applying Corollary 2.2 in AL:

$$\psi\left(M,n,\theta\right)=Mn^{1/2}h_2\left(M/\left(n^{1/2}\theta\right)\right),$$

where

$$h_2\left(\lambda\right)=\lambda/\left(2\left(1+\lambda/3\right)\right).$$

We additionally assume

$$M=M_n=Dn^{-k},$$

where D is an arbitrary constant and k is described below. Therefore,

$$\begin{aligned}
\psi\left(M_n,n,\theta\right) &= M_n n^{1/2}\frac{M_n/\left(n^{1/2}\theta\right)}{\left(2\left(1+M_n/\left(3n^{1/2}\theta\right)\right)\right)},\\
&= Dn^{-k}n^{1/2}\frac{Dn^{-k}/\left(n^{1/2}\theta\right)}{\left(2\left(1+Dn^{-k}/\left(3n^{1/2}\theta\right)\right)\right)},\\
&= \frac{n^{1/2}\theta}{n^{1/2}\theta}\cdot\frac{D^2n^{-2k}/\left(\theta\right)}{\left(2\left(1+Dn^{-k}/\left(3n^{1/2}\theta\right)\right)\right)},\\
&= \frac{D^2n^{1/2-2k}}{\left(\left(2n^{1/2}\theta+\frac{D}{3}n^{-k}\right)\right)}.
\end{aligned}$$

We want to bound $\exp\left(-\frac{1}{2}\psi\left(M,n,\theta\right)\right)$. To do this we replace θ with θ_n^* defined in Lemma 27, which gives an upper bound for (36). The result is:

$$\psi\left(M,n,\theta_n^*\right)=\frac{D^2n^{1/2-2k}}{\left(\left(2n^{1/2}\theta_n^*+\frac{D}{3}n^{-k}\right)\right)}.$$

This can be simplified:

$$\frac{D^2 n^{1/2-2k}}{\left(\left(2n^{1/2}\theta_n^* + \frac{D}{3}n^{-k}\right)\right)} = \frac{1}{\frac{2}{D^2}n^{2k}\theta_n^* + \frac{1}{3D}n^{k-1/2}}.$$

Using the bound on θ_n^* from Lemma 27, we get:

$$\frac{1}{\frac{2C}{D^2}n^{2k-2\delta} + \frac{1}{3D}n^{k-1/2}} \leq \frac{1}{\frac{2}{D^2}n^{2k}\theta_n^* + \frac{1}{3D}n^{k-1/2}}.$$

If $k \leq \delta$ and $k \leq 1/2$, then

$$\lim_{n\to\infty} \frac{1}{\frac{2C}{D^2}n^{2k-2\delta} + \frac{1}{3D}n^{k-1/2}} = \infty,$$

for arbitrary $D > 0$. This implies

$$\exp\left(-C\psi\left(M_n, n, \theta_n^*\right)\right) \to 0.$$

The AL result gives the rate of convergence for the centered empirical process multiplied by $n^{-1/2}$. This implies that the final rate should be $O_p\left(n^{-1/2+\delta-k}\right)$ because of the $n^{-1/2}$ difference and the n^{δ} adjustment discussed above. The restrictions on k and δ imply that $1/2 - \delta + k \leq 1/2$.

More conditions are needed for the above outline of the proof to hold. These conditions are verified here. By Lemma 28, the L^∞ metric entropy of all sets of functions $\mathcal{G}_{t_l,n}$, $n \in \mathbb{N}$ is uniformly bounded as follows:

$$H_\infty\left(t, \mathcal{G}_{t_l,n}\right) \leq A\epsilon^{-2(d+R)/\varpi}. \tag{37}$$

The exponential bounds from AL described above require the L^∞ metric entropy to be bounded with the form of (37). A sequence of these exponential bounds are applied in order to derive rates of convergence.

In Corollary 2.2 of AL, we fix $\epsilon = 1/2$ throughout. Because the entropy bound is uniform over all $\mathcal{G}_{t_l,n}$, the constants K_1, K_2 and K_3 in this corollary are fixed for all n. These constants only depend on ϵ, A and $2\left(d+R\right)/\varpi$, which are fixed for all $\mathcal{G}_{t_l,n}$. We now have the prerequisites to apply AL Corollary 2.2 to our situation.

The assumed $1 > \left(d+R\right)/\varpi$ is required for a faster than $1/4$ rate of convergence. The

conditions in (2.7) from AL need to hold for the bounds (36) to be used to derive convergence rates as outlined above. The following conditions should be satisfied:

$$M_n = Dn^{-k} \geq K_1 C \left(n^{-2\delta}\right)^{(2-2((d+R)/\varpi))/4}, \tag{38}$$

and

$$M_n = Dn^{-k} \geq K_3 n^{[2((d+R)/\varpi)-2]/[2(2((d+R)/\varpi)+2)]}. \tag{39}$$

(38) implies

$$-k \geq -2\delta \left(\frac{1}{2} - \left(\frac{d+R}{2\varpi}\right)\right),$$

and therefore

$$\delta \geq \left(\frac{\varpi}{\varpi - d - R}\right) k. \tag{40}$$

(39) implies

$$-k \geq \frac{2\left(\left(d+R\right)/\varpi\right) - 2}{2\left(2\left(\left(d+R\right)/\varpi\right) + 2\right)},$$

and therefore

$$\frac{(\varpi - d - R)}{2\left(\varpi + d + R\right)} \geq k. \tag{41}$$

(40) and (41) were assumed in (B3). Therefore, the rate of convergence of terms in (35) is $O_p\left(n^{-1/2+\delta-k}\right)$ by our previous outline of the convergence rate proof given above. The above arguments can be made for each subinterval $[t_l, t_{l+1}]$. Therefore, the rate for the sum is:

$$\sum_{l=0}^{m} O_p\left(n^{-1/2+\delta-k}\right),$$

which implies (34) is $O_p\left(n^{-1/2+\delta-k}\right)$. This argument can be repeated for each j. ■

Lemma 30 *Make the assumptions (A1)-(A4), (B1)-(B4). For each j:*

$$\left|Q_n'\left(\widetilde{\phi}_n\right)[\widetilde{u}_n] - Q_n'\left(\widetilde{\phi}_n\right)[u_n]\right| = O_p\left(n^{-1/2-\varkappa_1/2}\right) + O\left(g_n\right), \tag{42}$$

$$\left|Q_n'\left(\widetilde{\phi}_n\right)[u_n] - Q_n'\left(\alpha_0 + s_n u_n\right)[u_n]\right| = O\left(g_n\right), \tag{43}$$

Proof. The first term can be bounded as follows:

$$\begin{aligned}
& \left|Q_n'\left(\widetilde{\phi}_n\right)[\widetilde{u}_n] - Q_n'\left(\widetilde{\phi}_n\right)[u_n]\right|, \\
= & \left|\frac{1}{n}\sum_{i=1}^{n}\int_0^T\left[\left(h_0-\widetilde{h}_n\right)(u)\right]dN_u^i - \int_0^T \exp\left(\widetilde{\phi}_n\right)\left[\left(h_0-\widetilde{h}_n\right)(u)\right]\mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}}du\right|, \\
\leq & \left|\frac{1}{n}\sum_{i=1}^{n}\int_0^T\left[\left(h_0-\widetilde{h}_n\right)(u)\right]dN_u^i\right| + C\int_0^T\left|\left(h_0-\widetilde{h}_n\right)(u)\right|du.
\end{aligned}$$

This uses boundedness of $\exp\left(\widetilde{\phi}_n\right)$, which follows from our assumptions. The second term can be bounded:

$$\begin{aligned}
& \left|Q_n'\left(\widetilde{\phi}_n\right)[u_n] - Q_n'\left(\alpha_0+s_nu_n\right)[u_n]\right|, \\
= & \left|\frac{1}{n}\sum_{i=1}^{n}\int_0^T\left\{\exp\left\{\widetilde{\phi}_n\right\} - \exp\left(\alpha_0+s_nu_n\right)\right\}\times\left\{\left(\widehat{h}-h_0\right)(u) + \left(\widehat{\beta}-\beta_0\right)'Z_u^i\right\}\mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}}du\right|, \\
\leq & C\int_0^T\left|\left(\widetilde{h}_n-h_0\right)(u)\right|du.
\end{aligned}$$

This follows from our boundedness assumptions and a Taylor expansion of the terms $\exp\left\{\widetilde{\phi}_n\right\} - \exp\left(\alpha_0+s_nu_n\right)$. Therefore, rates for these terms depend on:

$$\int_0^T\left|\left(\widetilde{h}_n-h_0\right)(u)\right|du, \tag{44}$$

and

$$\left|\frac{1}{n}\sum_{i=1}^{n}\int_0^T\left[\left(h_0-\widetilde{h}_n\right)(u)\right]dN_u^i\right|. \tag{45}$$

The rate for (44) follows from assumption (B4). For (45), as the event N_t^i can happen at any time in $[0,T]$, It would seem this term must be controlled with the sup norm:

$$\sup_{t\in[0,T]}\left|\left(h_0-\widetilde{h}_n\right)(t)\right|.$$

However, using Burkholder's inequality, the requirement that we approximate in the sup norm can

be weakened. See Protter (2005) for a statement of Burkholder's inequality. Specifically:

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}\int_0^T \left[\left(h_0-\widetilde{h}_n\right)(u)\right] dN_u^i, \\
= \; & \frac{1}{n}\sum_{i=1}^{n}\int_0^T \left[\left(h_0-\widetilde{h}_n\right)(u)\right] dM_u^i, \\
& +\frac{1}{n}\sum_{i=1}^{n}\int_0^T \left[\left(h_0-\widetilde{h}_n\right)(u)\right] \left\{\exp\left(h_0(u)+\beta_0' Z^i(u)\right)\right\} \mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}} du.
\end{aligned}$$

The first term here is a martingale as described in the beginning of this appendix. This term can be handled with the Markov and Burkholder inequalities. Choose $\varkappa_1 > 0$ as in assumption (B4). Consider the following probability bound:

$$\begin{aligned}
& P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\int_0^T \left[\left(h_0-\widetilde{h}_n\right)(u)\right] dM_u^i\right| \geq n^{-1/2-\varkappa_1/2}\epsilon\right\}, \\
\leq \; & P\left\{\sup_{t\in[0,T]}\left|\frac{1}{n}\sum_{i=1}^{n}\int_0^t \left[\left(h_0-\widetilde{h}_n\right)(u)\right] dM_u^i\right| \geq n^{-1/2-\varkappa_1/2}\epsilon\right\}, \\
= \; & P\left\{\sup_{t\in[0,T]}\left[\left|\sum_{i=1}^{n}\int_0^t \left[\left(h_0-\widetilde{h}_n\right)(u)\right] dM_u^i\right|^2\right] \geq n^2 n^{-1-\varkappa_1}\epsilon^2\right\}, \\
\leq \; & \frac{1}{n^{1-\varkappa_1}\epsilon^2} E\left\{\sup_{t\in[0,T]}\left[\left|\sum_{i=1}^{n}\int_0^t \left[\left(h_0-\widetilde{h}_n\right)(u)\right] dM_u^i\right|^2\right]\right\}.
\end{aligned}$$

This follows from simple manipulations, properties of the supremum and Markov's inequality. Now we apply Burkholder's inequality:

$$\begin{aligned}
& \frac{1}{n^{1-\varkappa_1}\epsilon^2} E\left\{\sup_{t\in[0,T]}\left[\left|\sum_{i=1}^{n}\int_0^t \left[\left(h_0-\widetilde{h}_n\right)(u)\right] dM_u^i\right|^2\right]\right\} \\
\leq \; & \frac{1}{n^{1-\varkappa_1}\epsilon^2} C_p E\left\{\left[\sum_{i=1}^{n}\int_0^{\cdot} \left[\left(h_0-\widetilde{h}_n\right)(u)\right] dM_u^i\right]_T\right\}.
\end{aligned}$$

See Protter (2005) for a definition of the bracket process $[\cdot]_t$. Properties of the bracket process

give us:

$$\begin{aligned}\left[\sum_{i=1}^{n}\int_0^{\cdot}\left[\left(h_0-\widetilde{h}_n\right)(u)\right]dM_u^i\right]_T &= \sum_{i=1}^{n}\int_0^T\left[\left(h_0-\widetilde{h}_n\right)(u)\right]^2 dN_u^i,\\ &\leq n\sup_{t\in[0,T]}\left[\left(h_0-\widetilde{h}_n\right)(t)\right]^2,\\ &\leq Cnn^{-\varkappa_1-\varkappa_2}.\end{aligned}$$

The final inequality follows by assumption (B4). Therefore

$$\begin{aligned}\frac{1}{n^{1-\varkappa_1}\epsilon^2}C_pE\left\{\left[\sum_{i=1}^{n}\int_0^t\left[\left(h_0-\widetilde{h}_n\right)(u)\right]dM_u^i\right]_T\right\} &\leq \frac{1}{n^{1-\varkappa_1}\epsilon^2}C_pCnn^{-\varkappa_1-\varkappa_2},\\ &= \frac{1}{n^{\varkappa_2}\epsilon^2}C_pC\to 0.\end{aligned}$$

This shows that the term $\frac{1}{n}\sum\int_0^T\left[\left(h_0-\widetilde{h}_n\right)(u)\right]dM_u^i = O_p\left(n^{-1/2-\varkappa_1/2}\right)$ which is faster than $O_p\left(n^{-1/2}\right)$. Finally, the rate of

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^T\left[\left(h_0-\widetilde{h}_n\right)(u)\right]\left\{\exp\left(h_0(u)+\beta_0'Z_u^i\right)\right\}\mathbf{1}_{\left\{\tau_i^{(1)}\geq u\right\}}du,$$

or

$$C\int_0^T\left[\left(h_0-\widetilde{h}_n\right)(u)\right]du,$$

needs to be controlled. By our assumption (B4), the final rate for term (45) is $O_p\left(n^{-1/2-\varkappa_1/2}\right)+O(g_n)$. ■

Proof (Theorem 10). Because of the form of the Fisher norm and our assumptions (A1)-(A3) and (B1)-(B2), for each j, $\left\|\alpha^j-\alpha_0^j\right\|^2=0$ if and only if $\alpha^j=\alpha_0^j$. Therefore, $\left\|\alpha^{1:R}-\alpha_0^{1:R}\right\|_{1:R}=0$ if and only if $\alpha^{1:R}=\alpha_0^{1:R}$. As a result, as argued in CLS, $clsp\left(\Theta^1\times\cdots\times\Theta^R\right)-\alpha_0^{1:R}$ is a Hilbert space w.r.t. the norm $\|\cdot\|_{1:R}$.

In what follows we show that for each j, $\left\|\widehat{\alpha}^j-\alpha_0^j\right\|$ has the required rate. This implies that $\left\|\widehat{\alpha}^{1:R}-\alpha_0^{1:R}\right\|_{1:R}$ has the claimed rate by the form of $\|\cdot\|_{1:R}$. The argument is for an arbitrary j in what follows. We no longer index by j in this proof.

From our assumptions, the following holds:

$$Q_0'(\alpha_0)[u_n] \leq 0, \tag{46}$$

$$Q_n\left(\pi_n\alpha_0 - \frac{1}{2}\widetilde{u}_n\right) - Q_n(\widehat{\alpha}) \leq O_p\left(\varepsilon_n^2\right). \tag{47}$$

(46) follows from Lemma 19. (47) follows from the definition of the estimator and Lemma 22. By the consistency result and the form of the norm $\|\cdot\|$, with probability approaching one, $\|u_n\| < \epsilon_0$. Using (47) and the mean value theorem from Lemma 20, there exists s_n with $\frac{1}{2} < s_n < 1$ such that:

$$-Q_n'(\pi_n\alpha_0 + s_n\widetilde{u}_n)[\widetilde{u}_n] \leq 2O_p\left(\varepsilon_n^2\right),$$

or

$$-Q_n'\left(\widetilde{\phi}_n\right)[\widetilde{u}_n] \leq 2O_p\left(\varepsilon_n^2\right), \tag{48}$$

where s_n is defined above.

By Lemma 21 the following holds: there exists $\widetilde{t}_n \in [0, s_n]$ such that:

$$Q_0'(\alpha_0 + s_nu_n)[u_n] = Q_0'(\alpha_0)[u_n] + s_nQ_0''\left(\alpha_0 + \widetilde{t}_nu_n\right)[u_n].$$

Therefore

$$\begin{aligned}
-Q_0''\left(\alpha_0 + \widetilde{t}_nu_n\right)[u_n] &= \frac{1}{s_n}\left\{Q_0'(\alpha_0)[u_n] - Q_0'(\alpha_0 + s_nu_n)[u_n]\right\}, \\
&\leq 2\left\{Q_0'(\alpha_0)[u_n] - Q_0'(\alpha_0 + s_nu_n)[u_n]\right\}, \\
&= 2\left\{\begin{array}{c} Q_0'(\alpha_0)[u_n] - Q_n'\left(\widetilde{\phi}_n\right)[\widetilde{u}_n] \\ +Q_n'\left(\widetilde{\phi}_n\right)[\widetilde{u}_n] - Q_n'\left(\widetilde{\phi}_n\right)[u_n] \\ +Q_n'\left(\widetilde{\phi}_n\right)[u_n] - Q_n'(\alpha_0 + s_nu_n)[u_n] \\ +Q_n'(\alpha_0 + s_nu_n)[u_n] - Q_0'(\alpha_0 + s_nu_n)[u_n] \end{array}\right\}.
\end{aligned} \tag{49}$$

Consider the first line of (49). By (46) and (48), the first line in (49) is $O_p\left(\varepsilon_n^2\right)$. Therefore:

$$-Q_0''\left(\alpha_0+\widetilde{t}_n u_n\right)[u_n] \leq 2\left\{\begin{array}{c} O_p\left(\varepsilon_n^2\right) \\ +Q_n'\left(\widetilde{\phi}_n\right)[\widetilde{u}_n]-Q_n'\left(\widetilde{\phi}_n\right)[u_n] \\ +Q_n'\left(\widetilde{\phi}_n\right)[u_n]-Q_n'\left(\alpha_0+s_n u_n\right)[u_n] \\ +Q_n'\left(\alpha_0+s_n u_n\right)[u_n]-Q_0'\left(\alpha_0+s_n u_n\right)[u_n] \end{array}\right\}.$$

Next, subtract and add $\|u_n\|^2$ from the left side:

$$\|u_n\|^2-\|u_n\|^2-Q_0''\left(\alpha_0+\widetilde{t}_n u_n\right)[u_n] \leq 2\left\{\begin{array}{c} O_p\left(\varepsilon_n^2\right) \\ +Q_n'\left(\widetilde{\phi}_n\right)[\widetilde{u}_n]-Q_n'\left(\widetilde{\phi}_n\right)[u_n] \\ +Q_n'\left(\widetilde{\phi}_n\right)[u_n]-Q_n'\left(\alpha_0+s_n u_n\right)[u_n] \\ +Q_n'\left(\alpha_0+s_n u_n\right)[u_n]-Q_0'\left(\alpha_0+s_n u_n\right)[u_n] \end{array}\right\}.$$

Rearranging terms gives:

$$\|u_n\|^2 \leq 2\left\{\begin{array}{c} O_p\left(\varepsilon_n^2\right) \\ +Q_n'\left(\widetilde{\phi}_n\right)[\widetilde{u}_n]-Q_n'\left(\widetilde{\phi}_n\right)[u_n] \\ +Q_n'\left(\widetilde{\phi}_n\right)[u_n]-Q_n'\left(\alpha_0+s_n u_n\right)[u_n] \\ +Q_n'\left(\alpha_0+s_n u_n\right)[u_n]-Q_0'\left(\alpha_0+s_n u_n\right)[u_n] \end{array}\right\}+\|u_n\|^2-\left\{-Q_0''\left(\alpha_0+\widetilde{t}_n u_n\right)[u_n]\right\}.$$

Then take absolute values of each side:

$$\begin{aligned}
\|u_n\|^2 &\leq 2\left\{\begin{array}{c} |O_p(\varepsilon_n^2)| \\ +\left|Q_n'\left(\widetilde{\phi}_n\right)[\widetilde{u}_n]-Q_n'\left(\widetilde{\phi}_n\right)[u_n]\right| \\ +\left|Q_n'\left(\widetilde{\phi}_n\right)[u_n]-Q_n'(\alpha_0+s_nu_n)[u_n]\right| \\ +\left|Q_n'(\alpha_0+s_nu_n)[u_n]-Q_0'(\alpha_0+s_nu_n)[u_n]\right| \\ \frac{1}{2}\left|\|u_n\|^2-\left\{-Q_0''\left(\alpha_0+\widetilde{t}_nu_n\right)[u_n]\right\}\right| \end{array}\right\}, \\
&\leq 2\left\{\begin{array}{c} |O_p(\varepsilon_n^2)| \\ +\left|Q_n'\left(\widetilde{\phi}_n\right)[\widetilde{u}_n]-Q_n'\left(\widetilde{\phi}_n\right)[u_n]\right| \\ +\left|Q_n'\left(\widetilde{\phi}_n\right)[u_n]-Q_n'(\alpha_0+s_nu_n)[u_n]\right| \\ +\left|Q_n'(\alpha_0+s_nu_n)[u_n]-Q_0'(\alpha_0+s_nu_n)[u_n]\right| \\ \frac{1}{2}C_2\|u_n\|^{2+\varrho} \end{array}\right\}. \qquad (50)
\end{aligned}$$

(50) holds with probability approaching one by Corollary 26. Rearranging this gives:

$$\|u_n\| \leq 2\left\{\begin{array}{c} \|u_n\|^{-1}|O_p(\varepsilon_n^2)| \\ +\|u_n\|^{-1}\left|Q_n'\left(\widetilde{\phi}_n\right)[\widetilde{u}_n]-Q_n'\left(\widetilde{\phi}_n\right)[u_n]\right| \\ +\|u_n\|^{-1}\left|Q_n'\left(\widetilde{\phi}_n\right)[u_n]-Q_n'(\alpha_0+s_nu_n)[u_n]\right| \\ +\|u_n\|^{-1}\left|Q_n'(\alpha_0+s_nu_n)[u_n]-Q_0'(\alpha_0+s_nu_n)[u_n]\right| \\ \frac{1}{2}C_2\|u_n\|^{1+\varrho} \end{array}\right\}.$$

We divide the situation into two cases: either $\|u_n\| < n^{-\delta}$ or $\|u_n\| \geq n^{-\delta}$. In the first case, there is no need for additional arguments to bound the rate of convergence. In the second case there is something to prove. Notice that,

$$\|u_n\| - \frac{1}{2}C_2\|u_n\|^{1+\varrho} = \|u_n\|\left(1-\frac{1}{2}C_2\|u_n\|^{\varrho}\right).$$

With probability approaching one, the following holds:

$$\left(1-\frac{1}{2}C_2\|u_n\|^{\varrho}\right) \geq \left(1-\frac{1}{2}C_2\epsilon_0^{\varrho}\right).$$

Define $C_3 = \left(1 - \frac{1}{2}C_2\epsilon_0^{\varrho}\right)$. With probability approaching one, we have:

$$\|u_n\| \leq \frac{2}{C_3}\left\{\begin{array}{c} \|u_n\|^{-1}\left|O_p\left(\varepsilon_n^2\right)\right| \\ +\|u_n\|^{-1}\left|Q_n'\left(\widetilde{\phi}_n\right)[\widetilde{u}_n] - Q_n'\left(\widetilde{\phi}_n\right)[u_n]\right| \\ +\|u_n\|^{-1}\left|Q_n'\left(\widetilde{\phi}_n\right)[u_n] - Q_n'\left(\alpha_0 + s_n u_n\right)[u_n]\right| \\ +\|u_n\|^{-1}\left|Q_n'\left(\alpha_0 + s_n u_n\right)[u_n] - Q_0'\left(\alpha_0 + s_n u_n\right)[u_n]\right| \end{array}\right\}.$$

Hence, either $\|u_n\| < n^{-\delta}$ or

$$\|u_n\| \leq \frac{2}{C_3}\left\{\begin{array}{c} n^{\delta}\left|O_p\left(\varepsilon_n^2\right)\right| \\ +n^{\delta}\left|Q_n'\left(\widetilde{\phi}_n\right)[\widetilde{u}_n] - Q_n'\left(\widetilde{\phi}_n\right)[u_n]\right| \\ +n^{\delta}\left|Q_n'\left(\widetilde{\phi}_n\right)[u_n] - Q_n'\left(\alpha_0 + s_n u_n\right)[u_n]\right| \\ +\|u_n\|^{-1}\left|Q_n'\left(\alpha_0 + s_n u_n\right)[u_n] - Q_0'\left(\alpha_0 + s_n u_n\right)[u_n]\right| \end{array}\right\}. \tag{51}$$

Lemma 30 gives us the convergence rates for the second and third terms in (51). These are $O_p\left(n^{-1/2-\varkappa_1/2}\right) + O\left(g_n\right)$ and $O\left(g_n\right)$ respectively. Finally, we must deal with the last term. This can be bounded by:

$$\sup_{\alpha\in\Theta_n, n^{-\delta}\leq\|\alpha-\alpha_0\|\leq\epsilon_0}\left|\|\alpha-\alpha_0\|^{-1}\left\{\begin{array}{c} Q_n'\left(\alpha_0 + s_n\left(\alpha-\alpha_0\right)\right)[\alpha-\alpha_0] \\ -Q_0'\left(\alpha_0 + s_n\left(\alpha-\alpha_0\right)\right)[\alpha-\alpha_0] \end{array}\right\}\right|.$$

The rate of convergence for this term was derived in Lemma 29. The rates of these four elements give the rates for the case $\|u_n\| > n^{-\delta}$. Adding in the other case $\|u_n\| \leq n^{-\delta}$ gives the final rate. We have $\|\widehat{\alpha} - \alpha_0\| = O_p\left(\max\left\{n^{-\delta}, n^{\delta}\varepsilon_n^2, n^{\delta}g_n, n^{\delta-1/2-\varkappa_1/2}, n^{-1/2+\delta-\kappa}\right\}\right)$ as claimed. The above proof can be repeated for each j. ■

References

[1] Ai, C. & X. Chen (2003) Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions. *Econometrica* 71, 1795-1843.

[2] Ai, C. & X. Chen (2007) Estimation of Possibly Misspecified Semiparametric Conditional Moment Restriction Models with Different Conditioning Variables. *Journal of Econometrics* 141, 5-43.

[3] Alexander, K.S. (1984) Probability Inequalities for Empirical Processes and a Law of the Iterated Logarithm. *Annals of Probability* 12, 1041-1067.

[4] Andersen, P.K., Ø. Borgan, R.D. Gill & N. Keiding (1993) *Statistical Models Based on Counting Processes*. Springer.

[5] Bielecki, T.R. & M. Rutkowski (2004) *Credit Risk: Modeling, Valuation and Hedging*. Springer.

[6] Black, F. & M. Scholes (1973) The Pricing of Options and Corporate Liabilities. *Journal of Political Economy* 81, 637-654.

[7] Brémaud, P. (1981) *Point Processes and Queues: Martingale Dynamics*. Springer-Verlag.

[8] Chen, X. (2007) Large Sample Sieve Estimation of Semi-Nonparametric Models, Chapter 76 in Handbook of Econometrics, Vol. 6B, 2007, eds. J.J. Heckman and E.E. Leamer, North-Holland.

[9] Chen, X. & Z. Liao (2014) Sieve M Inference on Irregular Parameters. *Journal of Econometrics* 182, 70-86.

[10] Chen, X., Z. Liao & Y. Sun (2014) Sieve Inference on Possibly Misspecified Semi-Nonparametric Time Series Models. *Journal of Econometrics* 178, 639-658.

[11] Chen, X. & D. Pouzo (2009) Efficient Estimation of Semiparametric Conditional Moment Models with Possibly Nonsmooth Residuals. *Journal of Econometrics* 152, 46-60.

[12] Chen, X. & D. Pouzo (2012) Estimation of Nonparametric Conditional Moment Models with Possibly Nonsmooth Generalized Residuals. *Econometrica* 80, 277-321.

[13] Chen, X. & X. Shen (1998) Sieve Extremum Estimates for Weakly Dependent Data. *Econometrica* 66, 289-314.

[14] Chen, X., E. Tamer & A. Torgovitsky (2011) Sensitivity Analysis in Semiparametric Likelihood Models. Cowles Foundation working paper.

[15] Chui, C.K. (1992) *An Introduction to Wavelets.* Academic Press.

[16] Collin-Dufresne, P., R. Goldstein & J. Hugonnier (2004) A General Formula for Valuing Defaultable Securities. *Econometrica* 72 (5), 1377-1407.

[17] Conway, J.B. (1990) *A Course in Functional Analysis.* Springer.

[18] Duffie, D., L. Saita & K. Wang (2007) Multi-Period Corporate Default Prediction With Stochastic Covariates. *Journal of Financial Economics* 83, 635-665.

[19] Duffie, D., M. Schroder & C. Skiadas (1996) Recursive Valuation of Defaultable Securities and the Timing of Resolution of Uncertainty. *Annals of Applied Probability* 6, 1075-1090.

[20] Duffie, D. & K.L. Singleton (1999) Modeling Term Structures of Defaultable Bonds. *Review of financial Studies* 12, 687-720.

[21] Duffie, D. & K.L. Singleton (2003) *Credit Risk: Pricing, Measurement and Management.* Princeton.

[22] Fleming, T.R. & D.P. Harrington (1991) *Counting Processes and Survival Analysis.* Wiley.

[23] Folland, G.B. (1999) *Real Analysis: Modern Techniques and Their Applications.* Wiley-Interscience.

[24] Hall, P. (1992) *The Bootstrap and Edgeworth Expansion.* Springer-Verlag.

[25] Karr, A.F. (1987) Maximum Likelihood Estimation in the Multiplicative Intensity Model Via Sieves. *Annals of Statistics* 15 (2), 473-490.

[26] Kolmogorov, A.N. & V.M. Tikhomirov (1959) ϵ-entropy and ϵ-capacity of Sets in a Functional Space. [English Translation, *American Mathematical Society Translations* 17 (2), 277-364 (1961).]

[27] Linton, O.B., J.P. Nielsen & S. van de Geer (2003) Estimating Multiplicative and Additive Hazard Functions by Kernel Methods. *Annals of Statistics* 31 (2), 464-492.

[28] Martinussen, T. & T.H. Scheike (2010) *Dynamic Regression Models for Survival Data.* Springer.

[29] Merton, R. (1973) The Theory of Rational Option Pricing. *Bell Journal of Economics and Management Science* 4, 141-183.

[30] Nielsen, J.P. & O.B. Linton (1995) Kernel Estimation in a Nonparametric Marker Dependent Hazard Model. *Annals of Statistics* 23 (5), 1735-1748.

[31] Nielsen, J.P., O.B. Linton & P.J. Bickel (1998) On a Semiparametric Survival Model with Flexible Covariate Effect. *Annals of Statistics* 26 (1), 215-241.

[32] Pollard, D. (1984) *Convergence of Stochastic Processes.* Springer-Verlag.

[33] Protter, P.E. (2005) *Stochastic Integration and Differential Equations.* Springer.

[34] Shen, X. (1997) On Methods of Sieves and Penalization. *Annals of Statistics* 25, 2555-2591.

[35] Shen, X. & W. Wong (1994) Convergence Rate of Sieve Estimates. *Annals of Statistics* 22, 580-615.

[36] Singleton, K.J. (2006) *Empirical Dynamic Asset Pricing: Model Specification and Econometric Assessment.* Princeton University Press.

[37] Van den Berg, G.J., L. Janys, E. Mammen, J.P. Nielsen (2014) A General Semiparametric Approach to Inference with Marker-Dependent Hazard Rate Models. IZA working paper.

[38] Van der Vaart, A.W. (1998) *Asymptotic Statistics.* Cambridge.

[39] Van der Vaart, A.W. & J.A. Wellner (1996) *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer.

[40] Wolter, J.L. (2015) Separating the Impact of Macroeconomic Variables and Global Frailty in Event Data. Manuscript.

[41] Wong, W.H. & T. Severini (1991) On Maximum Likelihood Estimation in Infinite Dimensional Parameter Spaces. *Annals of Statistics* 19, 603-632.