Maximum approximate Bernstein likelihood estimation in proportional hazard model for interval-censored data

Zhong Guan

Department of Mathematical Sciences, Indiana University South Bend, South Bend, Indiana,

Correspondence
Zhong Guan, Department of Mathematical Sciences, Indiana University South Bend, South Bend, IN, USA
Email: zguan@iusb.edu

Maximum approximate Bernstein likelihood estimates of the baseline density function and the regression coefficients in the proportional hazard regression models based on interval-censored event time data result in smooth estimates of the survival functions which enjoys an almost $n^{1/2}$-rate of convergence faster than the $n^{1/3}$-rate for the existing estimates. The proposed method was shown by a simulation to have better finite sample performance than its main competitors. Some examples including real data are used to illustrate the usage of the proposed method.

KEYWORDS
approximate likelihood, Bernstein polynomial model, Cox's proportional hazard regression model, density estimation, interval censoring, mixture beta model, survival curve

1 | INTRODUCTION

Traditionally in semi- and non-parametric statistics, we approximate an unknown smooth distribution function, an infinite-dimensional parameter, by a step function and parameterize it by the jump sizes of the step function at the observed values. Therefore, the working model is actually a multinomial distribution, a parameter of finite but varying dimension. The resulting estimate is a step function and does not possess a density. This approach is appropriate only for discrete unknown distribution and works fine when the infinite-dimensional parameter is nuisance. However, in the situation when such parameters such as survival, hazard, and density functions are our concerns the traditional approach resulting in a jagged step-function estimate is not satisfactory especially when sample size is small that is usually the case for survival analysis of rare diseases and for reliability estimation of expensive products. Besides the roughness of the estimate it is difficult to parameterize the unknown survival function and not easy to find the nonparametric maximum likelihood estimate when data are incompletely observed due to the complication of assigning probabilities and the large number of parameters (usually the same as the sample size) to be estimated. Moreover, the roughness of the estimate of the nonparametric component could affect the accuracy of the estimates of parameters in semiparametric models. Turnbull\(^1\) presented an expectation-maximization algorithm (see Dempster et al.,\(^2\) also) to compute the discrete nonparametric maximum likelihood estimate of the distribution function from grouped, censored, and truncated data without covariates (see Groeneboom and Wellner,\(^3\) also). The method has been generalized to obtain semiparametric maximum likelihood estimate of the survival function to models including Cox’s proportional hazards model.\(^4,7\) Finkelstein and Wolfe\(^8\) proposed some semiparametric models for interval censored data. Asymptotic results about some semiparametric models can be found in Huang and Wellner\(^6\) and Schick and Yu,\(^9\) and so on. With interval censored data, the assignment of the probabilities within the Turnbull interval cannot be uniquely determined.\(^10,11\) Groeneboom and Wellner\(^3\) suggested an iterative convex minorant algorithm, which was improved or generalized by Wellner and Zhan,\(^12\) Pan,\(^7\) and Anderson-Bergman.\(^13\) Grouped failure time data have been studied by, among others, Prentice and Gloeckler\(^14\) and Pierce et al.\(^15\) Unfortunately, the non- or semi-parametric maximum likelihood estimate of the survival function based on interval censored...
data is a step-function and only has $n^{1/3}$-rate of convergence.\textsuperscript{6} Parametric models and kernel smoothing methods\textsuperscript{16,17} have been applied to obtain smooth estimator of survival function.\textsuperscript{18-20} Another continuous estimation was due to Becker and Melbye\textsuperscript{21} who assumed piecewise constant intensity model. Carstensen\textsuperscript{22} generalized this method to regression models by assuming piecewise constant baseline rate. The most recent and the most relevant work to this article is Wang et al\textsuperscript{23} in which the cumulative baseline hazard function is estimated by a monotone splines with questionable knot selection.

Goetghebeur and Ryan\textsuperscript{24} indicated that many of the expectation-maximization-like methods have the relatively ad hoc nature of the procedure used to impute missing data and proposed a method using approximate likelihood to avoid such problem that retains some of the appealing features of the nonparametric smoothing methods such as the regression spline smoothing of Kooperberg and Clarkson\textsuperscript{25} and the local likelihood kernel smoothing of Betensky et al.\textsuperscript{20}

Nonparametric density estimation is rather difficult due to the lack of information contained in sample data about it.\textsuperscript{26,27} Kernel method is usually unsatisfactory when sample size is small even for complete data. Some authors have studied the estimation of density function based on censored data without covariate (see, eg, Braun et al,\textsuperscript{28} Harlass,\textsuperscript{29} and the references therein).

A useful working statistical model must be finite-dimensional and approximates the true underlying distribution (see p. 1 of Bickel et al\textsuperscript{26}). Instead of approximating the underlying continuous distribution function by a step-function, Guan\textsuperscript{30} suggested a Bernstein polynomial approximation\textsuperscript{31,32} that is actually a mixture of some specific beta distributions. This Bernstein polynomial model performs much better than the classical kernel method for estimating density even from grouped data\textsuperscript{33} and data with measurement errors.\textsuperscript{34} The maximum approximate Bernstein likelihood estimate can be viewed as a continuous version of the non- or semi-parametric maximum likelihood estimate. In this article, such estimates of the conditional survival function and density function given covariate are proposed by fitting interval censored data with Cox's proportional hazards model. Bernstein polynomials are used by Hothorn et al\textsuperscript{15} to parameterize the log-cumulative baseline hazard function rather than the density function in the Cox model.

The rest of this article is organized as follows. Section 2 describes the proposed method and the algorithm to find maximum likelihood estimates. Some asymptotic results are given in Section 3. Simulation results and applications to some real datasets are presented in Section 4. Section 5 concludes the article with some discussions. Proofs of the asymptotic results are relegated in the Appendix.

2 | METHODOLOGY

Let $T$ be an event time and $X$ be an associated $d$-dimensional covariate with distribution $H(x)$ on $\mathcal{X}$. We denote the marginal and the conditional survival functions of $T$, respectively, by $S(t) = \overline{F}(t) = 1 - F(t) = Pr(T > t)$ and $S(t|x) = \overline{F}(t|x) = 1 - F(t|x) = Pr(T > t|X = x)$. Let $f(t|x)$ denote the conditional density of a continuous $T$ given $X = x$. The conditional cumulative hazard function and odds ratio are, respectively, $\Lambda(t|x) = -\log S(t|x)$ and $O(y|x) = S(y|x)/[1 - S(y|x)]$. Consider the Cox's proportional hazard regression model\textsuperscript{36}

$$S(t|x) = S(t|x; y, f_0) = S_0(t)^{\exp(\gamma^\top x)},$$

where $\gamma \in \mathbb{G} \subset \mathbb{R}^d$, $\mathbf{x} = \mathbf{x} - \mathbf{x}_0$, $\mathbf{x}_0$ is any fixed covariate value, $f_0(\cdot) = f(\cdot|x_0)$ is the unknown baseline density and $S_0(t) = \int_{t}^{\infty} f_0(s)ds$. This is equivalent to

$$f(t|x) = f(t|x; y, f_0) = \exp(\gamma^\top x)S_0(t)^{\exp(\gamma^\top x)-1}f_0(t).$$

It is clear that Equations (1) and (2) are also true if we change the “baseline” covariate $\mathbf{x}_0$ to any $\mathbf{x}_0 \in \mathcal{X}$ with the same $\gamma$ but $\mathbf{x}$ being replaced by $\mathbf{x}' = \mathbf{x} - \mathbf{x}_0$. For a given $\gamma \in \mathbb{G}$, define a $\gamma$-related “baseline” as an $\mathbf{x}_\gamma \in \arg \min_{\mathbf{x} \in \mathcal{X}} \gamma^\top \mathbf{x}$ and denote $\mathbf{x}_\gamma = \mathbf{x} - \mathbf{x}_\gamma$. Define $r = \inf \{t : F(t|\mathbf{x}_0) = 1\}$. It is true that $r$ is independent of $\mathbf{x}_0$, $0 < r \leq \infty$, and $f(t|x)$ have the same support $[0, r]$ for all $x \in \mathcal{X}$. It is obvious that for any strictly increasing continuous function $\psi$, $Pr[\psi(T) > t|x] = Pr[\psi(T) > t|x_\gamma]|_{x=x_\gamma}$. Thus the transformed event time $\psi(T)$ also satisfies the Cox model (1).

We will consider the general situation where the event time is subject to interval censoring. The observed data are $Z = (\Delta, X, Y)$, where $Y = (Y_1, Y_2]$ and $\Delta$ is the censoring indicator, that is, $T = Y = Y_1 = Y_2$ is uncensored if $\Delta = 0$ and $T \in Y = (Y_1, Y_2], 0 \leq Y_1 < Y_2 \leq \infty$, is interval censored if $\Delta = 1$. The reader is referred to Huang and Wellner\textsuperscript{6} for a review and more references about interval censoring. The right-censoring $Y_2 = \infty$ and left-censoring $Y_1 = 0$ are included as special cases. For any individual observation $z = (\delta, x, y)$, where if $\delta = 0$ then $y = y = t$ else if $\delta = 1$ then $y = (y_1, y_2] \ni t$, $0 \leq y_1 < y_2 \leq \infty$,
let the full loglikelihood, up to an additive term independent of \((\gamma, f_0)\), be

\[
\ell^\prime(\gamma, f_0; z) = (1 - \delta)[\gamma^T \bar{x} + \log f_0(y) + (\alpha^T \bar{x} - 1) \log S_0(y)] + \delta \log[S_0(y_1)^{\alpha^T \bar{x}} - S_0(y_2)^{\alpha^T \bar{x}}].
\] (3)

Let \((\delta_i, x_i, y_i), i = 1, \ldots, n\), be independent observations of \((\Delta, X, Y)\). If \(\tau\) is either unknown or \(\tau = \infty\) and \(\tau_\nu\) is at least the last finite observed time, that is, \(\tau_\nu \geq y_{\nu j} = \max\{y_{\nu 1}, y_{\nu 2}: y_{\nu 2} < \infty; i, j = 1, \ldots, n\}\) then \([\tau_{\nu n}, \infty)\) is contained in the last Turnbull interval.\(^1\) It is well known that if the last event time was right censored then the distribution of \(T\) is not “nonparametrically estimable” on \([\tau_{\nu n}, \infty)\). Thus all finite observed times are in \([0, \tau_{\nu n})\) only on which can \(f(t|x)\) be estimated.

The full likelihood (3) cannot be maximized unless \(S(t|x_0)\) is specified by a finite dimensional model. Traditional method approximates \(S(t|x_0)\) by step-function and treats the jumps at observations as unknown parameters. Let \(\dot{S}_n\) denote such a non- or semi-parametric maximum likelihood estimate of \(S\) based on interval-censored event time data and \(\dot{y}\) denote the associated maximum likelihood estimate of \(y\) when covariates present.\(^6\)

However the Bernstein polynomial approximation makes the parametrization simple and much easy.\(^{30,33}\) Given any \(x_0\), denote \(\pi = \pi(x_0) = 1 - S_0(\tau_{\nu n})\). For integer \(m \geq 1\) we define \(S_{m+1} = \{(u_0, \ldots, u_m)^T \in \mathbb{R}^{m+1}: u_i \geq 0, \sum_{i=0}^m u_i = 1\}\). We can approximate \(f_\nu(t)\) on \([0, \tau_{\nu n}]\) by \(f_\nu(t; \nu, p) = r_{\nu n}^{-1} \sum_{i=0}^m p_i \beta_m(t/\tau_{\nu n})\), where \(\beta_m(u) = (m+1)^m u^m\) is the density of the beta distribution with shapes \((i+1, m+i+1)\). \(p = \pi(x_0) = (p_0, \ldots, p_{m+1})\) are subject to constraints \(p \in S_{m+1}\) and \(p_{m+1} = 1 - \pi\). Here the dependence of \(\pi\) and \(p\) on \(x_0\) will be suppressed. If \(\pi < 1\), although we cannot estimate the values of \(f_\nu(t)\) on \((\tau_{\nu n}, \infty)\), we can put an arbitrary guess on them such as \(f_\nu(t; \nu, p) = p_{m+1} \alpha(t - \tau_{\nu n}), t \in (\tau_{\nu n}, \infty)\), where \(\alpha(\cdot)\) is a density on \([0, \infty)\) satisfying \((1 - \pi)\alpha(0) = (m+1)p_{m+1}/\tau_{\nu n}\). Thus \(S_\nu(t)\) on \([0, \infty)\), can be “estimated” by

\[
S_{\nu m}(t; p) = I(t \leq \tau_{\nu n}) \sum_{i=0}^m p_i [\bar{S}_\nu(t/\tau_{\nu n}) + I(t > \tau_{\nu n})p_{m+1}\bar{A}(t - \tau_{\nu n})].
\] (4)

where \(\bar{S}_\nu(t) = 1 - \nu(t) = \int_0^t \bar{\beta}_\nu(s)ds\), \(i = 0, \ldots, m\), \(\bar{S}_{m+1}(t) \equiv 1, \bar{A}(t) = \int_t^\infty \alpha(u)du\). Thus we can approximate \(S(t|x)\) and \(f(t|x)\) on \([0, \tau_{\nu n}]\) by \(S_{\nu m}(t|x; \pi, p) = S(t|x; \nu, f_\nu(m; \nu, p))\) and \(f_{\nu m}(t|x; \nu, p) = f(t|x; \nu, f_\nu(m; \nu, p))\), respectively.

If \(\tau\) is finite and known we choose \(\tau_n = \tau\) and specify \(p_{m+1} = 0\). Generally \(\tau\) is either infinite or unknown. Then, in case \(y_{\nu j}\) is not censored \(\tau_{\nu n}\) is chosen to be a little bigger than \(y_{\nu j}\) but is believed not larger than \(\tau\), otherwise, \(\tau_{\nu n} = \nu(y_{\nu j})\). In this case, from Equation (3), we see that for data without right-censoring or covariate we also have to specify \(p_{m+1} = 0\) due to its unidentifiability. If \(\tau_{\nu n} \neq 1\) we divide all the observed times by \(\tau_{\nu n}\). Thus we assume \(\tau_n = 1\) in the following. We define \(m^* = m\) or \(m + 1\) according to whether we specify \(p_{m+1} = 0\) or not. Thus \(p = (p_0, \ldots, p_{m^*})\) satisfies constraints

\[
p = \pi(x_0) = (p_0, \ldots, p_{m^*}) \in S_{m^*}. \quad 0 \leq p_{m+1} < 1.
\] (5)

Then \(\ell(\gamma, f_0) = \sum_{i=1}^n \ell(\gamma, f_0; z_i)\) can be approximated by \(\ell_{\nu m}(\gamma, p) = \sum_{i=1}^n \ell_{\nu m}(\gamma, p; z_i)\), where \(\ell_{\nu m}(\gamma, p; z) = \ell(\gamma, f_\nu(m; \nu, p) z)\). For a given degree \(m\), if \((\gamma, \hat{p})\) maximizes \(\ell_{\nu m}(\gamma, p)\) subject to constraints in Equation (5) for some \(x_0\) then \((\gamma, \hat{p})\) is called the maximum approximate Bernstein (or beta) likelihood estimator of \((\gamma, p)\). This is a full likelihood method. The maximum approximate Bernstein (or beta) likelihood estimators of \(f(t|x)\) and \(S(t|x)\) are, respectively, \(\hat{f}_{\nu}(t|x) = f_{\nu m}(t|x; \gamma, \hat{p})\) and \(\hat{S}_{\nu}(t|x) = S_{\nu m}(t|x; \gamma, \hat{p})\).

For a given \(\gamma\), let \(\hat{\pi}(\gamma)\) denote the maximizer of \(\ell_{\nu m}(\gamma, p)\) with respect to \(p = (p_0, \ldots, p_{m^*})^T\) subject to constraints in Equation (5). The derivative of \(\ell_{\nu m}(\gamma, p; z)\) with respect to \(p\) is \(\partial \ell_{\nu m}(\gamma, p; z)/\partial p = \Psi(\gamma, p; z) = [\Psi_0(\gamma, p; z), \ldots, \Psi_{m^*}(\gamma, p; z)]^T\), where

\[
\Psi_j(\gamma, p; z) = \frac{\partial \ell_{\nu m}(\gamma, p; z)}{\partial p_j}, \quad j = 0, \ldots, m^*.
\] (6)

About the second derivatives we have the following lemma.

**Lemma 1.** The Hessian matrix \(H(\gamma, p) = \partial^2 \ell_{\nu m}(\gamma, p)/\partial p \partial p^T\) is nonpositive, that is, all entries are nonpositive. For any fixed \(\gamma\) if \(\gamma^T x_0 \leq \min_{1 \leq i \leq n} \gamma^T x_i\) then \(H(\gamma, p)\) is negative semi-definite for each \(p \in S_{m^*}\). If, in addition, the vectors \([\Psi_0(\gamma, p; z), \ldots, \Psi_{m^*}(\gamma, p; z)]^T, j = 0, \ldots, m^*\), are linearly independent, then \(H(\gamma, p)\) is negative definite.

Similar to Peters and Walker\(^3\) we have the following result.
Theorem 1. For any fixed $\gamma$ if $\gamma^\top x_0 \leq \min_{1 \leq i \leq n} \{\gamma^\top x_i\}$ then $\hat{p}(\gamma)$ is a maximizer of $\ell_m(\gamma, p)$ if and only if

$$\lambda_n(\gamma) = \sum_{i=1}^{n} e_i^\top x_i \geq \sum_{i=1}^{n} \Psi_j(\gamma, \hat{p}(\gamma); z_i) \tag{7}$$

for all $j = 0, \ldots, m^*$ with equality if $\hat{p}_j(\gamma) > 0$. If, in addition, the vectors $[\Psi_j(\gamma, p; z_1), \ldots, \Psi_j(\gamma, p; z_n)]^\top$, $j = 0, \ldots, m^*$, are linearly independent for all $p$ in the interior of $S_m$, then $\hat{p}(\gamma)$ is unique.

Therefore it is necessary that $\hat{p}_j(\gamma) = \hat{p}_j(\gamma)\overline{\Psi}_j(\gamma, \hat{p}(\gamma))$, where $\overline{\Psi}_j(\gamma, p) = [\lambda_n(\gamma)]^{-1} \sum_{i=1}^{n} \Psi_j(\gamma, p; z_i)$, $j = 0, \ldots, m^*$. We have fixed-point iteration

$$p_{j}^{(s+1)} = p_{j}^{(s)} \overline{\Psi}_j(\gamma, p^{(s)}), \quad j = 0, \ldots, m^*; \ s = 0, 1, \ldots . \tag{8}$$

If $\gamma^\top x_0 \leq \min_{1 \leq i \leq n} \{\gamma^\top x_i\}$ then $\overline{\Psi}_j(\gamma, p) \geq 0$ for all $j = 0, \ldots, m^*$ and $p \in \mathbb{S}_m$. Similar to the proof of theorem 4 of Peters and Walker,37 we can prove the convergence of $p^{(s)}$.

Theorem 2. Suppose for any fixed $\gamma$ that $\gamma^\top x_0 \leq \min_{1 \leq i \leq n} \{\gamma^\top x_i\}$ and the vectors $[\Psi_j(\gamma, p; z_1), \ldots, \Psi_j(\gamma, p; z_n)]^\top$, $j = 0, \ldots, m^*$, are linearly independent. If $p^{(0)}$ is in the interior of $S_m$, then the sequence $\{p^{(s)}\}$ of Equation (8) converges to $\hat{p}(\gamma)$.

For a fixed $p$ we have the following concavity of $\ell_m(\gamma, p)$ as a function of $\gamma$.

Lemma 2. Suppose the vectors $(\hat{x}_1, \ldots, \hat{x}_d)$, $d = 1, \ldots, d$, are linearly independent and $0 < y_i < 1$ if $\delta_i = 0$. For any fixed $p \in \mathbb{S}_m$, the matrix $\partial^2 \ell_m(\gamma, p)/(\partial \gamma \partial \gamma^\top)$ is negative definite.

Define an empirical $\gamma$-related “baseline” $\hat{x}_0 = \hat{x}_0(\gamma)$ such that $\gamma^\top \hat{x}_0 = \min_{1 \leq i \leq n} \{\gamma^\top x_i\}$. For a fixed $\gamma$ we choose $x_0 = \hat{x}_0(\gamma)$ to obtain $\hat{p} = \hat{p}(\gamma)$. Therefore we can also estimate $f(t(x))$ and $S(t(x))$, respectively, by $\hat{f}_m(t(x)) = f_m(t(x; \gamma, \hat{p})$ and $\hat{S}_m(t(x)) = S_m(t(x; \gamma, \hat{p})$ for dataset without covariate, we have $\gamma = 0$. Then we have $\hat{f}_m(t) = f_m(t; 0, \hat{p})$ and $\hat{S}_m(t) = S_m(t; 0, \hat{p})$. Because $\gamma$ is consistent, $(\hat{\gamma}, \hat{p})$ are “close” to $(\hat{\gamma}, \hat{p})$. Thus $(\hat{\gamma}, \hat{p})$ can be used as initial values for finding $(\hat{\gamma}, \hat{p})$ by the following algorithm. Such procedure was also suggested by Huang.5

Algorithm for finding $(\hat{\gamma}, \hat{p})$:

Step 0. Start with an initial guess $\gamma^{(0)}$ of $\gamma$. Choose $x_0^{(0)} = \hat{x}_0(\gamma^{(0)})$. Use Equation (8) with $\gamma = \gamma^{(0)}, x_0 = x_0^{(0)}$, and starting point $p^{(0)}$ to get $p^{(0)} = \hat{p}(\gamma)$. Set $s = 0$.

Step 1. Find the maximizer $\gamma^{(s+1)}$ of $\ell_m(\gamma, p^{(s)})$ using the Newton-Raphson method.

Step 2. Choose $x_0^{(s+1)} = \hat{x}_0(\gamma^{(s+1)})$, $\gamma = \gamma^{(s+1)}$, and a proper $p^{(0)}$. Then use Equation (8) with $x_0 = x_0^{(s+1)}$ to get $p^{(s+1)} = \hat{p}(\gamma)$. Set $s = s + 1$.

Step 3. Repeat Steps 1 and 2 until convergence. The final $\hat{\gamma}$ and $p^{(s)}$ are taken as $(\hat{\gamma}, \hat{p})$ with baseline $\hat{x}_0 = x_0^{(s)}$.

The concavities of $\ell_m(\gamma, p)$ with respect to $\gamma$ and $p$ ensure that the above iterative algorithm is a point-to-point map and the solution set contains single point. Convergence of $(\hat{\gamma}^{(s)}, p^{(s)})$ to $(\hat{\gamma}, \hat{p})$ is guaranteed by the Global Convergence Theorem.38

An optimal model degree $m$ can be selected using the method of Guan.30 For some properly chosen $m_0$ and $k$ let $m_i = m_0 + i, i = 0, \ldots, k$. For each $i$, fit the data to obtain $(\hat{\gamma}, \hat{p})$ and $\ell_i = \ell_m(\gamma, p)$. The optimal degree $m$ is $m = \arg\max_{1 \leq i \leq k} (R(m_i))$, the maximizer $\hat{m}$ of $R(m_i) = k \log(\ell_k - \ell_0)/k - i \log(\ell_i - \ell_0)/i - (k - i) \log(\ell_k - \ell_i)/(k - i), i = 1, \ldots, k$, where $R(m_k) = 0$. Alternatively, we can replace $\ell_i$ by $\ell_m(\gamma, p)$ for all $i$. The resulting optimal degree is denoted by $\hat{m}$. The proposed method has been implemented in R as a component of package mable35 which is available on CRAN.

3 | ASYMPTOTIC RESULTS

The following assumptions will be used to prove some asymptotic results.

Assumption 1. The support $\mathcal{X}$ of covariate $X$ is compact and for each $x_0 \in \mathcal{X}$, $E(XX^\top)$ is positive definite, where $\hat{X} = X - x_0$.
Assumption 2. For each \(x_0 \in \mathcal{X}\) and \(\tau_n > 0\), there exist \(f_m(t; p_0)\) and \(\rho > 0\) such that, uniformly in \(t \in [0, \tau_n]\),

\[
\frac{f_m(t; p_0) - f_0(t)}{f_0(t)} = O(m^{-\rho/2}),
\]

where \(p_0 = (p_{01}, \ldots, p_{0m}, p_{0,m+1})^T\), and \(p_{0,m+1} = 1 - \pi(x_0) = S(r_n | x_0)\).

For any \(y\), the compactness of \(\mathcal{X}\) ensures the existence of \(x_\gamma \in \arg \min \{ \gamma^T x : x \in \mathcal{X}\} \). Boundedness of \(\mathcal{X}\) is assumed in the literature, for example, (A3)(b) of Huang and Wellner. The positive definiteness of \(E(\tilde{X}X^T)\) assures the identifiability of \(\gamma\).

Lemma 3. Suppose that \(\phi(t) = t^\rho (1 - t)^\alpha \phi_0(t)\) is a density on \([0, 1]\). A function \(f\) is said to be \(\alpha-Hölder\) continuous with \(\alpha \in (0, 1]\) if \(|f(x) - f(y)| \leq C|x - y|^\alpha\) for some constant \(C > 0\). We have the following result.

**Lemma 3.** Suppose that \(\phi(t) = t^\rho (1 - t)^\alpha \phi_0(t)\) is a density on \([0, 1]\), \(a\) and \(b\) are nonnegative integers, \(\phi_0 \in C^{(r)}[0, 1], r \geq 0, \phi_0(t) \geq b_0 > 0\), and \(\phi_0^{(r)}\) is \(\alpha-Hölder\) continuous with \(\alpha \in (0, 1]\). Then there exists \(p_0 \in S_m\) such that uniformly in \(t \in [0, 1]\), with \(\rho = r + \alpha\),

\[
\frac{f_m(t; p_0) - \phi(t)}{\phi(t)} = O(m^{-\rho/2}).
\]

This lemma was proved in Wang and Guan. This is a generalization of the result of Lorentz which requires a positive lower bound for \(\phi\), that is, \(a = b = 0\). If \(\phi(t) = \tau_0 f_0(t)/\pi(x_0)\) as a density on \([0, 1]\) fulfills the condition of Lemma 3, then Assumption 2 is fulfilled. The condition of Lemma 3 seems only sufficient for Assumption 2.

In the following, all expectations \(E(\cdot)\) are taken with respect to the (joint) distribution of random variable(s) in upper case. The following are the conditions for cases considered in the asymptotic results. We reduce the general interval censoring to Case 2 so that the examination times \(U\) and \(V\) are defined as in sections 1.2 and 1.3 of Huang and Wellner.

**Condition 0.** The event time \(T\) is uncensored and \(\tau_n = \tau < \infty\).

**Condition 1.** The event time \(T\) is subject to Case 1 interval censoring. Given \(X = x\) the examination time \(U\) has cdf \(G(\cdot|x)\) on \([\tau_l, \tau_u] = [\tau_n < \tau < \infty\), and \(E(O(U|X)) < \infty\).

**Condition 2.** The event time \(T\) is subject to Case 2 interval censoring. Given \(X = x\) the observed examination times \((U, V)\) have joint cdf \(G(\cdot, \cdot|x)\) on \((u, v) : 0 < \tau_l \leq u < v \leq \tau_u\), \(\tau_n = \tau_u < \tau\), and \(E(O(U|X)S(U|X)) < \infty\).

The conditions about the support of the examination times are similar to those of Huang and Wellner. Under Condition 0, we define statistical distance \(D_{01}^2(\gamma_0; p, x_0) = \chi_0^2(\gamma_0; p, x_0) + D_{01}^2(\gamma_0; p, x_0)\), where \(\gamma_0\) is the true value of \(\gamma\),

\[
\chi_0^2(\gamma_0; x_0) = E\left\{ f_m(T; p) \right\} = \int_0^\tau f_m(y; p)dy,
\]

\[
D_{01}(\gamma; p, x_0) = E\left\{ |e^{\gamma_0}\bar{X} - 1| \right\}.
\]

Under Condition 1, we define distance

\[
D_{11}^2(p; x_0) = E\left\{ \frac{S_m(U; p)}{S_0(U)} - 1 \right\}^2 O(U|X).
\]

Under Condition 2, we define \(D_{21}^2(p; x_0) = \max\{D_{21}^2(p; x_0) : i = 1, 2\}\), where

\[
D_{21}^2(p; x_0) = E\left\{ \frac{S_m(U; p)}{S_0(U)} - 1 \right\}^2 O(U|X)S(U|X),
\]

\[
D_{22}^2(p; x_0) = E\left\{ \frac{S_m(V; p)}{S_0(V)} - 1 \right\}^2 S(V|X).
\]
The same symbols $C$ and $C'$ may represent different constants in different places.

**Theorem 3.** Let $(\hat{\gamma}, \hat{\rho})$ be the maximum approximate Bernstein likelihood estimate of $(\gamma, \rho)$ with degree $m \geq Cn^{1/\rho}$ for some constant $C > 0$. Suppose that Assumptions 1 and 2 are satisfied. For each $i = 0, 1, 2$, and any $\epsilon \in (0, 1/2)$, under Condition $i$, we have $||\hat{\gamma} - \gamma_i||^2 \leq Cn^{-1+\epsilon}$, a.s. and $D^2_I(\hat{\gamma}, \hat{\rho}) \leq Cn^{-1+\epsilon}$, a.s.

**Theorem 4.** Suppose that Assumptions 1 and 2 are satisfied. Let $\hat{\gamma} = \hat{\gamma}_m(\gamma_0)$ be the maximizer of $\ell_m(\gamma, \rho_0)$ for some $\rho_0$ that satisfies Assumption 2. For each $i = 0, 1, 2$, under Condition $i$, $\sqrt{n}(\hat{\gamma} - \gamma_0)$ converges in distribution to $N(0, I^{-1})$ as $n \to \infty$, where $x_0 \in \arg\min_{x \in \mathbb{R}} \gamma_0 x$, $I = E(\tilde{X}X^T)$ under Condition 0; $I = E\left\{[O(U|X)\Lambda^2(U|X)] \tilde{X} \tilde{X}^T\right\}$ under Condition 1; and

$$I = E \left[ \frac{\Lambda^T M A}{S(U|X) - S(V|X)} \tilde{X} \tilde{X}^T \right]$$

$$\geq E \left\{ [O(U|X)S(U|X)\Lambda^2(U|X) + S(V|X)\Lambda^2(V|X)] \tilde{X} \tilde{X}^T \right\},$$

under Condition 2, where $\Lambda = [\Lambda(U|X), \Lambda(V|X)]^T$ and

$$M = S(U|X)S(V|X) \begin{pmatrix} O(U|X) & -1 \\ O(U|X) & 1 \end{pmatrix}.$$

**Remark 1.** The asymptotic variance of Cox’s maximum partial likelihood estimator $\hat{\gamma}$ from uncensored data is $I^{-1}$, where, by the law of total covariance, $I_{\text{cox}} = E[\text{var}(X|T)] \leq E(XX^T) - I$, with equality if and only if $E(X|T = t)$ is constant.

Because $\ell_m(\gamma, \rho)$ depends on $\rho$ through $f_m(\cdot; \rho)$ and $f_m(\cdot; \rho_0) \approx f_m(\cdot; \hat{\rho})$, although $\rho_0$ is unknown, we have $\hat{\gamma} \approx \hat{\gamma}_m$. We can estimate the information $I$ by $\hat{I} = n^{-1}\partial^2 \ell_m(\hat{\gamma}, \hat{\rho})/\partial \gamma \partial \rho^T$, with $x_0 = x_0$.

### 4 SIMULATION AND EXAMPLES

#### 4.1 Simulation

Assume that, given $X = x$, $T$ is Weibull $W(\theta, \sigma e^{-\gamma\gamma/\theta})$ so that the baseline ($x = 0$) distribution is $W(\theta, \sigma)$ with shape and scale $\theta = \sigma = 2$. The function `simIC_weib()` of R package icenReg was used to generate interval censored data of sizes $n = 30, 50, 100$ with censoring probability 70% from Weibull distributions. For data with covariate, $X = (X_1, X_2)$, where $X_1$ and $X_2$ are independent, $X_1$ is uniform(-1,1) and $X_2 = \pm 1$ is uniform, with coefficients $\gamma_1 = 0.5, \gamma_2 = -0.5$. For data without covariate, the R ICC package was used to obtain Braun et al.’s kernel estimator $\hat{S}_K$. The calculation of the proposed estimate was done using the R package `mable`. In each case, 1000 samples were generated and used to estimate $\gamma, f(\cdot|0)$ and $S(\cdot|0)$ on $[0, 7]$. If $r_n = y_{(m)} < 7$ we use exponential $\alpha(\cdot)$ on $(r_n, 7)$ as in Equation (4).

The pointwise mean squared errors of the estimated survival functions are plotted in Figure 1. The proposed estimator is compared with the parametric maximum likelihood estimator $\hat{S}_p$, the non- or semi-parametric maximum likelihood estimator $\hat{S}_E$, and the kernel estimator of Braun et al. (see Figure 1). The overall performance of the $\hat{S}_p$ is close, and getting closer, as sample size increases, to that of $\hat{S}_E$ and much better than that of $\hat{S}_E$ and $\hat{S}_K$. The simulation results on the estimation of the regression coefficients are summarized in Table 1. Because the proposed $\hat{S}_p$ has smaller mean squared errors than $\hat{S}_E$ especially when sample size is not large, the new estimator $\hat{\gamma}$ may have smaller mean squared error than the traditional one. This is convinced by the simulation. From these results we see that the proposed $\hat{\gamma}$ is better than $\hat{\gamma}$ and is close to the parametric maximum likelihood estimator especially for small sample data. The proposed estimates using $m = \hat{m}$ and $m = \tilde{m}$ are similar.

Comparison of the proposed method with a spline-smoothing method of Wang et al. implemented in R package `ICsurv` is also conducted using their models of simulation I and II. Note that model I is not in favor of the proposed method because the baseline density in unbounded at $t = 0$. Model II right-censors a very heavy right tail of the event times. From Table 2, we see that the two methods are similar in terms of estimation of the regression coefficients. However the proposed method outperforms the spline method in estimating the survival function. Unfortunately, `ICsurv` does not provide a density estimation.
FIGURE 1  Simulated pointwise mean squared errors (%). Left panels: \( \hat{S}_{B1} \) with \( m = \hat{m} \), \( \hat{S}_{B2} \) with \( m = \hat{\hat{m}} \), \( \hat{S}_E \) and the parametric maximum likelihood estimator \( \hat{S}_P \) at baseline \( x = 0 \). Right panels: \( \hat{S}_E \), \( \hat{S}_{B1} \) with \( m = \hat{m} \), the kernel estimator \( \hat{S}_K \), and the \( \hat{S}_P \) without covariate [Colour figure can be viewed at wileyonlinelibrary.com]

| \( n \) | \( r_1 \) | \( r_2 \) |
|------|------|------|
| 30   | 0.279 | 0.239 |
| 50   | 0.120 | 0.109 |
| 100  | 0.047 | 0.046 |

| \( n \) | \( r_1 \) | \( r_2 \) |
|------|------|------|
| 30   | 0.104 | 0.088 |
| 50   | 0.048 | 0.044 |
| 100  | 0.018 | 0.017 |

TABLE 1  (i) Mean squared errors of estimates of the regression coefficients using semiparametric method (SP), the proposed method using \( m = \hat{m} \) (B1), the proposed method using \( m = \hat{\hat{m}} \) (B2), and the parametric method (P); (ii) Mean \( \mu_m \) and standard deviation \( \sigma_m \) of \( \hat{m} \)
TABLE 2 Root mean (integrated) squared errors of estimates of the regression coefficients (survival function) using spline method and the proposed method

| Model | n | γ       | \(\hat{y}_1\)     | \(\hat{y}_2\)     | \(\hat{S}(0)\)     |
|-------|---|---------|-------------------|-------------------|-------------------|
| I     | 50 | (1, 1)  | 0.716 (0.518)     | 0.661 (0.534)     | 0.166 (0.164)     |
| II    | 50 | (1, 1)  | 0.733 (0.714)     | 1.381 (1.356)     | 2.655 (2.935)     |
| II    | 50 | (1,-1)  | 0.691 (0.688)     | 1.336 (1.354)     | 2.666 (2.887)     |
| II    | 50 | (-1,1)  | 0.837 (0.809)     | 1.954 (1.962)     | 2.999 (3.059)     |
| II    | 50 | (1, 0)  | 0.821 (0.785)     | 1.585 (1.557)     | 2.507 (2.813)     |

FIGURE 2 Gentleman and Geyer’s example. Left panel: the \(\hat{S}_E(x)\), the \(\hat{S}_H\)'s using initial \(p_1^{(0)}\), and right panel: the \(\hat{S}_H\)'s using initial \(p_1^{(0)}\), where \(m = 6\) and \(i = 1, 2, 3\), \(p_1^{(0)} = (1, 2, \ldots, 7)^T/28\), \(p_2^{(0)} = (1, 1, \ldots, 1)^T/7\), and \(p_3^{(0)} = (1, 2, 3, 4, 3, 2, 1)^T/16\) [Colour figure can be viewed at wileyonlinelibrary.com]

If \(f_o\) such as the Weibull density function has all continuous derivatives then based on Lemma 3 we expect an optimal degree smaller than \(CN^{1/p}\) for any \(p > 0\) and some \(C > 0\). Meanwhile an optimal degree chosen by the method of change-point subject to larger variation for smaller sample size \(n\). In this simulation of 1000 runs, for \(n \in \{30, 50, 100\}\), the sample means of \(\hat{m}\) are, respectively, 12.46, 10.80, and 7.27, and the standard deviations are, respectively, 9.55, 8.96, 7.14. The interested reader is referred to Guan for simulation results about dependence of optimal \(m\) on \(n\).

4.2 An artificial example

The artificial dataset of Gentleman and Geyer shows that the Turnbull’s nonparametric maximum likelihood estimate \(\hat{F}(t)\) may not be unique. The data consist of six intervals \((0, 1), (0, 2), (0, 2), (1, 3), (1, 3), (2, 3)\). Specifying \(p_{m + 1} = 0\) and choosing \(r_n = 3\) we have the transformed intervals \((0, 1/3), (0, 2/3), (0, 2/3), (1/3, 1), (1/3, 1), (2/3, 1)\).

It is easy to see that \(\hat{p}\) is unique and uniform if \(m = 1, 2\) but not unique if \(m \geq 3\). Figure 2 shows the nonparametric maximum likelihood estimate \(\hat{S}_E(t)\) and the \(\hat{S}_H(t)\) and \(\hat{f}_H(t)\) when \(m = 6\) with different starting points \(p_1^{(0)} = (1, 2, \ldots, 7)^T/28\), \(p_2^{(0)} = (1, 1, \ldots, 1)^T/7\), and \(p_3^{(0)} = (1, 2, 3, 4, 3, 2, 1)^T/16\). Although the \(\hat{p}\) is not unique, as shown in Figure 2, the resulting estimated survival functions are almost identical. A kernel density estimate for this dataset was discussed in Braun et al.

4.3 Stanford heart transplant data

To illustrate the use of the proposed method for right-censored data with binary covariate, we used the Stanford heart transplant data contained in R survival package. More information about this dataset can be found in Crowley and Hu. We choose \(X\), the indicator of prior bypass surgery, as covariate and \(r_n = y_{(m)} = 1799\). The Cox’s partial likelihood estimate \(\gamma\) is \(\hat{\gamma} = -0.74072\) (s.e. 0.3591). With fixed \(\gamma = \hat{\gamma}\), we obtained \(\hat{m} = 14\) and \(\hat{p} = (\hat{p}_0, \ldots, \hat{p}_{15})^T\), where \(\hat{p}_0 = 0.470490, \hat{p}_6 = 1.3 \times 10^{-6}, \hat{p}_7 = 0.151148, \hat{p}_8 = 2.8 \times 10^{-5}, \hat{p}_{10} = 1.1 \times 10^{-7}, \hat{p}_{11} = 0.038977, \hat{p}_{15} = 1 - \hat{\alpha} = 0.339359\), and all the other \(\hat{p}_i\)'s are smaller than \(10^{-9}\).
with the chosen $\hat{m} = 14$, the maximizer $(\hat{\gamma}, \hat{p})$ of $\ell_m(\gamma, p)$ was found to be $\hat{\gamma} = -0.95151$ (s.e. 0.12309) and $\hat{p} = (\hat{p}_0, \ldots, \hat{p}_{14})^T$, where $\hat{p}_0 = 0.40848, \hat{p}_2 = 4.5 \times 10^{-6}, \hat{p}_3 = 3.4 \times 10^{-6}, \hat{p}_8 = 1.1 \times 10^{-6}, \hat{p}_9 = 0.14646, \hat{p}_2 = 2.3 \times 10^{-6}, \hat{p}_{10} = 1.3 \times 10^{-6}, \hat{p}_{11} = 0.03827, \hat{p}_{12} = 1.2 \times 10^{-6}, \hat{p}_{15} = 1 - \hat{\alpha} = 0.40677$, and all the other $\hat{p}_i$'s are smaller than $10^{-6}$. The optimal degree is $\hat{m} = 13$ based on full likelihood $\ell_m(\gamma, p)$. We obtained $\hat{\gamma} = -1.00679$ (s.e. 0.12309) and $\hat{p} = (\hat{p}_0, \ldots, \hat{p}_{14})^T$, where $\hat{p}_0 = 0.39837, \hat{p}_6 = 0.03944, \hat{p}_7 = 0.10975, \hat{p}_8 = 1.3 \times 10^{-6}, \hat{p}_9 = 1.3 \times 10^{-6}, \hat{p}_{10} = 0.03011, \hat{p}_{11} = 3.0 \times 10^{-6}, \hat{p}_{14} = 1 - \hat{\alpha} = 0.42232$, and all the other $\hat{p}_i$'s are smaller than $10^{-6}$. The results are shown in Figure 3. The proposed estimates of survival probabilities for those who had (no) by-pass surgery are much larger (a little smaller) than the semiparametric maximum likelihood estimators.

4.4 Ovarian cancer data

The ovarian cancer dataset contained in the R package Survival\textsuperscript{42} was originally reported by Edmonson,\textsuperscript{43} and was used as real data example by several authors (eg, Collett\textsuperscript{44} and Huang and Ghosh\textsuperscript{45}). In this study, $n = 26$ patients with advanced ovarian carcinoma (stages IIIB and IV) were treated using either cyclophosphamide alone (1 g/m\textsuperscript{2}) or cyclophosphamide (500 mg/m\textsuperscript{2}) plus Adriamycin (40 mg/m\textsuperscript{2}) by i.v. injection every 3 weeks in order to compare the treatment effect in prolonging the time of survival. Twelve observations are uncensored and the rest is right-censored. We choose $X = \text{Age}$ so that we have an example of right-censored data with continuous covariate. The Cox's partial likelihood estimate of $\gamma$ is $\hat{\gamma} = 0.16162$ (s.e. 0.04974). Using the proposed method we obtained optimal degree $m = 23$ based on either $\ell_m(\gamma, p)$ or $\ell_m(\hat{\gamma}, \hat{p})$ (see upper panels of Figure 4). With $m = 23$, we have $\hat{\gamma} = 0.17665$ (s.e. 0.01218), and $\hat{\alpha} = 38.89$. The components of $\hat{p}$ are $\hat{p}_2 = 0.00226, \hat{p}_9 = 0.02789, \hat{p}_{10} = 0.00277, \hat{p}_{24} = 0.96707$, and all the other $\hat{p}_i < 10^{-6}$. The estimated survival curves at ages 60 and 65 are shown in Figure 4.

**FIGURE 3** Stanford heart transplant data. Upper left panel: log-likelihood $\ell_m(\gamma, p)$; Upper right panel: likelihood ratio for choosing model degree using change-point estimate. Lower panels: the $\hat{S}_m(\gamma, p)$'s using $\hat{m} = 13$ and $\hat{m} = 14$, and $\hat{S}_m(\gamma, p)$ using $\hat{m} = 14$, with prior surgery $x = 1$ (left) and without prior surgery $x = 0$ (right) [Colour figure can be viewed at wileyonlinelibrary.com]
CONCLUDING REMARKS

The proposed estimator of the survival function has a rate of convergence arbitrarily close to \( n^{1/2} \). This rate is much faster than the \( n^{1/3} \)-rate for the existing estimators. The asymptotic results are also supported by the simulation study. The approximate Bernstein polynomial model may be applied to other semiparametric models to obtain smooth estimate of the underlying distribution function and to improve the estimates of parameters. The traditional discrete non- or semi-parametric maximum likelihood estimate is still useful to obtain initial starting points for the proposed estimates of survival function and the regression coefficients. The only drawback of the proposed method is its computational cost of the EM like algorithm which clearly worth its accuracy and efficiency.

ACKNOWLEDGMENTS

The author was grateful for the reviewer’s suggestions, comments, and corrections that greatly improved the presentation of the manuscript.

ORCID

Zhong Guan © https://orcid.org/0000-0003-4808-7400

REFERENCES

1. Turnbull BW. The empirical distribution function with arbitrarily grouped, censored and truncated data. *J R Stat Soc Ser B Stat Methodol.* 1976;38(3):290-295.
2. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B Stat Methodol.* 1977;39(1):1-38.
3. Groeneboom P, Wellner JA. *Information Bounds and Nonparametric Maximum Likelihood Estimation.* DMV Seminar19. Basel: Birkhäuser Verlag; 1992.
How to cite this article: Guan Z. Maximum approximate Bernstein likelihood estimation in proportional hazard model for interval-censored data. *Statistics in Medicine.* 2021;40:758–778. https://doi.org/10.1002/sim.8801

APPENDIX

Proof of Lemma 1

Let \( p \) be any point in the interior of \( S_m \). For any nonzero vector \( v = (v_0, \ldots, v_{m^*})^T \in \mathbb{R}^{m^*+1} \), define

\[
w(y; v) = \sum_{k=0}^{m} v_k \beta_{mk}(y), \quad W(y; v) = \sum_{k=0}^{m^*} v_k \overline{\beta}_{mk}(y).
\]

By Equation (6), the \((j,k)\)-entry of \( H(y,p) \) is

\[
H_{jk}(z) = \frac{\partial^2 \mathcal{E}_m(y; p, z)}{\partial p_j \partial p_k} = \frac{\partial^2 \mathcal{Q}_m(y; p, z)}{\partial p_j \partial p_k}
\]

\[
= -(1 - \delta) \left[ \frac{f_k(y|x_0; p)}{f_k^2(y|x_0; p)} + \frac{(e^{\tau x} - 1)\overline{\beta}_{mk}(y)\overline{\beta}_{mk}(y)}{S_m^2(y|x_0; p)} \right]
\]

\[
+ \delta (e^{\tau x} - 1) \left\{ \frac{S_m(y_1|x_0; p)e^{\tau x} - 2\overline{\beta}_{mk}(y_1)\overline{\beta}_{mk}(y_2)}{S_m^2(y|x_0; p)e^{\tau x}} - \frac{S_m(y_1|x_0; p)e^{\tau x} - 2\overline{\beta}_{mk}(y_1)\overline{\beta}_{mk}(y_2)}{S_m^2(y|x_0; p)e^{\tau x}} \right\}
\]

\[
= \left\{ \frac{S_m(y_1|x_0; p)e^{\tau x} - 2\overline{\beta}_{mk}(y_1)\overline{\beta}_{mk}(y_2)}{S_m^2(y|x_0; p)e^{\tau x}} - \frac{S_m(y_1|x_0; p)e^{\tau x} - 2\overline{\beta}_{mk}(y_1)\overline{\beta}_{mk}(y_2)}{S_m^2(y|x_0; p)e^{\tau x}} \right\}.
\]

Denote temporarily \( \eta = e^{\tau x}, B_{ij} = \overline{\beta}_{ij}(y), \) and \( V_j = S_m(x_j|0; p) \) \((i = 0, \ldots, m^*; j = 1, 2)\). We know \( V_1 \geq V_2 \) and \( B_{12} \geq B_{22} \). In order to show that \( H_{jk}(z) \leq 0 \) for all \( j,k = 0, \ldots, m^* \), it suffices to show \( A \leq B \), where \( A = (\eta - 1)(V_1^{\eta - 2}B_{12}B_{21} - V_1^{\eta - 2}B_{12}B_{22})(V_1^{\eta - 2} - V_1^{\eta - 2}B_{12}B_{22}) + (V_1^{\eta - 2}B_{12}B_{21} - V_1^{\eta - 2}B_{12}B_{22})(V_1^{\eta - 2} - V_1^{\eta - 2}B_{12}B_{22}) \). This is obvious if \( V_1^{\eta - 2}B_{12}B_{21} - V_1^{\eta - 2}B_{12}B_{22} \leq 0 \). Now assuming \( V_1^{\eta - 2}B_{12}B_{21} - V_1^{\eta - 2}B_{12}B_{22} > 0 \), because \( \eta \geq 1 \), we have \( V_1^{\eta - 2} - V_2^{\eta - 2} \geq \eta V_2^{\eta - 1}(V_1 - V_2) \), and

\[
B - A = \eta(V_1V_2)^{\eta - 2}[V_1^2B_{12}B_{21} + V_2^2B_{12}B_{22} - V_1V_2(B_{12}B_{21} + B_{12}B_{22})]
\]

\[
+ (V_1^{\eta - 2} - V_2^{\eta - 2})(V_1^{\eta - 2}B_{12}B_{21} - V_1^{\eta - 2}B_{12}B_{22})
\]

\[
\geq \eta(V_1V_2)^{\eta - 2}[V_1^2B_{12}B_{21} + V_2^2B_{12}B_{22} - V_1V_2(B_{12}B_{21} + B_{12}B_{22})]
\]

\[
+ \eta V_2^{\eta - 1}(V_1 - V_2)(V_1^{\eta - 2}B_{12}B_{21} - V_1^{\eta - 2}B_{12}B_{22})
\]
For any $v \in \mathbb{R}^{m+1}$, denoting $W_i = W(y_i; \psi) (i = 1, 2)$, we have $v^T H(\psi; \psi) v = \sum_{i=1}^n v^T H(\psi; \psi; z_i) v$, where, shown by simple algebra,

$$v^T H(\psi; \psi; z_i) v = -(1 - \delta) \left[ \frac{w^2(y; v)}{f_m^2(y|x_0; \psi)} + \frac{(\eta - 1)W^2(y; v)}{S_m^2(y|x_0; \psi)} \right] - \eta \left[ \frac{V_{\psi}^2 - V_{\psi_2}^2 - V_{\psi_2}^2}{V_{\psi}^2 - V_{\psi_2}^2} + \eta \frac{(V_{\psi}^2 V_{\psi_1} - V_{\psi_2}^2 V_{\psi_2})^2}{V_{\psi}^2 V_{\psi_2}^2} \right] .$$

Because $\eta \geq 1$, we have

$$v^T H(\psi; \psi; z_i) v \leq -(1 - \delta) \left[ \frac{w^2(y; v)}{f_m^2(y|x_0; \psi)} + \frac{(\eta - 1)W^2(y; v)}{S_m^2(y|x_0; \psi)} \right] - \eta \left( \frac{V_{\psi}^2 - V_{\psi_2}^2}{V_{\psi}^2 - V_{\psi_2}^2} \right)^2 .$$

where

$$U_0(\psi; \psi; z_i) = (1 - \delta) \left[ \frac{\beta_m(y)}{f_m^2(y|x_0; \psi)} + \frac{e^\tau x - 1}{S_m^2(y|x_0; \psi)} \right] + e^\tau x \left[ \frac{S_m(y|x_0; \psi) e^{\alpha x} - S_m(y|x_0; \psi) e^{\alpha x}}{S_m(y|x_0; \psi) e^{\alpha x} - S_m(y|x_0; \psi) e^{\alpha x}} \right] .$$

where $u^{\otimes 2} = uu^T$ for a vector $u$. Now $v^T \sum_{i=1}^n U_0(\psi; \psi; z_i) v = 0$ implies, for all $i = 1, \ldots, n, \sum_{j=0}^n v_j^2 = 0$. The proof of Lemma 1 is complete.

**Proof of Lemma 2**

The derivatives of $\ell_m(\psi; \psi; z)$ with respect to $\psi$ are

$$\frac{\partial \ell_m(\psi; \psi; z)}{\partial \psi} = (1 - \delta)[1 + e^{\alpha x} \log(S_m(y|x_0; \psi))] x + \delta \left( \frac{S_m(y|x_0; \psi)}{S_m(y|x_0; \psi)} - S_m(y|x_0; \psi) \right) ,$$

$$\frac{\partial^2 \ell_m(\psi; \psi; z)}{\partial \psi \partial \psi^T} = (1 - \delta)e^{\alpha x} \log(S_m(y|x_0; \psi)) x x^T + \delta \left( \frac{S_m(y|x_0; \psi)}{S_m(y|x_0; \psi)} - S_m(y|x_0; \psi) \right) - \left( \frac{S_m(y|x_0; \psi)}{S_m(y|x_0; \psi)} - S_m(y|x_0; \psi) \right)^2 \right) .$$

where

$$S_m(t|x; \psi; \psi) = e^{\alpha x} S_m(t|x_0; \psi) e^{\alpha x} \log(S_m(t|x_0; \psi)) x .$$

$$S_m(t|x; \psi; \psi) = e^{\alpha x} S_m(t|x_0; \psi) e^{\alpha x} \log(S_m(t|x_0; \psi)) x x^T + S_m(t|x; \psi; \psi) .$$

Clearly Equation (A3) is negative semidefinite for $\delta = 0$. Denote temporally $\eta = e^{\alpha x}$, and $V_j = S_m(y|x_0; \psi)$, $j = 1, 2$. We know $1 \geq V_1 \geq V_2 \geq 0$. It is easy to show that, for $\delta = 1,$

$$\frac{\partial^2 \ell_m(\psi; \psi; z)}{\partial \psi \partial \psi^T} = \left[ \frac{\eta(V_{1}^n \log V_1 - V_{2}^n \log V_2)}{V_{1}^n - V_{2}^n} + \frac{\eta^2(V_{1}^n \log V_1 - V_{2}^n \log V_2)}{V_{1}^n - V_{2}^n} - \frac{\eta^2(V_{1}^n \log V_1 - V_{2}^n \log V_2)^2}{(V_{1}^n - V_{2}^n)^2} \right] x x^T$$
From Lemma 1 and Equation (A1) it follows that

\[
S = x \log x - x + 1 \log x/(x - 1)^2
\]
is decreasing on [1, \( V_2^{-\eta} \)]. Therefore, by inequality \( x \log x > x - 1 \), for \( x > 1 \),

\[
\xi \left( \frac{V_1}{V_2} \right)^\eta - \eta \log V_2 \geq \xi (V_2^{-\eta}) - \eta \log V_2 = \frac{[V_2^{-\eta} \log V_2^{-\eta} - (V_2^{-\eta} - 1) \log V_2^{-\eta}]}{(V_2^{-\eta} - 1)^2} > 0.
\]

For any \( v \in \mathbb{R}^d \), \( v^T [\partial^2 \mathcal{E}_m(y, \mathbf{p}; z)/(\partial y \partial y^T)] v = 0 \) implies \( v^T \mathbf{x}_i = 0 \) for all \( i \) and thus \( v = 0 \). The lemma is proved.

**Proof of Theorem 1**

If \( y^T x_0 \leq \min_{y \in S_m}(y^T x_1) \), we have \( y^T \mathbf{x}_i \geq 0 \). By Lemma 1, \( \mathcal{E}_m(y, \mathbf{p}) \) is strictly concave on the compact and convex set \( S_m \) for the fixed \( y \). By the optimality condition for convex optimization,\(^{36}\) we have that \( \mathbf{p}(y) \) is the unique maximizer of \( \mathcal{E}_m(y, \mathbf{p}) \) if and only if

\[
\nabla \mathcal{E}_m(y, \mathbf{p}) = 0,
\]

where \( \nabla \mathcal{E}_m(y, \mathbf{p}) = \partial \mathcal{E}_m(y, \mathbf{p}) / \partial \mathbf{p} \). Therefore \( \mathbf{p}(y) \) is a maximizer of \( \mathcal{E}_m(y, \mathbf{p}) \) for the fixed \( y \) if and only if

\[
\sum_{i=1}^n e_i^T \mathbf{x}_i \geq \sum_{i=1}^n \frac{\partial \mathcal{E}_m}{\partial p_i \mathbf{y}}[y, \mathbf{p}(y)] = \sum_{i=1}^n \Psi_i[y, \mathbf{p}(y); \mathbf{z}_i].
\]

for all \( j = 0, \ldots, m^* \) with equality if \( \mathbf{p}_j(y) > 0 \). The proof is complete.

**Proof of Theorem 2**

Following the proof of Theorems 1 and 2 and the Corollary of Peters and Walker,\(^{37}\) we define \( \Pi = \text{diag}(\mathbf{p}) \) and \( A(y, \mathbf{p}) = \Pi \nabla \mathcal{E}_m(y, \mathbf{p}) \), where \( \Psi(y, \mathbf{p}) = [\Psi_0(y, \mathbf{p}), \ldots, \Psi_m(y, \mathbf{p})]^T \). Then

\[
A(y, \mathbf{p}) = \frac{1}{\lambda_n(y)} \Pi \nabla \mathcal{E}_m(y, \mathbf{p}).
\]

Its gradient is

\[
\nabla A(y, \mathbf{p}) = \frac{\partial A(y, \mathbf{p})}{\partial \mathbf{p}^T} = \frac{1}{\lambda_n(y)} \text{diag} [\nabla \mathcal{E}_m(y, \mathbf{p})] + \frac{1}{\lambda_n(y)} \Pi \frac{\partial \nabla \mathcal{E}_m(y, \mathbf{p})}{\partial \mathbf{p}^T}.
\]

For any norm on \( \mathbb{R}^{m^*+1} \), we have \( A(y, \mathbf{p}) - \mathbf{p} = \nabla A(y, \mathbf{p})(\mathbf{p} - \mathbf{p}) + \mathcal{O}(\|\mathbf{p} - \mathbf{p}\|^2) \). Consider \( \nabla A(y, \mathbf{p}) \) as an operator on subspace \( \mathbb{Z}_m = \{ \mathbf{z} \in \mathbb{R}^{m^*+1} : 1^T \mathbf{z} = 0 \} \). If all components of \( \mathbf{p}(y) \) are positive then \( \nabla \mathcal{E}_m(y, \mathbf{p}(y)) = \lambda_n(y) \mathbf{1} \), and \( \nabla A(y, \mathbf{p}) \mathbf{y} = I_{m^*+1} - \mathbf{Q} \), where

\[
\mathbf{Q} = -\frac{1}{\lambda_n(y)} \Pi \frac{\partial^2 \mathcal{E}_m(y, \mathbf{p})}{\partial \mathbf{p} \partial \mathbf{p}^T}.
\]

From Lemma 1 and Equation (A1) it follows that \( \mathbf{Q} \) is a left stochastic matrix and \( \mathbf{p}^T \frac{\partial^2 \mathcal{E}_m(y, \mathbf{p})}{\partial \mathbf{p} \partial \mathbf{p}^T} = -\frac{\partial \mathcal{E}_m(y, \mathbf{p})}{\partial \mathbf{p}^T} = -\lambda_n(y) \mathbf{1}^T \). So \( \mathbb{Z}_m \) is invariant under \( \mathbf{Q} \).
Define an inner product \( \langle \cdot, \cdot \rangle \) by \( \langle u, v \rangle = u^\top \Pi^{-1} v \) for \( u, v \) in \( \mathbb{Z}_m \). It can be easily shown that, with respect to this inner product, \( Q \) is symmetric and positive semidefinite on \( \mathbb{Z}_m^\prime \):

\[
\langle u, Qu \rangle = -\frac{1}{\lambda_n(y)} u^\top \frac{\partial^2 \ell_m(y, \tilde{p})}{\partial \tilde{p} \partial \tilde{p}^\top} u \geq 0.
\]

\[
\langle u, Qv \rangle = -\frac{1}{\lambda_n(y)} u^\top \frac{\partial^2 \ell_m(y, \tilde{p})}{\partial \tilde{p} \partial \tilde{p}^\top} v = \langle Qu, v \rangle.
\]

Let \( \mu_0 \) and \( \mu_m \) be the smallest and largest eigenvalues of \( Q \) associated with eigenvectors in \( \mathbb{Z}_m^\prime \). Then the operator norm of \( \nabla^2 \ell_m(y, y) \) on \( \mathbb{Z}_m^\prime \) w.r.t. this inner product equals \( \max\{1 - \mu_0, |1 - \mu_m|\} \). It is clear that \( 0 \leq \mu_0 \leq \mu_m \leq 1 \) because \( Q \) is a left stochastic matrix. By Lemma 1 we have \( \mu_0 > 0 \). Similar to the proof of theorem 2 of Peters and Walker,\(^7\) the assertion of the theorem follows. If \( \tilde{p} \) contains zero component(s), say \( \tilde{p}_j = 0, j \in J_0 \), deleting the \( j \)th row and \( j \)th column of the vectors and matrices in the above proof for all \( j \in J_0 \), we can show that the iterates \( \tilde{p}^{[s]}_j \), \( s = 0, 1, \ldots \), converge to \( \tilde{p}_j \) as \( s \to \infty \) for all \( j \notin J_0 \). Because \( \sum_j \tilde{p}^{[s]}_j = 1 \) and \( \tilde{p}^{[s]}_j \geq 0, j = 0, 1, \ldots, m' \), for those \( j \in J_0, \tilde{p}^{[s]}_j \) converges to zero as \( s \to \infty \). The proof of Theorem 2 is complete.

**Proof of Theorem 3**

Let \( \mathbb{B}_2(r) = \{ y : ||y - y_0|| \leq r \} \), where \( || \cdot || \) denotes the Euclidean norm in \( \mathbb{R}^d \). For a decreasing positive sequence \( \varepsilon_n \downarrow 0 \) as \( n \to \infty \), let \( A_m(\varepsilon_n) \) be a subset of \( S_{m+1} \) so that, for all \( i \in [0, r] \), \( f_m(t|x_0; p) - f(t|x_0)/f(t|x_0) \leq \varepsilon_n \). Clearly, for all \( p \in A_m(\varepsilon_n) \), we have \( |S_m(t|x_0; p) - S(t|x_0)|/S(t|x_0) \leq \varepsilon_n \). We need the following lemma for the proof.

**Lemma 4.** Suppose that Assumptions 1 and 2 with \( m \geq C_0 r^{1/2} \) for some constant \( C_0 \), and Condition i are satisfied for an \( i = 0, 1, 2 \) and an \( \varepsilon \in (0, 1/2) \). If \( ||y - y_0|| \leq Cn^{-1+\varepsilon} \) then for any \( \varepsilon' \in (\varepsilon, 1/2) \) and \( n \) large enough the maximizer \( \hat{p}(y) \) of \( \ell_m(y, p) \) almost surely satisfies \( D^2[\hat{p}(y); x_0] \leq C' n^{-1+\varepsilon'} \), for some constant \( C' > 0 \), where \( x_0 = x_0, \hat{p}(y) \in A_m(\varepsilon_n) \), and \( \varepsilon_n = n^{-1-r}/2 \). Conversely, if \( D^2[\hat{p}(y); x_0] \leq Cn^{-1+\varepsilon} \), for some \( x_0 \), then for any \( \varepsilon' \in (\varepsilon, 1/2) \) and \( n \) large enough the maximizer \( \hat{p}(y) \) of \( \ell_m(y, p) \) for the fixed \( p \) almost surely satisfies \( ||\hat{p}(y) - y_0|| \leq C'n^{-1+\varepsilon'} \), for some constant \( C' > 0 \).

**Proof of Lemma 4.** Define \( \ell(y, f_0) = \sum_i \ell(y, f_0; z_i) \) and \( R(y, p) = \ell(y, f_0) - \ell_m(y, p) \). By Taylor expansion \( \log x = x - 1 + (x-1)^{2/2} + o((x-1)^2) \) we have, for all \( p \in A(\varepsilon_n) \),

\[
R(y, p) = -\sum_{i=1}^n \frac{1}{\delta_i} \left( y - y_0 \right)^\top \tilde{x}_i + \log \frac{f_m(y|x_0; p)}{f(y|x_0)} + (e^{\tilde{x}_i} - 1) \log \frac{S_m(y|x_0; p)}{S(y|x_0)} + \sum_{i=1}^n \log \left[ \frac{S_m(y_1|x_0; p)e^{\tilde{x}_i} - S_m(y_2|x_0; p)e^{\tilde{x}_i}}{S(y_1|x_0)e^{\tilde{x}_i} - S(y_2|x_0)e^{\tilde{x}_i}} \right] + \frac{1}{2} \sum_{i=0}^1 \frac{1}{r_2(\xi_1(p))} + \frac{1}{2} \sum_{i=0}^1 \frac{1}{r_1(\xi_1(p))},
\]

where \( \tilde{x}_0(\xi_1(p)) = \sum_{i=1}^n [(1 - \delta_i)U_0(\xi_1(p)) + \sum_{i=1}^n (1 - \delta_i)[U_1(\xi_1(p)) + (e^{\tilde{x}_i} - 1)U_2(\xi_1(p))], \tilde{x}_1(\xi_1(p)) = \sum_{i=1}^n \delta_i U_0(\xi_1(p)) + \sum_{i=1}^n \delta_i U_2(\xi_1(p)). U_0(\xi_1(p)) = (y_0 - y)^\top \tilde{x}_i + (e^{\tilde{x}_i} - e^{\tilde{x}_i}) \log S(y|x_0), \)

\[
U_1(\xi_1(p)) = \frac{S_m(y_1|x_0; p)e^{\tilde{x}_i} - S_m(y_2|x_0; p)e^{\tilde{x}_i}}{S(y_1|x_0)e^{\tilde{x}_i} - S(y_2|x_0)e^{\tilde{x}_i}},
\]

It is clear, for all real \( \xi \),

\[
|e^{\xi} - 1 - \sum_{i=1}^n \frac{1}{r_1(\xi)}| \leq \frac{e^{|\xi|}}{(j+1)!} O(|\xi|^{j+1}) \quad (j = 1, 2, \ldots).
\]
Under Condition 0: The data are uncensored so that all $\delta_i = 0$. By integration by parts we have

$$
E[U_0(y)] = \int_X \left( (y_0 - y)^\top \dot{x} - (e^{\dot{r}^\top \dot{x}} - e^{r^\top \dot{x}}) \right) \int_0^\infty \log S(y|x_0) dS(y|x) dH(x)
$$

$$
= \int_X \left( (y_0 - y)^\top \dot{x} - (e^{\dot{r}^\top \dot{x}} - e^{r^\top \dot{x}}) \right) \int_0^\infty f(y|x) dy dH(x)
$$

$$
= \int_X \left( (y_0 - y)^\top \dot{x} - (e^{\dot{r}^\top \dot{x}} - e^{r^\top \dot{x}}) \right) dH(x)
$$

$$
= \int_X \left( \sum_{j=2}^\infty \frac{(y-y_0)^\top \dot{x}^j}{j!} + e((y-y_0)^\top \dot{x}) \right) dH(x).
$$

(A10)

where $j \geq 3$. Because $X$ is bounded we have, for all $y \in \mathbb{B}_d(n^{-1/2})$,

$$
\lambda_0 \| y - y_0 \|^2 \leq E[U_0(y)] - \alpha(\| y - y_0 \|^2) \leq \lambda_d \| y - y_0 \|^2,
$$

(A11)

where $\lambda_0 > 0$ and $\lambda_d > 0$ are, respectively, the minimum and maximum eigenvalues of $E(XX^\top)$. Similarly, repeated integration by parts implies

$$
E[U_0^2(y)] = \int_X \left( (y - y_0)^\top \dot{x}^2 + (e^{r^\top \dot{x}} - e^{\dot{r}^\top \dot{x}})^2 \right) \int_0^\infty \log^2 S(y|x_0) f(y|x) dy dH(x)
$$

$$
- 2 \int_X (y - y_0)^\top \dot{x} (e^{r^\top \dot{x}} - e^{\dot{r}^\top \dot{x}}) \int_0^\infty \log S(y|x_0) f(y|x) dxdH(x)
$$

$$
= \int_X \left( (y - y_0)^\top \dot{x} - (e^{r^\top \dot{x}} - e^{\dot{r}^\top \dot{x}} - 1)^2 \right) + (e^{r^\top \dot{x}} - e^{\dot{r}^\top \dot{x}} - 1)^2 dH(x).
$$

By Equation (A9), we have $|e^i - 1 - x| \leq |x|^{2\beta|x|}/2$, and

$$
\text{var}[U_0(y)] \leq \int_X \left[ \frac{1}{4} (y - y_0)^\top \dot{x}^4 e^{2\beta|x|} + (e^{r^\top \dot{x}} - e^{\dot{r}^\top \dot{x}} - 1)^2 \right] dH(x).
$$

(A12)

Consequently

$$
\text{var}[U_0(y)] \leq \eta \lambda_d n^{1+\epsilon}.
$$

(A13)

Therefore by the law of iterated logarithm we have, for all $y \in \mathbb{B}_d(n^{-1/2})$,

$$
\tilde{R}_{00}(y, p) = \sum_{i=1}^n U_{0i}(y) = nE[U_0(y)] + O(n \text{var}[U_0(y)] \log \log n)^{1/2})
$$

$$
\leq \lambda_d n^{\epsilon} + O((n \log \log n)^{1/2}), \text{ a.s.}
$$

(A14)

For $j = 1, 2$ and $i = 1, \ldots, n$, we define

$$
V_{ji}(p) = U_{1i}^j(p) + (e^{\tilde{r}^\top \tilde{x}_i} - 1)U_{2i}^j(p),
$$

(A15)

$$
W_{ji}(y, p) = U_{1i}^j(y) + (e^{\tilde{r}^\top \tilde{x}_i} - 1)U_{2i}^j(p) = V_{ji}(p) + (e^{\tilde{r}^\top \tilde{x}_i} - e^{\tilde{r}^\top \tilde{x}_i})U_{2i}^j(p).
$$

(A16)

Integration by parts implies

$$
E[V_{11}(p)] = E[U_{11}(p) + (e^{\tilde{r}^\top \tilde{x}_i} - 1)U_{21}(p)]
$$

$$
= \int_X e^{r^\top \dot{x}} \int_0^\infty \left\{ \frac{f_m(y|x_0; p)}{f(y|x_0)} - 1 + (e^{r^\top \dot{x}} - 1) \left[ \frac{S_m(y|x_0; p)}{S(y|x_0)} - 1 \right] \right\} dH(x).
$$
\[ \times S(y|x_0)e^{y^T \Delta - 1}dyH(x) \]
\[ = \int_\mathcal{X} e^{y^T \tilde{x}} \int_0^\infty \left\{ \left[ f_m(y|x_0; p) - f(y|x_0) \right] S(y|x_0)e^{y^T \Delta - 1}dy \right\} dH(x) \]
\[ - \left[ S_m(y|x_0; p) - S(y|x_0) \right] dS(y|x_0)e^{y^T \Delta - 1} dH(x) = 0. \]

We also have, by integration by parts,
\[ 2E[U_{11}(p)U_{22}(p)|x_i] = -E[(e^{y^T \tilde{x}} - 2)U_{22}^2(p)|x_i]. \]  
(A17)

Therefore by Equation (A17) we have $2E[(e^{y^T \tilde{x}} - 1)U_{11}(p)U_{22}(p)] = -E[(e^{y^T \tilde{x}} - 1)(e^{y^T \tilde{x}} - 2)U_{22}^2(p)]$, and \[ \text{var}[V_1(p)] = E[V_{11}^2(p)] = E[(U_{11}(p) + (e^{y^T \tilde{x}} - 1)U_{22}(p))^2] = E[U_{11}^2(p) + (e^{y^T \tilde{x}} - 1)U_{22}^2(p)]. \]

Thus
\[ \text{var}[V_1(p)] = E[V_{11}^2(p)] = E[V_{22}(p)] = \chi_0^2(p; x_0) + E[(e^{y^T \tilde{x}} - 1)U_{22}^2(p)]. \]

If $T$ is independent of covariate $X$, then $y_0 = 0$ and $E[V_{11}^2(p)] = \chi_0^2(p; x_0)$. If $y_0 \neq 0$, we have $\gamma \neq 0$ for large $n$ and
\[ E[W_{21}(\gamma; p)] = \chi_0^2(p; x_0) + D_{01}^2(\gamma; p; x_0). \]  
(A19)

By the triangular inequality
\[ D_{01}^2(\gamma_0; p; x_0) \leq D_{01}^2(\gamma; p; x_0) + E[|e^{\gamma^T \tilde{x}} - e^{\gamma_0^T \tilde{x}}|U_{22}^2(p)]. \]  
(A20)

By Equation (A16),
\[ \text{var}[W_1(\gamma; p)] = E[W_{21}(\gamma; p)] + \{E[W_{21}(\gamma; p)]\}^2 + E[(e^{y^T \tilde{x}} - 1)(e^{y^T \tilde{x}} - e^{y_0^T \tilde{x}})U_{22}^2(p)]. \]

By Assumption 1, for all $\gamma \in \mathbb{B}_d(n^{-1-\epsilon}/2)$,
\[ |E[(e^{y^T \tilde{x}} - 1)(e^{y^T \tilde{x}} - e^{y_0^T \tilde{x}})U_{22}^2(p)]| \leq E[|e^{y^T \tilde{x}} - e^{y_0^T \tilde{x}}|(e^{y^T \tilde{x}} - 1)U_{22}^2(p)] = O(n^{-1-\epsilon/2})E[(e^{y^T \tilde{x}} - 1)U_{22}^2(p)]. \]

Therefore, if $E[W_{21}(\gamma; p)] \leq n^{-1+\epsilon'}$, for any $\epsilon' \in (\epsilon, 1/2)$, then by Equations (A18) to (A20), we have, for all $\gamma \in \mathbb{B}_d(n^{-1-\epsilon}/2)$ and all $p \in A_m(\epsilon_n)$,
\[ E[W_{21}(\gamma; p)] = E[(e^{y^T \tilde{x}} - e^{y_0^T \tilde{x}})U_{22}^2(p)] = O(n^{-1+\epsilon'+\epsilon'/2}) = o(n^{-1+\epsilon'}), \]  
(A21)

\[ \text{var}[W_{11}(\gamma; p)] = E[W_{21}(\gamma; p)] + O(n^{-2+\epsilon'+\epsilon'/2}) + O(n^{-3/2+\epsilon'+\epsilon'/2}) \]
\[ = E[W_{21}(\gamma; p)] + o(n^{-3/2+3\epsilon'/2}). \]

(A22)

For any $\epsilon' \in (\epsilon, 1/2)$, if
\[ E[W_{21}(\gamma; p)] = E[U_{11}^2(p) + (e^{y^T \tilde{x}} - 1)U_{22}^2(p)] = n^{-1+\epsilon'} \]
then we have, by Equations (A21), (A22), and the law of iterated logarithm,
\[ \tilde{R}_{01}(\gamma, p) = -nE[W_{11}(\gamma; p)] + O\left(\{n\text{var}[W_{11}(\gamma; p)]\} \log \log n\right)^{1/2} \]
\[ = o(n^{\epsilon'}), \quad \text{a.s.}, \]

and, by the strong law of large numbers\(^47\) for rowwise independent and identically distributed arrays $W_{21}(\gamma; p)$,
\[ \tilde{R}_{02}(\gamma, p) = \frac{1}{2} \sum_{i=1}^n W_{21}(\gamma, p) = \frac{n}{2} E[W_{21}(\gamma, p)] + o\{nE[W_{21}(\gamma, p)]\}, \quad \text{a.s.} \]
Thus, by Equation (A8), for all \( \mathbf{p} \) that satisfy (A23), there is an \( \eta > 0 \) so that \( R(\mathbf{r}, \mathbf{p}) = \sum_{j=0}^{2} \tilde{R}_{j}(\mathbf{r}, \mathbf{p}) \geq \eta n^{-c}, \) a.s., but at \( \mathbf{p} = \mathbf{p}_{0}, m \geq C_{n}^{n_{0}^{'}/p}, R(\mathbf{r}, \mathbf{p}_{0}) = O(n_{0}') = o(n_{0}^{'}) , \) a.s.. By Equation (A20), the minimizer \( \mathbf{\tilde{p}} \) of \( R(\mathbf{r}, \mathbf{p}) \) for the fixed \( \mathbf{r} \) satisfies \( D_{0}^{'}(\mathbf{\tilde{p}}; \mathbf{x}_{0}) \leq E[W_{2}(\mathbf{r}, \mathbf{p})] + E[e^{\mathbf{r}^{\top} \mathbf{x}_{0}} - e^{\mathbf{r}^{\top} \mathbf{\tilde{p}}} | U_{2}^{2}(\mathbf{p})] \leq C' n^{-1+c} \) for some constant \( C' \) and \( \mathbf{\tilde{p}} \in A_{m}(\epsilon_{n}). \)

Similarly, for any \( \mathbf{p} \) that satisfies \( D_{0}^{'}(\mathbf{p}; \mathbf{x}_{0}) \leq C n^{-1+c} \), we can prove that the maximizer \( \mathbf{\tilde{y}} \) of \( \mathbf{\tilde{c}}(\mathbf{r}, \mathbf{p}) \) for the fixed \( \mathbf{p} \) satisfies \( \| \mathbf{\tilde{y}} - \mathbf{y}_{0} \|^{2} \leq C' n^{-1+c} \), for all \( c' \in (\epsilon, 1/2) \), almost surely. The proof under Condition 0 is complete.

**Under Condition 1:** All the data are current status and all \( \delta_{i} = 1 \). We have

\[
E[U_{3}(\mathbf{r}, \mathbf{p})] = \left\{ \begin{aligned}
\frac{S_{m}(0|\mathbf{x}_{0}; \mathbf{p}) e^{\mathbf{r}^{\top} \mathbf{x}_{0}}}{1 - S(U|X)} - I(0 \leq T \leq U|X) \\
\frac{S_{m}(U|\mathbf{x}_{0}; \mathbf{p}) e^{\mathbf{r}^{\top} \mathbf{x}_{0}}}{S(U|X)} - I(U < T \leq U|X) - 1
\end{aligned} \right\}
\]

\[
= \int_{X} \int_{0}^{\infty} dG(u|x)|H(x) = 0,
\]

\[
E[U_{3}(\mathbf{r}, \mathbf{p})] = \int_{X} \int_{0}^{\infty} \frac{|S_{m}(u|\mathbf{x}; \mathbf{r}, \mathbf{p}) - S(u|\mathbf{x})|^{2}}{S(u|x)} dG(u|x) dH(x)
\]

\[
= \int_{X} \int_{0}^{\infty} \frac{S_{m}(u|\mathbf{x}_{0}; \mathbf{p}) e^{\mathbf{r}^{\top} \mathbf{x}_{0}}}{S(u|x)} - 1 \right\} O(u|x) dG(u|x) dH(x). \tag{A24}
\]

The law of iterated logarithm and the strong law of large numbers for \( U_{3} \)'s imply, for all \( \mathbf{p} \in A(\epsilon_{n}) \),

\[
R(\mathbf{r}, \mathbf{p}) = \mathbf{\tilde{R}}_{11}(\mathbf{r}, \mathbf{p}) + \mathbf{\tilde{R}}_{12}(\mathbf{r}, \mathbf{p})
\]

\[
= O(\sigma(U_{3})(n \log \log n)^{1/2}) + n \text{var}(U_{3}) + o[\text{var}(U_{3})], \quad \text{a.s.}
\]

By Taylor expansion, with \( u = e^{\mathbf{r}^{\top} \mathbf{x}}, a = e^{\mathbf{e}^{\top} \mathbf{x}}, v = S_{m}(y|\mathbf{x}_{0}; \mathbf{p}), b = S(y|\mathbf{x}_{0}), \)

\[
\frac{v^{b} - 1}{b^{a}} - 1 = (u - a) \log b + a \left( \frac{v}{b} - 1 \right) + R_{2}(\mathbf{r}, \mathbf{p}), \tag{A25}
\]

where

\[
R_{2}(\mathbf{r}, \mathbf{p}) = \frac{b}{2b^{a}} \left\{ \left[ (\log b)(u - a) + a \frac{v - b}{b} \right]^{2} + \left[ 2(u - a) - a \frac{v - b}{b} \right] \frac{v - b}{b} \right\}.
\]

for some \( (\tilde{a}, \tilde{b}) \) on the line segment joining \( (u,v) \) and \( (a,b) \), that is, \( \tilde{a} = (1 - \theta) e^{\mathbf{r}^{\top} \mathbf{x}} + \theta e^{\mathbf{r}^{\top} \mathbf{x}} \) and \( \tilde{b} = (1 - \theta) S(y|\mathbf{x}_{0}) + \theta S_{m}(y|\mathbf{x}_{0}; \mathbf{p}) \), \( 0 \leq \theta \leq 1 \). For all \( \mathbf{p} \in A_{m}(\epsilon_{n}), |v - b|/b \leq \epsilon_{n}, \)

\[
\frac{\tilde{b}}{b^{a}} = \frac{[b + \theta(v - b)]^{a + \theta(u - a)} - 1}{b^{a}} \leq (1 - \theta \epsilon_{n})^{a + \theta(u - a)} \leq C(1 - |u - a| \log b).
\tag{A26}
\]

For \( k = 1, 2, \)

\[
\frac{b^{k}}{b^{a}} = \frac{b}{b^{a}} \left( 1 + \theta \frac{v - b}{b} \right)^{-k} \leq (1 - \theta \epsilon_{n})^{-k} \leq C' \tag{A27}
\]

By the expansion \( \log (1 + z) = \sum_{k=1}^{\infty} (-1)^{k+1} z^{k} / k, |z| < 1 \), we have, for all \( \mathbf{p} \in A_{m}(\epsilon_{n}), \)

\[
\bar{b} = \log[b + \theta(v - b)] = \log b + O(|v - b|/b) = \log b + O(\epsilon_{n}). \tag{A28}
\]
For all positive integer \( k \), we have

\[
|z(\log z)^k| \leq k^k e^{-k}, \quad z \in [0, 1].
\] (A29)

For any \( y \in \mathbb{B}(n^{-1-\varepsilon}/2), \varepsilon \in (0, 1/2) \), and \( x_0 \) such that \( y^\top x_0 = \max_{x \in \mathcal{X}} y^\top x \). If, for \( \varepsilon' \in (\varepsilon, 1/2) \),

\[
D_2^2(p; x_0) = \int_{\mathcal{X}} \int_0^{\infty} \left[ \frac{S_m(u|x_0; p)}{S(u|x_0)} - 1 \right]^2 O(u|x)dG(u|x)dH(x) = n^{-1+\varepsilon'},
\]

then it follows from Equations (A25)-(A29), and the triangular inequality, that, for all \( p \in A_m(\varepsilon_n) \),

\[
\text{var}(U_{3i}) \geq |D_1(p; x_0) - \left\{ \int_{\mathcal{X}} \int_0^{\infty} \left[ e^{y^\top x} (e^{y^\top x} - 1) \log S(u|x_0) \right] \right\}^2 O(u|x)dG(u|x)dH(x) |^{1/2} + o(n^{-1+\varepsilon'}), \quad a.s.
\] (A30)

By Equation (A29), \( \text{var}(U_{3i}) \geq n^{-1+\varepsilon'/2} - 2e^{-1}E^{1/2}[O(Y|X)]O(n^{-1-\varepsilon/2})^2 + o(n^{-1+\varepsilon'}) \). Thus, there is an \( \eta_0 > 0 \), so that, for all \( p \) that satisfy \( D_2^2(p; x_0) = n^{-1+\varepsilon'} \), we have \( R(y, p) \geq \eta_0 n^{\varepsilon'} \), a.s. At \( p = p_0 \), with \( m \geq C_0 n^{1/\rho} \), \( R(y, p_0) = \mathcal{O}(n^\varepsilon) \), a.s. Therefore \( R(y, p) \) is minimized by \( \bar{p} = \bar{p}(y) \) such that

\[
D_2^2(\bar{p}; x_0) = \int_{\mathcal{X}} \int_0^{\infty} \left[ \frac{S_m(u|x_0; \bar{p})}{S(u|x_0)} - 1 \right]^2 O(u|x)dG(u|x)dH(x) < n^{-1+\varepsilon'}.
\] (A31)

Similarly, by Equation (A30), if \( D_2^2(\bar{p}; x_0) < n^{-1+\varepsilon} \) for an \( x_0 \in \mathcal{X} \), then the minimizer \( \bar{y} = \bar{y}(p) \) of \( R(y, p) \) satisfies \( \bar{y} \in \mathbb{B}(n^{-1-\varepsilon'/2}) \) for all \( \varepsilon' \in (\varepsilon, 1/2) \).

**Under Condition 2**: The data are Case 2 interval censored data \( \delta_i = 1, i = 1, \ldots, n \).

\[
E[U_{3i}(y, p)] = E \left\{ \left[ \frac{S_m(0|x_0; p)e^{y^\top X} - S_m(U|x_0; p)e^{y^\top X}}{1 - S(U|X)} I(0 \leq T \leq U|X) \right] \right. \\
+ \left. \left[ \frac{S_m(U|x_0; p)e^{y^\top X} - S_m(V|x_0; p)e^{y^\top X}}{S(U|X) - S(V|X)} I(U < T \leq V|X) \right] \right. \\
+ \left. \left[ \frac{S_m(V|x_0; p)e^{y^\top X} - S_m(\infty|x_0; p)e^{y^\top X}}{S(V|X)} I(V < T < \infty|X) \right] - 1 \right\} \\
= \int_{\mathcal{X}} \int_0^{\infty} \int_0^0 0dG(u, v|x)dH(x) = 0.
\]

Similarly

\[
E[U_{3i}^2(y, p)] = \int_{\mathcal{X}} \int_0^{\infty} \int_0^{\infty} \sum_{i=1}^3 \left[ \frac{S_m(y_{i-1}|x; y, p) - S_m(y_i|x; y, p)}{S(y_{i-1}|x) - S(y_i|x)} - 1 \right] \\
\times [S(y_{i-1}|x) - S(y_i|x)]dG(y_i, y_{i+1})dH(x),
\]

where \( y_0 = 0 \) and \( y_3 = \infty \). Simplifying notations, \( W_1 = U, W_2 = V, S_t = S_m(W_t|X; y, p), S_t = S(W_t|X), \) and \( \Lambda_t = \Lambda(W_t|X) \) \( (i = 1, 2) \), we have, clearly,

\[
E[U_{3i}^2(y, p)] \geq \left[ \frac{(S_1 - S_1)^2}{1 - S_1} + \frac{(S_2 - S_2)^2}{S_2} \right] \\
= E \left\{ \left[ \frac{S_m(Y_1|X; y, p)}{S(Y_1|X)} - 1 \right]^2 O(Y_1|X)S(Y_1|X) \right\} + E \left\{ \left[ \frac{S_m(Y_2|X; y, p)}{S(Y_2|X)} - 1 \right]^2 S(Y_2|X) \right\}.
\]

Thus the proof under Condition 2 can be done by the argument similar to the proof under Condition 1. The proof of Lemma 4 is complete.
Now we prove Theorem 3. If \( \gamma^{(0)} \) is chosen to be an efficient and asymptotically normal estimator of \( \gamma \) as in Cox and Huang and Wellner, then, under the conditions of the theorem, for large \( n \), almost surely \( \| \gamma^{(0)} - \gamma_0 \| < n^{-1+\epsilon} \). Lemma 4 and the convergence of \( \gamma^{(0)}, p^{(3)} \) imply that \( \| \gamma^{(0)} - \gamma_0 \| < n^{-1+\epsilon} \), and \( \hat{p} \in A_\varepsilon (\varepsilon_n) \). The proof is complete.

**Proof of Theorem 4**

**Uncensored Data:** all \( \delta_i = 0 \). Expansion of \( Q(\tilde{\gamma}, S_m) = \partial \ell_m(\tilde{\gamma}, p_0)/\partial \gamma \) at \( \gamma_0 \) gives

\[
0 = n^{-1/2} Q(\tilde{\gamma}, S_m) = Z_n - J_n n^{-1/2} (\tilde{\gamma} - \gamma_0) + n^{-1/2} R_n (\tilde{\gamma}),
\]

where

\[
Z_n = n^{-1/2} \sum_{i=1}^{n} (1 + e^{T_i \tilde{x}_i} \log [S_m(y_i|x_0;p_0)]) \tilde{x}_i,
\]

\[
J_n = -\frac{1}{n} \sum_{i=1}^{n} e^{T_i \tilde{x}_i} \log [S_m(y_i|x_0;p_0)] \tilde{x}_i \tilde{x}_i^T,
\]

\[
R_n (\tilde{\gamma}) = \frac{1}{2} \sum_{i=1}^{n} e^{T_i \tilde{x}_i} \log [S_m(y_i|x_0;p_0)] (\tilde{\gamma} - \gamma_0)^T \tilde{x}_i,
\]

and \( \tilde{\gamma} = \gamma_0 + \theta (\tilde{\gamma} - \gamma_0) \) for some \( \theta \in [0, 1] \). If \( m = m_n \) satisfies \( n^{1/2} m^{-\rho/2} = o(1) \) then \( J_n \to -E [\log S(T|X)XX^T] = E(XX^T) = I \) and \( Z_n \) converges in distribution to normal with mean \( 0 \) and variance \( I \). For any \( \varepsilon > 0 \) and large \( n \), \( R_n (\tilde{\gamma}) = O(n^\varepsilon) \), a.s. Therefore \( \sqrt{n} (\tilde{\gamma} - \gamma_0) = J_n^{-1} Z_n + O(n^{-1/2+\varepsilon}) \) converges in distribution to normal with mean \( 0 \) and variance \( I^{-1} \).

**Interval censored Data:** all \( \delta_i = 1 \). Expansion of \( Q(\tilde{\gamma}, S_m) = \partial \ell_m(\tilde{\gamma}, p_0)/\partial \gamma \) at \( \gamma_0 \) gives

\[
0 = n^{-1/2} Q(\tilde{\gamma}, S_m) = Z_n - J_n n^{-1/2} (\tilde{\gamma} - \gamma_0) + n^{-1/2} R_n (\tilde{\gamma}),
\]

where

\[
Z_n = n^{-1/2} \sum_{i=1}^{n} \frac{\tilde{S}_m(y_{i1}|x_i; \gamma_0; p_0) - \tilde{S}_m(y_{i2}|x_i; \gamma_0; p_0)}{\tilde{S}_m(y_{i1}|x_i; \gamma_0; p_0) - \tilde{S}_m(y_{i2}|x_i; \gamma_0; p_0)} \tilde{x}_i,
\]

\[
J_n = -\frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\tilde{S}_m(y_{i1}|x_i; \gamma_0; p_0) - \tilde{S}_m(y_{i2}|x_i; \gamma_0; p_0)}{\tilde{S}_m(y_{i1}|x_i; \gamma_0; p_0) - \tilde{S}_m(y_{i2}|x_i; \gamma_0; p_0)} \right\}
\]

\[
\cdot \frac{1}{[\tilde{S}_m(y_{i1}|x_i; \gamma_0; p_0) - \tilde{S}_m(y_{i2}|x_i; \gamma_0; p_0)]^2}
\]

\[
\sum_{i=1}^{n} \frac{\tilde{S}_m(y_{i1}|x_i; \gamma_0; p_0) - \tilde{S}_m(y_{i2}|x_i; \gamma_0; p_0)}{[\tilde{S}_m(y_{i1}|x_i; \gamma_0; p_0) - \tilde{S}_m(y_{i2}|x_i; \gamma_0; p_0)]^2} \tilde{x}_i \tilde{x}_i^T,
\]

for any \( \varepsilon > 0 \) and large \( n \), \( R_n (\tilde{\gamma}) = O(n^\varepsilon) \), a.s. If \( m = m_n \) satisfies \( n^{1/2} m^{-\rho/2} = o(1) \) then, for Case 1 interval censored data,

\[
J_n \to -E \left\{ \frac{[-S(U|X)\Lambda^2(U|X) + S(U|X)\Lambda^2(U|X)]XX^T}{1 - S(U|X)} \right\}
\]

\[
+ E \left\{ \frac{S(U|X)\Lambda^2(U|X)}{1 - S(U|X)} \right\} XX^T \equiv I,
\]

and for Case 2 interval censored data,

\[
J_n \to -E \left\{ \frac{[-S(U|X)\Lambda^2(U|X) + S(U|X)\Lambda^2(U|X) - S(V|X)\Lambda^2(V|X) + S(V|X)\Lambda^2(V|X)]XX^T}{1 - S(U|X)} \right\}
\]

\[
+ E \left\{ \frac{S(U|X)\Lambda^2(U|X)}{1 - S(U|X)} \right\} XX^T \equiv I,
\]

\[
\frac{S(U|X)\Lambda^2(U|X) - S(V|X)\Lambda^2(V|X)}{S(V|X)} XX^T \right\} \equiv I,
\]

\[
\frac{S(U|X)\Lambda^2(U|X)}{S(V|X)} XX^T \right\} \equiv I.
\]
where \( S_i = S(W_i|X) \), \( \Lambda_i = \Lambda(W_i|X) \), \( i = 1, 2 \), \( W_1 = U \), and \( W_2 = V \). It is clear

\[
I \geq E \left\{ \left[ \frac{S^2(U|X)\Lambda^2(U|X)}{1 - S(U|X)} + \frac{S^2(V|X)\Lambda^2(V|X)}{S(V|X)} \right] \bar{X}\bar{X}^T \right\}
= E \left\{ \left[ O(U|X)S(U|X)\Lambda^2(U|X) + S(V|X)\Lambda^2(V|X) \right] \bar{X}\bar{X}^T \right\}.
\]

In both cases, \( \mathbf{Z}_n \) converges in distribution to normal with mean 0 and variance \( I \). For any \( \epsilon > 0 \) and large \( n \), \( R_n(\hat{y}) = \mathcal{O}(n^\epsilon) \), a.s. Hence \( \sqrt{n}(\hat{y} - \gamma_0) = J_n^{-1}[\mathbf{Z}_n + \mathcal{O}(n^{-1/2+\epsilon})] \) converges in distribution to normal with mean 0 and variance \( I^{-1} \).