The Annals of Statistics
2000, Vol. 28, No. 5, 1472–1501

SEQUENTIAL CONFIDENCE REGIONS FOR MAXIMUM LIKELIHOOD ESTIMATES

BY A. DMITRIENKO AND Z. GOVINDARAJULU

Eli Lilly and Company and University of Kentucky

The goal of this paper is to develop a general framework for constructing sequential fixed size confidence regions based on maximum likelihood estimates. Asymptotic properties of the sequential procedure for setting up the confidence regions are analyzed under very broad assumptions on the underlying parametric model. It is shown that the proposed sequential procedure is asymptotically optimal in the sense that it approximates the optimal fixed-sample size procedure. It is further shown that the "cost of ignorance" associated with the sequential procedure is bounded. Applications are made to estimation problems arising in prospective and retrospective studies.

1. Introduction. Sequential techniques have been in the arsenal of modern statistical methods for over fifty years and found applications in a wide variety of problems. Most commonly, sequential estimation methods are utilized when one is interested in statistical inferences with fixed precision, for example, fixed size confidence interval estimation with a given coverage probability [Stein (1945), Chow and Robbins (1965)].

In general, one can successfully approach statistical problems involving fixed precision using the principle of accumulated information. In the scalar parameter case, the observed Fisher information is approximately inversely proportional to the mean squared error of the ML estimate of an unknown parameter. The mean squared error of the ML estimate can therefore be controlled if one designs a sequential sampling scheme which achieves a certain level of observed Fisher information. The same reasoning holds if the unknown parameter is multivariate. In the latter case, the observed Fisher information is usually replaced by the minimum eigenvalue of the observed Fisher information matrix. It is important to point out that this principle applies to a wide variety of situations including the cases of non-identically distributed and dependent observations. Lai and Siegmund (1983) show that a stopping rule based on the observed Fisher information can be utilized for constructing a fixed-accuracy estimate of an autoregressive parameter.

In this paper, the described principle of accumulated information is used for studying sequential maximum likelihood (ML) estimation. We develop a general framework for constructing fixed size confidence regions based on ML estimates. Section 2 considers properties of sequential ML estimation in a general setting and general conditions under which the ML estimates and their sequential versions are asymptotically normal. Section 3 explains how

Received February 1999; revised June 2000.
AMS 1991 *subject classifications*. Primary 62L12; secondary 62F10.
Key words and phrases. Sequential methods, asymptotic consistency, asymptotic efficiency.

utilizing these results and the principle of accumulated Fisher information one can set up a sequential fixed size multivariate confidence region for the unknown parameter vector. The confidence region has asymptotically the correct coverage probability and the expected number of observations required for constructing the confidence region is asymptotically equivalent to the (hypothetical) best fixed sample size computed under the assumption that the parameter vector is known in advance. It is also shown in Section 4 that under additional regularity assumptions the sequential procedure possesses an important property typically referred to as "bounded cost of ignorance." In Section 5, the general conditions of Sections 2 and 3 are verified for the ML estimates of unknown vector parameters in generalized linear models. The Appendix contains mathematical details of some proofs presented in Sections 2 and 3.

2. General properties of sequential ML estimates. In this section we will discuss asymptotic properties of maximum likelihood estimates in a general setting. A similar general setup for non-sequential ML estimation has been considered by Weiss (1971, 1973), Sweeting (1980), Kaufmann (1987). Their sufficient conditions have been modified in this section to provide a very general framework for establishing asymptotic normality of the sequential ML estimates and studying asymptotic properties of sequential confidence regions.

Specialized versions of these results have appeared in the literature. Anscombe (1952) used similar arguments for establishing asymptotic normality of sequential ML estimates of an unknown scalar parameter [see Govindarajulu (1987), Section 4.11, for details and references]. Grambsch (1983, 1989), Chang and Martinsek (1992) and Chang (1995) studied sequential ML estimation of a multivariate parameter with applications to fixed size confidence regions. Grambsch (1983, 1989) outlined a framework for sequential ML inference based on Gleser's (1969) multivariate extension of Anscombe's theorem which provides conditions for replacing a fixed-sample size by a random stopping time. Grambsch (1989) proposed a sequential sampling scheme for estimation of the parameters of a logistic regression in retrospective case-control studies. Chang and Martinsek (1992) and Chang (1995) considered a similar problem of ML estimation of the parameters of logistic and general binary response models. Siegmund and Sellke (1983) considered a more complicated application of sequential methods to the analysis of maximum partial likelihood estimates in proportional hazards models.

Let Z_n denote the vector (matrix) of observations. These may include both response and explanatory variables in the statistical model (note that explanatory variables may have random components). Assuming that the distribution of the response variables in Z_n belongs to a known parametric class indexed by a r-dimensional parameter λ ($\lambda \in \Lambda$), let $L_n(Z_n; \lambda)$ denote the likelihood function of the observations. If some of the explanatory variables in the data matrix Z_n are random, the likelihood function $L_n(Z_n; \lambda)$ is defined conditionally on those random components. The true value λ_0 of the vector parameter λ is unknown to the statistician and needs to be estimated from Z_n by the

method of maximum likelihood. Assume that the likelihood function is twice continuously differentiable with respect to λ and let $\ell_n(Z_n;\lambda) = \ln L_n(Z_n;\lambda)$. Let $\dot{\ell}_n(Z_n;\lambda)$ denote the score statistic, that is, $\dot{\ell}_n(Z_n;\lambda) = \partial\ell_n(Z_n;\lambda)/\partial\lambda$, and define the expected and observed Fisher information matrices $J_n(\lambda)$ and $\widehat{J}_n(\lambda)$, respectively, to be

$$J_n(\lambda) \stackrel{\text{def}}{=} \left[-E\frac{\partial^2\ell_n(Z_n;\lambda)}{\partial\lambda_i\partial\lambda_j}\right]_{1\le i,j\le r}, \quad \widehat{J}_n(\lambda) \stackrel{\text{def}}{=} \left[-\frac{\partial^2\ell_n(Z_n;\lambda)}{\partial\lambda_i\partial\lambda_j}\right]_{1\le i,j\le r},$$

where λ is a parameter point in Λ. It is important to note that the information matrices are assumed to depend on the unknown parameter and therefore the described general framework includes statistical models with a fairly complicated structure, for example, generalized linear models. Next, define the normalized expected and observed Fisher information matrices as

$$K_n(\lambda) \stackrel{\text{def}}{=} Q_n^{-1}J_n(\lambda)(Q_n')^{-1}, \quad \widehat{K}_n(\lambda) \stackrel{\text{def}}{=} Q_n^{-1}\widehat{J}_n(\lambda)(Q_n')^{-1}, \tag{2.1}$$

where Q_n is a square, non-singular matrix of normalizing coefficients (which may depend on λ_0). The elements of the normalizing matrix represent the rates at which the elements of the observed Fisher information matrix increase as $n \to \infty$.

The following conditions will be used throughout this paper. They will be utilized for proving asymptotic normality of the ML estimate $\widehat{\lambda}_n$ and its sequential version $\widehat{\lambda}_T$ computed at some random stopping time T. In Section 3, these conditions will be used for establishing asymptotic optimality of confidence regions constructed on the basis of the sequential ML estimate.

Before stating the conditions, we will need the following definitions.

DEFINITION 2.1. *A sequence of positive integer-valued random variables* $\{T_k,\ k \ge 1\}$ *is said to be regular if there exist positive integers* $\{t_k, k \ge 1\}$ *such that* $t_k \to \infty$ *and* $T_k/t_k \stackrel{p}{\to} 1$ *as* $k \to \infty$, *where* $\stackrel{p}{\to}$ *denotes convergence in probability.*

DEFINITION 2.2 [Hsu and Robbins (1947)]. *A sequence of random variables* $\{\zeta_k, k \ge 1\}$ *is said to converge completely to a constant* ζ *if* $\sum_{k=1}^{\infty} P\{|\zeta_k - \zeta| > \varepsilon\} < \infty$ *for any* $\varepsilon > 0$.

ASSUMPTIONS. Let λ_0 denote the true value of the unknown parameter λ and assume that, as $n \to \infty$,

A. $\|\dot{\ell}_n(Z_n;\lambda_0)\|/\phi_{\min}(J_n(\lambda_0))$ converges completely to 0, where $\phi_{\min}(A) =$ minimum eigenvalue of the matrix A.
B. $J_n(\lambda)$ is a continuous function of λ, $K_n(\lambda_0) \to K(\lambda_0)$, where $K(\lambda_0)$ is a positive definite matrix, and $\|\widehat{K}_n(\lambda) - K_n(\lambda_0)\| \to 0$ (completely) uniformly in a shrinking neighborhood of λ_0 in the sense that, for any $\varepsilon > 0$, there exists $\delta_\varepsilon > 0$ such that $\sum_{n=1}^{\infty} P\left\{\sup_{\|\lambda-\lambda_0\|\le\delta_\varepsilon} \|\widehat{K}_n(\lambda) - K_n(\lambda_0)\| > \varepsilon\right\} < \infty$.

C. $Q_n^{-1}\dot{\ell}_n(Z_n;\lambda_0) \xrightarrow{d} N(0, K(\lambda_0))$, where $\xrightarrow{d}$ denotes convergence in distribution.

D. $Q_{T_n}^{-1}\dot{\ell}_{T_n}(Z_{T_n};\lambda_0) \xrightarrow{d} N(0, K(\lambda_0))$ for any regular sequence of integer-valued random variables $\{T_n, n \geq 1\}$.

Conditions A–D will be illustrated in Section 5 where they will be verified for the ML estimates of unknown parameters in generalized linear models.

Theorem 2.1 shows that conditions A–D imply asymptotic normality of both the fixed sample and sequential ML estimates. We use the computed Fisher information matrix $\widehat{J}_n(\widehat{\lambda}_n)$ to normalize the deviation of the ML estimate since the marginal distribution of random components in the data matrix Z_n is generally unknown and thus one cannot compute the Fisher information matrix $J_n(\widehat{\lambda}_n)$.

The proof of Theorem 2.1 essentially follows from Cramèr's general methodology. It makes use of Lemma A.2 given in the Appendix. This lemma establishes conditions under which the ML estimate $\widehat{\lambda}_n$ converges completely λ_0. It is based upon a result of Fahrmeir and Kaufmann (1985).

THEOREM 2.1. *Assume that conditions* A, B *and* C *are satisfied and let* $\widehat{\lambda}_n$ *denote the ML estimate of* λ_0. *Then, as* $n \to \infty$,

$$(\widehat{\lambda}_n - \lambda_0)'\widehat{J}_n(\widehat{\lambda}_n)(\widehat{\lambda}_n - \lambda_0) \xrightarrow{d} \chi_r^2, \tag{2.2}$$

where χ_r^2 *denotes the chi-square distribution with* r *degrees of freedom. Further, if Condition* D *is also fulfilled then, for any sequence of regular integer-valued random variables* $\{T_n, n \geq 1\}$, *as* $n \to \infty$,

$$(\widehat{\lambda}_{T_n} - \lambda_0)'\widehat{J}_{T_n}(\widehat{\lambda}_{T_n})(\widehat{\lambda}_{T_n} - \lambda_0) \xrightarrow{d} \chi_r^2. \tag{2.3}$$

PROOF. Making use of a Taylor's series expansion and the likelihood equation, one obtains that $\dot{\ell}_n(Z_n;\lambda_0) = \widehat{J}_n(\eta_n)(\widehat{\lambda}_n - \lambda_0)$, where η_n is a point lying between λ_0 and $\widehat{\lambda}_n$. It follows from (2.1) that $Q_n^{-1}\dot{\ell}_n(Z_n;\lambda_0) = \widehat{K}_n(\eta_n)Q_n'(\widehat{\lambda}_n - \lambda_0)$. Therefore,

$$\begin{aligned}\widehat{K}_n^{1/2}(\widehat{\lambda}_n)Q_n'(\widehat{\lambda}_n - \lambda_0) &= \widehat{K}_n^{1/2}(\widehat{\lambda}_n)\widehat{K}_n^{-1}(\eta_n)Q_n^{-1}\dot{\ell}_n(Z_n;\lambda_0) \\ &= K^{-1/2}(\lambda_0)Q_n^{-1}\dot{\ell}_n(Z_n;\lambda_0) + R_n,\end{aligned} \tag{2.4}$$

where $R_n = \{\widehat{K}_n^{1/2}(\widehat{\lambda}_n)\widehat{K}_n^{-1}(\eta_n) - K^{-1/2}(\lambda_0)\}Q_n^{-1}\dot{\ell}_n(X_n;\lambda_0)$. By Condition C, the first term on the right-hand side of (2.4) is asymptotically normal $(0, I_r)$ and, by Lemma A.2, $\widehat{\lambda}_n \to \lambda_0$ a.s. Therefore, $\eta_n \to \lambda_0$ a.s. and, by Condition B,

$$\widehat{K}_n^{1/2}(\widehat{\lambda}_n)\widehat{K}_n^{-1}(\eta_n) \to K^{-1/2}(\lambda_0) \quad \text{a.s.} \tag{2.5}$$

Noting that $Q_n^{-1}\dot{\ell}_n(X_n;\lambda_0)$ is asymptotically normal, one can easily conclude that $R_n \xrightarrow{p} 0$. This proves that $\widehat{K}_n^{1/2}(\widehat{\lambda}_n)Q_n'(\widehat{\lambda}_n - \lambda_0) \xrightarrow{d} N(0, I_r)$ and together with (2.1) implies that (2.2) holds.

In order to prove (2.3), note that as in (2.4)

$$\widehat{K}_{T_n}^{1/2}(\widehat{\lambda}_{T_n})Q'_{T_n}(\widehat{\lambda}_{T_n}-\lambda_0)=K^{-1/2}(\lambda_0)Q_{T_n}^{-1}\dot{\ell}_{T_n}(Z_{T_n};\lambda_0)+R_{T_n}.$$

Utilizing (2.5) one can show that $\widehat{K}_{T_n}^{1/2}(\widehat{\lambda}_{T_n})\widehat{K}_{T_n}^{-1}(\eta_{T_n}) \to K^{-1/2}(\lambda_0)$ a.s. Together with Condition D this implies immediately that $R_{T_n} \xrightarrow{p} 0$. Therefore, making use of Condition D again implies that $\widehat{K}_{T_n}^{1/2}(\widehat{\lambda}_{T_n})Q'_{T_n}(\widehat{\lambda}_{T_n}-\lambda_0) \xrightarrow{d} (0, I_r)$ and (2.3) holds. □

3. Sequential fixed size confidence regions. The aim of this section is to demonstrate how sequential sampling schemes based on observed Fisher information can be used for making fixed precision inferences in the general setting described in Section 2. Motivated by Theorem 2.1, it is natural to define an approximate $(1-\alpha)100\%$ $(0<\alpha<1)$ confidence ellipsoid for λ_0 by

$$CR_n=\{\lambda\in R^r:(\widehat{\lambda}_n-\lambda)'\widehat{J}_n(\widehat{\lambda}_n)(\widehat{\lambda}_n-\lambda)\le\chi^2_{r,1-\alpha}\}, \tag{3.1}$$

where $\chi^2_{r,1-\alpha}$ is the $(1-\alpha)100\%$ percentile of the χ^2_r distribution. By (2.2), this confidence ellipsoid has asymptotically the projected coverage probability $(1-\alpha)$, that is, $P\{\lambda_0\in CR_n\}\to 1-\alpha$ as $n\to\infty$.

It can be shown that the size (defined as the length of the maximal axis) of CR_n is given by $2\{\chi^2_{r,1-\alpha}/\phi_{\min}(\widehat{J}_n(\widehat{\lambda}_n))\}^{1/2}$. Making a natural assumption that the observed information matrix does not remain bounded as $n\to\infty$, it is clear that the size of the confidence ellipsoid approaches 0. However, for any fixed value of n, the size cannot be controlled since $\phi_{\min}(\widehat{J}_n(\widehat{\lambda}_n))$ is a random quantity. In order to guarantee that the confidence ellipsoid is of fixed size in the sense that its maximal axis $\le 2d$, one can resort to sequential sampling and set the sample size equal to

$$N(d)=\min\{n\ge 1:\phi_{\min}(\widehat{J}_n(\widehat{\lambda}_n))\ge\chi^2_{r,1-\alpha}/d^2\}. \tag{3.2}$$

Note that this fixed precision sequential procedure is based on the principle of accumulated Fisher information because sampling is to be continued until the minimum eigenvalue of the observed information matrix $\widehat{J}_n(\widehat{\lambda}_n)$ achieves a certain threshold. This threshold depends on the characteristics of the confidence ellipsoid (namely, size and coverage probability) which should be specified prior to data collection.

The sample size $N(d)$ can be thought of as a random variable approximating the best fixed sample size

$$n(d)=\min\{n\ge 1:\phi_{\min}(J_n(\lambda_0))\ge\chi^2_{r,1-\alpha}/d^2\}. \tag{3.3}$$

Note that the best fixed sample size $n(d)$ is incomputable since the Fisher information matrix $J_n(\lambda_0)$ involves λ_0 and the expectation in $J_n(\lambda_0)$ is taken with respect to the unknown marginal distribution of the random components in the data matrix Z_n. No fixed-sample size procedure could accomplish the goal of constructing a confidence region for λ_0 of fixed size and prescribed

coverage probability. However, $J_n(\lambda_0)$ can be consistently estimated by the observed Fisher information matrix $\widehat{J}_{N(d)}(\widehat{\lambda}_{N(d)})$ and it is reasonable to expect that the sequential fixed size confidence ellipsoid

$$CR_{N(d)} = \{\lambda \in R^r : (\widehat{\lambda}_{N(d)} - \lambda)' \widehat{J}_{N(d)}(\widehat{\lambda}_{N(d)})(\widehat{\lambda}_{N(d)} - \lambda) \leq \chi^2_{r,1-\alpha}\}$$

has (at least asymptotically) the correct coverage probability $1 - \alpha$.

DEFINITION 3.1. *A sequential procedure associated with a stopping time $N(d)$ is said to be asymptotically equivalent to an optimal fixed-sample size procedure if, as $d \to 0$,* (i) $N(d)/n(d) \to 1$ *a.s.,* (ii) $EN(d)/n(d) \to 1$, (iii) $P\{\lambda_0 \in CR_{N(d)}\} \to 1 - \alpha$.

Conditions (ii) and (iii) are typically referred to as asymptotic efficiency and asymptotic consistency. These concepts were introduced by Chow and Robbins (1965) in the context of sequential interval estimation of the mean of i.i.d. observations. Their original ideas have later been extended to various settings including multivariate regression by Gleser (1965), Srivastava (1971) and generalized linear models by Chang and Martinsek (1992), Chang (1995). A comprehensive account of these and related topics can be found in Govindarajulu (1987), Chapter 5.

It is stated in Theorem 3.1 that the sequential procedure defined earlier in this section is asymptotically equivalent to the optimal fixed-sample size procedure. Therefore, without knowing λ_0, one obtains sequentially the results as good (in terms of sample size and coverage probability) as if one knew the true value of the parameter in advance.

THEOREM 3.1. *Assume that conditions* A–D *are satisfied and*

$$\lim_{\rho \to 1} \lim_{d \to 0} \{n(d\rho)/n(d)\} = 1. \tag{3.4}$$

Then the sequential procedure is asymptotically equivalent to the optimal fixed-sample size procedure.

REMARK 3.1. Condition (3.4) is satisfied if $n(d)$ is proportional to a power function or a slowly changing function of d which is the case in most applications.

It will be shown in Section 5 that the conditions of Theorem 3.1 are satisfied in generalized linear models with fixed and i.i.d. random covariates.

PROOF OF THEOREM 3.1. By (2.3), $(\widehat{\lambda}_{N(d)} - \lambda_0)' \widehat{J}_{N(d)}(\widehat{\lambda}_{N(d)})(\widehat{\lambda}_{N(d)} - \lambda_0) \xrightarrow{d} \chi^2_r$ as long as $\{N(d), d > 0\}$ is a regular sequence of random variables and therefore $P\{\lambda_0 \in CR_{N(d)}\} \to 1 - \alpha$ will follow from $N(d)/n(d) \to 1$.

Regarding the convergence of $N(d)/n(d)$ to 1, note that by the well-known properties of eigenvalues [see, e.g., Bellman (1960), Chapter 7], the eigenvalues of a sum of two non-negative definite matrices are uniformly larger

than the eigenvalues of either of these matrices. Therefore, recalling that $J_k(\lambda)$ is a sum of non-negative definite matrices, one can easily infer that $\{\phi_{\min}(J_k(\lambda)), k \geq 1\}$ is a non-decreasing sequence of positive numbers. Letting $U_k = \phi_{\min}(\widehat{J}_k(\widehat{\lambda}_k))$, $m_k = \phi_{\min}(J_k(\lambda_0))$, $b = \chi^2_{r,1-\alpha}/d^2$, it follows from Lemma A.3 that (3.4) and

$$\phi_{\min}(\widehat{J}_k(\widehat{\lambda}_k))/\phi_{\min}(J_k(\lambda_0)) \to 1 \quad \text{a.s.} \tag{3.5}$$

imply that $N(d)/n(d) \to 1$ a.s. Therefore it remains to prove that (3.5) holds. To this end one can utilize Lemma A.1 with $A_k = \widehat{J}_k(\widehat{\lambda}_k)$, $B_k = J_k(\lambda_0)$, $D_k = Q_k$, $a_k = \phi_{\min}(\widehat{J}_k(\widehat{\lambda}_k))$, $b_k = \phi_{\min}(J_k(\lambda_0))$. By conditions A and B, $D_k^{-1}A_k(D_k')^{-1} = \widehat{K}_k(\widehat{\lambda}_k) \to K(\lambda_0)$ and $D_k^{-1}B_k(D_k')^{-1} = K_k(\lambda_0) \to K(\lambda_0)$, where $K(\lambda_0)$ is a positive definite matrix. Hence (3.5) follows from Lemma A.1 and therefore $N(d)/n(d) \to 1$ a.s. as $d \to 0$.

In order to verify that $EN(d)/n(d) \to 1$, note first that (A.1) in Lemma A.1 implies

$$|\phi_{\min}(\widehat{J}_k(\widehat{\lambda}_k))/\phi_{\min}(J_k(\lambda_0)) - 1| \leq \|\widehat{K}_k(\widehat{\lambda}_k) - K_k(\lambda_0)\|/\phi_{\min}(K_k(\lambda_0)). \tag{3.6}$$

Furthermore, by Condition B, $\widehat{K}_k(\widehat{\lambda}_k) - K_k(\lambda_0) \to 0$ (completely) since $\widehat{\lambda}_k \to \lambda_0$ (completely). Observing that $\phi_{\min}(K_k(\lambda_0)) \to \phi_{\min}(K(\lambda_0)) > 0$, one obtains that the left-hand side of (3.6) also converges completely to 0. Therefore, for any $\varepsilon > 0$, $\sum_{k=1}^{\infty} P\{\phi_{\min}(\widehat{J}_k(\widehat{\lambda}_k)) < \varepsilon\phi_{\min}(J_k(\lambda_0))\} < \infty$ which in view of Lemma A.3 implies that $EN(d)/n(d) \to 1$ as $d \to 0$. □

4. Bounded cost of ignorance. This section is concerned with one interesting extension of the results established in Theorem 3.1. According to Theorem 3.1, the expected sample size $EN(d)$ which should be taken for setting up a fixed size confidence region for λ_0 is asymptotically equivalent to the best fixed sample size $n(d)$. Both $EN(d)$ and $n(d)$ obviously approach infinity as $d \to 0$ and therefore $EN(d)/n(d) \to 1$ does not necessarily imply that $EN(d) - n(d)$ is bounded. The question of whether $EN(d) - n(d)$ remains bounded as d approaches 0 is of much interest in sequential analysis. The difference $EN(d) - n(d)$ is called the *cost of ignorance* in not knowing the best fixed sample size $n(d)$ prior to the experiment. If the cost of ignorance is bounded from above by a constant as $d \to 0$ then the sequential procedure is said to have the property of *bounded cost of ignorance*.

In most simple problems of sequential point and interval estimation, the cost of ignorance can be analyzed using methods based on renewal theory under the assumption that the observations are i.i.d. [see Feller (1966), Chapter 11] and Siegmund [(1985), Chapter 8]. A comprehensive study of the asymptotic behavior of the cost of ignorance under broader assumptions was initiated by Lai and Siegmund (1977, 1979) and Woodroofe (1977) by developing *nonlinear renewal theory*. Nonlinear renewal theory shows that the cost of ignorance converges to a constant for certain stopping times generated by sums of dependent random variables. Specifically, the above-mentioned authors consider

stopping rules of the form

$$N(b) = \min\{k \geq 1 : U_k \geq b\} \text{ with } U_k = S_k + \xi_k, \qquad k \geq 1, \tag{4.1}$$

where $b > 0$, S_k is a partial sum of i.i.d. random variables with a positive mean and $\{\xi_k, k \geq 1\}$ is a "slowly changing sequence" satisfying certain regularity conditions. Second-order approximations to expected stopping times yielding asymptotic representations for the cost of ignorance have received a lot of attention in the literature. Recent developments in this area are discussed in Woodroofe (1982) and Siegmund (1985).

The central idea of nonlinear renewal theory is to express the random sequence $\{U_k, k \geq 1\}$ generating a stopping time as a sum of a random walk and a noise sequence [as in (4.1)]. Then, by controlling the noise sequence and analyzing the first term, one can generally show that the cost of ignorance for the stopping rule is asymptotically equal to a constant. A key role in these proofs is played by the assumption that the leading term (random walk) has a certain structure and the joint probability distribution of the summands belongs to a known class of distributions (e.g., sum of i.i.d. random variables, stationary process, Markov process, etc.). If a representation of this type is difficult to achieve or little can be assumed about the structure of the leading term, it appears problematic to apply the probabilistic methods of nonlinear renewal theory or appropriately modify them. It is not clear in most problems involving models with elaborate structure (time series and non-linear regression models) whether the difference between the expected sample size and the best fixed sample size (cost of ignorance) converges to a constant.

One of the ways to deal with this complication is to consider a broader problem of determining conditions under which the cost of ignorance is asymptotically bounded. The weaker assumptions on the underlying models may enable one to study the asymptotic behavior of the cost of ignorance in estimation problems with a fairly complex structure. Studying asymptotic efficiency of the sequential fixed-accuracy estimate of an autoregressive parameter considered by Lai and Siegmund (1983), Vexler and Konev (1995) demonstrate that nonlinear renewal theory cannot be directly applied in this situation. As an alternative, Vexler and Konev (1995) proposed to compute the Laplace-Stieltjes transform of the expected sample size and then utilize a Tauberian theorem to show that the cost of ignorance is bounded.

Building upon these results, Vexler and Dmitrienko (1999) developed a general framework for obtaining asymptotic approximations to expected stopping times with a bounded remainder term. Instead of relying on probabilistic methods (as in nonlinear renewal theory) based on assumptions of independence or stationarity, they heavily use Tauberian techniques to obtain expansions based on moment assumptions only. This new approach for deriving asymptotic expansions can greatly simplify asymptotic analysis of the cost of ignorance in a variety of sequential estimation problems.

The main theorem of Vexler and Dmitrienko (1999) is stated below as Lemma 4.1. We make use of this lemma for establishing the asymptotic bound-

edness of the cost of ignorance for the sequential procedure described in Section 2.

LEMMA 4.1. *Assume that* $\{U_k, k \geq 1\}$ *is a sequence of non-negative random variables and* $\{m_k, k \geq 1\}$ *is a sequence of non-negative real numbers such that* $m_k \to \infty$ *as* $k \to \infty$. *Assume further that, as* $b \to \infty$,

$$\sum_{k=1}^{\infty} P\{|U_k - m_k| > \delta m_k\} < \infty \quad \text{for some } \delta > 0, \tag{4.2}$$

$$b^{-1} \sum_{k=1}^{\infty} e^{-m_k/b} = O(1), \tag{4.3}$$

$$b^{-2} \sum_{k=1}^{\infty} e^{-m_k/b} E(U_k - m_k)^2 = O(1), \tag{4.4}$$

$$\sum_{k=1}^{\infty} \min(1, (b - m_k)^{-2}) = O(1). \tag{4.5}$$

Then $\sum_{k=1}^{\infty} (P\{U_k < b\} - I\{m_k < b\}) = O(1)$ *as* $b \to \infty$, *where* $I\{A\}$ *denotes the indicator function of the set* A.

Unlike nonlinear renewal theorems, this lemma requires no representation for the random sequence $\{U_k, k \geq 1\}$ and makes no assumptions about the joint distribution of the U_k's.

The first assumption of Lemma 4.1 is clearly satisfied if U_k/m_k converges completely to 1 and one can see that the other three assumptions hold under simple growth conditions on m_k and $E(U_k - m_k)^2$. For example, Vexler and Dmitrienko (1999) show that the assumptions of Lemma 4.1 are satisfied provided $\{m_k, k \geq 1\}$ is a non-decreasing sequence of positive real numbers such that

$$\liminf_{k \to \infty} (m_{k+s} - m_k) > 0 \qquad \text{for some integer } s \geq 1 \tag{4.6}$$

and

$$E|U_k - m_k|^t = O(k^{t/2}) \qquad \text{for some } t > 2 \text{ as } k \to \infty. \tag{4.7}$$

It is shown in Lai [(1996), Section 2] that (4.7) is not a restrictive condition and it is satisfied for a very broad class of stochastic sequences including martingales, moving averages and mixing sequences.

In order to see how Lemma 4.1 can be applied, let

$$U_k = \phi_{\min}(\widehat{J}_k(\widehat{\lambda}_k)), \quad m_k = \phi_{\min}(J_k(\lambda_0)), \quad b = \chi^2_{r,1-\alpha}/d^2. \tag{4.8}$$

One can see from (4.8) that in general U_k has no simple structure and therefore the methods of nonlinear renewal theory cannot be adapted so as to obtain an approximation to $EN(d)$. Chang (1995) points out that the sufficient conditions of nonlinear renewal theorems are very difficult to verify even in the

special case of sequential ML estimation of the parameter of a binary response model with i.i.d. random covariates.

Using the definitions of $N(d)$ and $n(d)$ it is easy to verify that

$$EN(d) \leq \sum_{k=1}^{\infty} P\{U_k < b\} \text{ and } n(d) = \sum_{k=1}^{\infty} I\{m_k < b\}.$$

If the conditions of Lemma 4.1 are satisfied then the difference $\sum_{k=1}^{\infty} P\{U_k < b\} - \sum_{k=1}^{\infty} I\{m_k < b\}$ is bounded from above uniformly in b which implies that $EN(d) - n(d)$ is less than some positive constant for any $d > 0$. In other words, this means that only a finite amount of information can be lost, if any, when the sequential procedure is applied instead of the hypothetical optimal procedure with the sample size $n(d)$.

Theorem 4.1, which constitutes a strengthening of Theorem 3.1, shows that under some additional assumptions one can utilize Lemma 4.1 and claim that the cost of ignorance in not knowing λ_0 and the marginal distribution of the random covariates for the proposed sequential procedure is bounded.

THEOREM 4.1. *Assume that conditions* A–D *of Section* 2 *are fulfilled. Assume further that*

$$\liminf_{k\to\infty} \{\phi_{\min}(J_{k+s}(\lambda_0)) - \phi_{\min}(J_k(\lambda_0))\} > 0 \quad \textit{for some integer } s \geq 1, \tag{4.9}$$

$$k^{-1}\, tr\, \{Cov\, \widehat{J}_k(\lambda_0)\} = O(1), \quad k \to \infty, \tag{4.10}$$

$$\phi^2_{\max}(Q_k Q_k')E\|\widehat{\lambda}_k - \lambda_0\|^4 = O(1), \quad k \to \infty, \tag{4.11}$$

$$\|\widehat{K}_k(\lambda) - \widehat{K}_k(\lambda_0)\| \leq \|\lambda - \lambda_0\| a_k \quad \textit{for any parameter point } \lambda, \tag{4.12}$$

where $\{a_k, k \geq 1\}$ *is a sequence of random variables such that*

$$k^{-2}\phi^2_{\max}(Q_k Q_k')Ea_k^4 = O(1), \qquad k \to \infty, \tag{4.13}$$

Then $\limsup_{d\to 0}(EN(d) - n(d)) < \infty$.

REMARK 4.1. Note that Conditions A–D ensure that $EN(d)/n(d) \to 1$ as $d \to 0$. The additional conditions are imposed in Theorem 4.1 in order to assert $\limsup_{d\to 0}(EN(d) - n(d)) < \infty$. Although the conditions of Theorem 4.1 may appear unwieldy, they can easily be verified for a variety of models. Roughly speaking, the theorem imposes conditions only on the growth rate of the minimum eigenvalue of the information matrix $J_k(\lambda_0)$ and the convergence rate of the ML estimate $\widehat{\lambda}_k$. Some examples will be provided in Section 5.

PROOF OF THEOREM 4.1. In order to prove that $EN(d) - n(d)$ is bounded from above, one needs to verify the conditions of Lemma 4.1 using the notation introduced in (4.8).

It has been shown in the proof of Theorem 3.1 [see the proof of $EN(d)/n(d) \to 1$] that (4.2) is a consequence of Conditions A and B of Section 2.

Next, by (4.9), there exists a positive number ε such that $m_k \geq \varepsilon k$ for any $k \geq k_0$. Let $q = e^{-\varepsilon/b}$ and note that $1 - q = O(1/b)$, $b \to \infty$. Then $\sum_{k=k_0}^{\infty} e^{-m_k/b} \leq \sum_{k=k_0}^{\infty} e^{-\varepsilon k/b} = \sum_{k=k_0}^{\infty} q^k = O(b)$ as $d \to 0$ or $b \to \infty$ and hence (4.3) holds.

Toward verifying (4.4) it suffices to show that, as $k \to \infty$,

$$E(U_k - m_k)^2 = E\{\phi_{\min}(\widehat{J}_k(\widehat{\lambda}_k)) - \phi_{\min}(J_k(\lambda_0))\}^2 = O(k). \tag{4.14}$$

Then making use of the same argument as in the verification of (4.3) will yield $\sum_{k=k_0}^{\infty} e^{-m_k/b} E(U_k - m_k)^2 = O(\sum_{k=k_0}^{\infty} kq^k) = O((1-q)^{-2}) = O(b^2)$ as $b \to \infty$. Now in order to establish (4.14), note that

$$\begin{aligned} &k^{-1}E\{\phi_{\min}(\widehat{J}_k(\widehat{\lambda}_k)) - \phi_{\min}(J_k(\lambda_0))\}^2 \\ &\quad \leq 2k^{-1}E\{\phi_{\min}(\widehat{J}_k(\lambda_0)) - \phi_{\min}(J_k(\lambda_0))\}^2 \\ &\qquad +2k^{-1}E\{\phi_{\min}(\widehat{J}_k(\widehat{\lambda}_k)) - \phi_{\min}(\widehat{J}_k(\lambda_0))\}^2 \end{aligned} \tag{4.15}$$

and consider the first term on the right-hand side of (4.15). It can be shown along the lines of Lemma A.1 that

$$\{\phi_{\min}(\widehat{J}_k(\lambda_0)) - \phi_{\min}(J_k(\lambda_0))\}^2 \leq \|\widehat{J}_k(\lambda_0) - J_k(\lambda_0)\|^2 \tag{4.16}$$

and hence, noting that $E\widehat{J}_k(\lambda_0) = J_k(\lambda_0)$ and using (4.10),

$$\begin{aligned} &k^{-1}E\{\phi_{\min}(\widehat{J}_k(\lambda_0)) - \phi_{\min}(J_k(\lambda_0))\}^2 \\ &\quad \leq k^{-1}E\|\widehat{J}_k(\lambda_0) - J_k(\lambda_0)\|^2 \\ &\quad = k^{-1}\ \mathrm{tr}\ \{\ \mathrm{Cov}\ \widehat{J}_k(\lambda_0)\} = O(1), \qquad k \to \infty. \end{aligned} \tag{4.17}$$

Next, utilizing an inequality similar to (4.16), the definition of the normalized information matrices $K_k(\lambda_0)$ and $\widehat{K}_k(\lambda_0)$ and then (4.13), one obtains that

$$\begin{aligned} &E\{\phi_{\min}(\widehat{J}_k(\widehat{\lambda}_k)) - \phi_{\min}(\widehat{J}_k(\lambda_0))\}^2 \\ &\quad \leq E\|\widehat{J}_k(\widehat{\lambda}_k) - \widehat{J}_k(\lambda_0)\|^2 \\ &\quad \leq \|Q_k\|^4 E\|\widehat{K}_k(\widehat{\lambda}_k) - \widehat{K}_k(\lambda_0)\|^2 \\ &\quad \leq r^2\phi_{\max}^2(Q_kQ_k')E\|\widehat{K}_k(\widehat{\lambda}_k) - \widehat{K}_k(\lambda_0)\|^2 \\ &\quad \leq r^2\{\phi_{\max}^2(Q_kQ_k')Ea_k^4\}^{1/2}\{\phi_{\max}^2(Q_kQ_k')E\|\widehat{\lambda}_k - \lambda_0\|^4\}^{1/2}. \end{aligned} \tag{4.18}$$

Now making use of (4.11) and (4.13), it follows from (4.18) that the second term on the right-hand side of (4.15) is bounded as $k \to \infty$ which together with (4.17) implies (4.14).

Finally, regarding (4.5), note that, by (4.9), one can find an $\eta > 0$ for which

$$m_{k+s} - m_k = \phi_{\min}(J_{k+s}(\lambda)) - \phi_{\min}(J_k(\lambda)) \geq \eta \tag{4.19}$$

for any $k \geq k_1$. Define $l(b) = \max\{k \geq 1 : m_k < b\}$. It is clear that $l(b) > k_1$ for any sufficiently large b. Assuming first that $k_1 \leq k \leq l(b)$, let $j(k)$ equal the integer part of $(l(b) - k)/s$, where s is defined by (4.9). Since $k \leq l(b) - j(k)s$, one obtains from the monotonicity of the sequence $\{m_k, k \geq 1\}$ that $m_k \leq m_{l(b)-j(k)s}$. From this inequality and (4.19), for any $k_1 \leq k \leq l(b)$, we have $b - m_k \geq m_{l(b)} - m_k \geq m_{l(b)} - m_{l(b)-j(k)s} \geq \eta j(k)$. Therefore, noting that $j(k)$ takes on the values $0, 1, 2, \ldots$, it can be seen that $\sum_{k=k_1}^{l(b)} \min(1, (b - m_k)^{-2}) \leq 1 + s\eta^{-2} \sum_{j=1}^{\infty} j^{-2} < \infty$ uniformly in b. Next, considering $k > l(b)$ and defining $j(k)$ as the integer part of $(k - l(b) + 1)/s$, one can show similarly that, uniformly in b, $\sum_{k=l(b)+1}^{\infty} \min(1, (b - m_k)^{-2}) \leq 1 + s\eta^{-2} \sum_{j=1}^{\infty} j^{-2} < \infty$ which implies that (4.5) is satisfied. □

5. Sequential estimation in generalized linear models. This section discusses sequential maximum likelihood estimation in generalized linear models. Generalized linear models (GLMs) have been introduced by Nelder and Wedderburn (1972) as a useful class of statistical models based on the exponential family of distributions. Formally, the term "generalized linear models" refers to statistical models in which observations follow a distribution from an exponential family and the natural parameter of this family is a function of a linear combination of unknown parameters. Assume that a vector $Y_n' = (y_1, \ldots, y_n)$ of response variables is observed whose components are independent and their probability density functions $f_{y_k}(y; X_k, \gamma)$, $1 \leq k \leq n$, with respect to a σ-finite measure ν on the real line belong to an exponential family of distributions, that is,

$$f_{y_k}(y; X_k, \gamma) = \exp\{\varphi_k(y\theta_k - b(\theta_k))\}, \tag{5.1}$$

where φ_k is a weight, $b(\cdot)$ is some specified function. Further, $\theta_k = h(X_k'\gamma)$, where $h(\cdot)$ is an increasing function, X_k is an r-vector of explanatory variables (covariates) and γ is an r-vector of unknown parameters to be estimated from the data. In the context of exponential families, θ_k is referred to as a natural parameter. The link function $h(u)$ is said to be a natural link function if $h(u) = u$.

It is of interest to note that it is also common to define GLMs in terms of the mean of the dependent variable y_k. Let $\mu_k = Ey_k$, then μ_k is assumed to be connected with X_k and γ through a function g, that is, $g(\mu_k) = X_k'\gamma$. It can be shown that the two definitions of GLMs are equivalent.

It will be assumed throughout this section that the covariate vectors $X_k, k \geq 1$, in (5.1) are fixed or stochastically independent. The assumption of fixed covariates is reasonable in prospectively designed studies when data on human patients or laboratory animals are collected over a period of time. Random-covariate GLMs are relevant in retrospective studies in which data are obtained by randomly sampling from existing data sets. Examples of GLMs with fixed and random covariates will be provided later in this section.

The major tool for the analysis of GLMs is the method of maximum likelihood (ML) estimation. Nelder and Wedderburn (1972) proposed an efficient

version of the scoring algorithm to implement ML estimation in GLMs. Berk (1972) gave sufficient conditions for consistency and asymptotic normality of the ML estimate for exponential models in the i.i.d. case. Haberman (1977) considered ML estimation of a vector of natural parameters in a general exponential family of distributions. Fahrmeir and Kaufmann (1985) improved Haberman's results by proving asymptotic normality of the ML estimate under weaker conditions.

Since this paper discusses GLMs with random covariates, it is instructive to compare two approaches to the definition of the ML estimators in statistical models with random effects considered in the literature. The first one assumes the knowledge of the marginal distribution of random explanatory variables. This approach may be too restrictive in situations where little information on the distribution of the random effects is available. In this paper, we adopt an alternative approach based on conditional arguments. Regardless of the form of the marginal distribution of random explanatory variables, we can introduce the *conditional likelihood function* (given the random covariates) and compute the *conditional ML estimates* of the unknown parameters. Conditional ML estimation in logistic regression models is discussed in Prentice and Breslow (1978), Stefanski and Carroll (1985). Sequential sampling in the conditional likelihood framework is studied in Grambsch (1989), Chang and Martinsek (1992) and Chang (1995).

Let $\ell_n(\gamma)$ denote the log-likelihood function conditional on the random covariates. The expected and observed Fisher information matrices $J_n(\gamma)$ and $\widehat{J}_n(\gamma)$ are given by

$$
\begin{aligned}
J_n(\gamma) &= \sum_{k=1}^{n} \varphi_k E\left(\ddot{b}(\theta_k)\{\dot{h}(X_k'\gamma)\}^2 X_k X_k'\right), \\
\widehat{J}_n(\gamma) &= \sum_{k=1}^{n} \varphi_k \left(\ddot{b}(\theta_k)\{\dot{h}(X_k'\gamma)\}^2 - (y_k - \dot{b}(\theta_k))\ddot{h}(X_k'\gamma)\right) X_k X_k',
\end{aligned}
\tag{5.2}
$$

where γ is a parameter point, $\dot{b}(\theta)$ and $\ddot{b}(\theta)$ denote the first and second derivatives of $b(\theta)$, respectively. The normalized expected and observed information matrices $K_n(\gamma)$ and $\widehat{K}_n(\gamma)$ are defined as in (2.1) with $Q_n = J_n^{1/2}(\gamma_0)$, where $J_n^{1/2}(\gamma_0)$ denotes the left Cholesky root of the positive definite matrix $J_n(\gamma_0)$ [see Horn and Johnson (1985), pages 406–407] and γ_0 denotes the true value of the parameter γ. The role played by the Cholesky representation in multivariate statistical analysis was emphasized by Fahrmeir and Kaufmann (1985) and Fahrmeir (1987). Note that $K_n(\gamma) = I_r$ ($r \times r$ identity matrix).

In what follows, we will verify Conditions A–D of Section 2 for the generalized linear model (5.1). The simpler conditions will be utilized later in the sequential ML analysis of GLMs encountered in prospective and retrospective studies.

Note that the conditions developed by Haberman (1977) and Fahrmeir and Kaufmann (1985) cannot be used directly in this section since the analysis of the sequential ML estimates (which follows from Section 2) relies on the com-

plete convergence of the non-sequential ML estimate and observed information matrix. Thus, the results of Haberman (1977) and Fahrmeir and Kaufmann (1985) need to be extended and strengthened in order to verify conditions A–D for generalized linear models.

CONDITION A. Generally, Condition A can be verified by applying Chebyshev's inequality and requiring that, for some $s \geq 1$, $\sum_{n=1}^{\infty} E\|\dot{\ell}_n(\gamma_0)\|^s / \{\phi_{\min}(J_n(\gamma_0))\}^s < \infty$.

CONDITION B. Lemma 5.1 describes conditions under which $\widehat{K}_n(\gamma) \to I_r$ (completely) as $n \to \infty$ uniformly in a neighborhood of γ_0.

LEMMA 5.1. *Assume that* $\phi_{\min}^{-1}(J_n(\gamma_0))\|\widehat{J}_n(\gamma) - \widehat{J}_n(\gamma_0)\| \to 0$ (*completely*) *uniformly in a neighborhood of* γ_0 *and* $\phi_{\min}^{-1}(J_n(\gamma_0))\|\widehat{J}_n(\gamma_0) - J_n(\gamma_0)\| \to 0$ (*completely*) *as* $n \to \infty$. *Then* $\widehat{K}_n(\gamma) \to I_r$ (*completely*) *as* $n \to \infty$ *uniformly in a shrinking neighborhood of* γ_0.

PROOF. It follows from the definition of the normalized information matrix $\widehat{K}_n(\gamma)$ that $\|\widehat{K}_n(\gamma) - I_r\| \leq \|J_n^{-1/2}(\gamma_0)\|\|(J_n^{-1/2}(\gamma_0))'\|\|\widehat{J}_n(\gamma) - J_n(\gamma_0)\|$. Since

$$\|J_n^{-1/2}(\gamma_0)\|\|(J_n^{-1/2}(\gamma_0))'\| = \operatorname{tr} J_n^{-1}(\gamma_0) \leq r\phi_{\max}(J_n^{-1}(\gamma_0)) = r/\phi_{\min}(J_n(\gamma_0)),$$

one can easily see that

$$\|\widehat{K}_n(\gamma) - I_r\| \leq r\phi_{\min}^{-1}(J_n(\gamma_0))\{\|\widehat{J}_n(\gamma) - \widehat{J}_n(\gamma_0)\| + \|\widehat{J}_n(\gamma_0) - J_n(\gamma_0)\|\}$$

which, in view of the assumptions, completes the proof of Lemma 5.1. □

CONDITION C. The normalized score statistic $J_n^{-1/2}(\gamma_0)\dot{\ell}_n(\gamma_0)$ is a sum of independent random vectors with mean zero and a finite covariance matrix. Therefore the asymptotic normality of $J_n^{-1/2}(\gamma_0)\dot{\ell}_n(\gamma_0)$ follows from the central limit theorem for triangular arrays of independent random variables [see, e.g., Chow and Teicher (1988), Chapter 9].

LEMMA 5.2. *Let* $Z_{kn} = \varphi_k(y_k - \dot{b}(\theta_k))\dot{h}(X_k'\gamma_0)u'J_n^{-1/2}(\gamma_0)X_k$, $1 \leq k \leq n$, *where* $u \neq 0$ *is an arbitrary* r*-vector and assume that, for any* $\varepsilon > 0$, $\sum_{k=1}^{n} E(Z_{kn}^2 I\{|Z_{kn}| > \varepsilon\}) \to 0$ *as* $n \to \infty$ (*Lindeberg condition*). *Then Condition* C *is satisfied.*

PROOF. It follows from (5.2) that $u'J_n^{-1/2}(\gamma_0)\dot{\ell}_n(\gamma_0) = \sum_{k=1}^{n} Z_{kn}$. It is easy to check that $EZ_{kn} = 0$. Next, it follows from (5.1) that

$$Ey_k = \dot{b}(\theta_k), \quad \operatorname{Var} y_k = \ddot{b}(\theta_k)/\varphi_k. \tag{5.3}$$

Hence $\sum_{k=1}^{n} \operatorname{Var} Z_{kn} = u'u$. By the central limit theorem and Lindeberg condition, this immediately implies that $u'J_n^{-1/2}(\gamma_0)\dot{\ell}_n(\gamma_0)$ is asymptotically normal

$(0, u'u)$. Since $u \neq 0$ is an arbitrary r-vector, making use of Cramèr-Wold's theorem [see Billingsley (1968), page 49] completes the proof of Lemma 5.2. □

CONDITION D. The verification of Condition D is based on arguments similar to those employed in Anscombe's (1952) theorem on asymptotic normality of randomly stopped statistics and Gleser's (1969) extension of Anscombe's results to the multivariate case which can be stated as follows. Assume that D_n is a non-singular matrix and S_n is a random vector, $n \geq 1$. Also, assume that $D_n^{-1}S_n$ is asymptotically normal $(0, I)$, where I denotes the identity matrix. Then, for any non-decreasing sequence $\{t_n, n \geq 1\}$ of positive integers with $t_n \to \infty$ as $n \to \infty$, $D_{t_n}^{-1}S_{T_n}$ is asymptotically normal $(0, I)$ provided T_n is an integer-valued random variable (indexed by n) such that $T_n/t_n \xrightarrow{p} 1$ and, for any $\varepsilon > 0$ and some $\delta_\varepsilon > 0$,

$$\limsup_{n\to\infty} P\left\{\max_{-t_n\delta_\varepsilon \leq k \leq t_n\delta_\varepsilon} \|D_{t_n}^{-1}(S_{t_n+k} - S_{t_n})\| > \varepsilon\right\} \leq \varepsilon. \tag{5.4}$$

It is important to note that the preceding results assume that the normalizing matrix D_{t_n} is based on a fixed sample size t_n. This assumption appears to be restrictive since in most problems of sequential analysis the fixed sample size is replaced with a random sample size and one is typically interested in proving the asymptotic normality of $D_{T_n}^{-1}S_{T_n}$.

In order to extend the main results of Gleser (1969) to this more realistic setting, write $D_{T_n}^{-1}S_{T_n}$ as a product of $D_{T_n}^{-1}D_{t_n}$ and $D_{t_n}^{-1}S_{T_n}$. Then, because of the preceding results, $D_{t_n}^{-1}S_{T_n}$ is asymptotically normal. The next step is to determine conditions on the normalizing matrix D_n such that $D_{T_n}^{-1}D_{t_n} \xrightarrow{p} I$. Finally, one can apply Slutsky's theorem in order to establish the asymptotic normality of $D_{T_n}^{-1}S_{T_n}$.

Lemma 5.3 implements the outlined plan. It establishes conditions that need to be imposed on the normalizing matrix $J_n^{1/2}(\gamma_0)$ (which plays the role of D_n) in order to prove the asymptotic normality of $J_{T_n}^{-1/2}(\gamma_0)\dot{\ell}_{T_n}(\gamma_0)$. The assertion of Lemma 5.3 easily follows from a general result stated in the Appendix (see Lemma A.4) since $\dot{\ell}_n(\gamma_0)$ is a sum of independent random vectors with mean zero and covariance matrix $J_n(\gamma_0)$.

LEMMA 5.3. *Assume that Condition* C *holds and, for any* $\eta > 0$*, there exists a* $\delta_\eta > 0$ *for which* $tr\,(J_{[n(1+\delta_\eta)]}(\gamma_0) - J_n(\gamma_0)) \leq \eta\phi_{\min}(J_n(\gamma_0))$*. Then Condition* D *is satisfied.*

PROSPECTIVE STUDIES. In this subsection we will illustrate the theoretical results obtained earlier in the paper with two examples of particular practical importance. These examples represent two different classes of GLMs. First we will consider a binary regression model with fixed covariates and then the case of GLMs with random covariates.

As an example of a prospective study, consider a quantal biological assay or a dose finding study. The response variables $y_1, \ldots, y_n$ are independent

Bernoulli with the probabilities of success p_k, $1 \le k \le n$, which are modelled as a function of several covariates:

$$p_k = G(\gamma_1 x_{1k} + \gamma_2 x_{2k} + \cdots + \gamma_r x_{rk}) = G(X_k'\gamma), \tag{5.5}$$

where G is a specified distribution function, $X_k' = (x_{1k}, \ldots, x_{rk})$ is a nonrandom covariate vector for the kth observation (kth subject or experimental unit) and $\gamma = (\gamma_1, \ldots, \gamma_p)$ is an unknown parameter vector to be estimated from the data.

It is easy to show that the distribution function of y_k can be embedded into the exponential family (5.1) with

$$\varphi_k = 1, \quad b(\theta) = \ln(1 + e^\theta), \quad h(x) = \ln\{G(x)/(1 - G(x))\}. \tag{5.6}$$

The general binary response model (5.5) includes, as special cases, logit and probit regression models. Further examples of binary response models with fixed covariates used in educational and industrial testing are discussed in Wu (1985).

Estimation in binary responses models has been considered in various papers [see Govindarajulu (1988) for details and references]. The ML estimates of the parameters in one-factor models of the form $p_k = G(\gamma_1 + \gamma_2 x_k)$ were shown by Church and Cobb (1973) to be equivalent to the Spearman-Karber estimates if the covariate values x_k, $1 \le k \le n$, are equally spaced, number of observations at each x_k is the same for all k and $p_1 \le \cdots \le p_k$. The asymptotic behavior of the Spearman-Karber estimates was studied by Miller (1973) and sequential versions of the Spearman-Karber estimates were introduced by Nanthakumar and Govindarajulu (1994, 1999) and Govindarajulu and Nanthakumar (2000). The above-mentioned results will be extended in this section to the binary regression model (5.5) with several covariates. The proposed sequential procedure can be applied to practical problems when the selected sample of subjects is not homogeneous and covariates need to be introduced in the model in order to describe the variability in the responses.

It will be shown that the sequential ML estimate of the unknown parameter in the general binary response model (5.5) possesses the optimal properties described in Section 2. In order to prove that the fixed size confidence regions based on the sequential ML estimate are asymptotically consistent and efficient, it remains to verify Conditions A–D of Section 2 using the auxiliary results established earlier in this section.

The above-mentioned conditions will be verified under the assumption that the fixed covariate vectors have a compact range, that is, the X_k's belong to a compact subset of the r-dimensional space. This implies, for example, that $\|X_k\|$, $k \ge 1$, are uniformly bounded. The assumption of compact range has been used by Fahrmeir and Kaufmann (1985) in their asymptotic analysis of the ML estimates in GLMs.

This assumption is not restrictive and can be justified by the following observation. It can be seen from (5.6) that the Fisher information matrix $J_n(\gamma)$

for the model (5.5) is given by

$$J_n(\gamma) = \sum_{k=1}^{n} g^2(X_k'\gamma)\{G(X_k'\gamma)(1 - G(X_k'\gamma))\}^{-1} X_k X_k', \tag{5.7}$$

where $g(x) = dG(x)/dx$. Therefore, if $x^2 g^2(x) = o\{G(x)(1 - G(x))\}$ as $|x| \to \infty$, then the contribution of a covariate vector X_k with a large norm in terms of the added information is essentially negligible. This condition is satisfied in most problems of practical interest, for example, problems involving logit and probit regression models.

Let $\phi_{\min}(n) = \phi_{\min}(\sum_{k=1}^{n} X_k X_k')$ and $\phi_{\max}(n) = \phi_{\max}(\sum_{k=1}^{n} X_k X_k')$. Making use of the compact range assumption, it is easy to show that, in any compact neighborhood of the true parameter point γ_0,

$$C_1 \phi_{\min}(n) \le \phi_{\min}(J_n(\gamma)) \le \phi_{\max}(J_n(\gamma)) \le C_2 \phi_{\max}(n), \tag{5.8}$$

where C_1 and C_2 are some positive constants. Therefore the asymptotic behavior of $\phi_{\min}(J_n(\gamma))$ and $\phi_{\max}(J_n(\gamma))$ in a neighborhood of γ_0 can be controlled by imposing conditions on the minimum and maximum eigenvalues of $\sum_{k=1}^{n} X_k X_k'$. Inequalities similar to (5.8) will be frequently used below in the proof of Theorem 5.1.

Theorem 5.1 formulates conditions under which the sequential fixed size confidence ellipsoid $CR_{N(d)}$ has asymptotically the correct coverage probability and the sequential sample size $N(d)$ is asymptotically equivalent to the best fixed sample size $n(d)$ as the size of the confidence ellipsoid becomes small.

THEOREM 5.1. *Assume that the covariate vectors X_k, $k \ge 1$, have a compact range,*

$$\sum_{k=1}^{n} \|X_k\|^2 = O(\phi_{\min}(n)), \quad n \to \infty, \tag{5.9}$$

$$\sum_{n=1}^{\infty} \phi_{\min}^{-q}(n) < \infty \quad \text{for some } q > 0, \tag{5.10}$$

$$\text{for any } \varepsilon > 0 \text{ and some } \delta_\varepsilon > 0, \quad \phi_{\min}^{-1}(n) \sum_{k=n+1}^{n(1+\delta_\varepsilon)} \|X_k\|^2 \le \varepsilon \quad \text{for any large } n. \tag{5.11}$$

Assume also that $g(x)$ is twice continuously differentiable, $g(x) > 0$ for any x, and $\lim_{\rho \to 1} \lim_{d \to 0} \{n(d\rho)/n(d)\} = 1$. Then the sequential procedure is asymptotically equivalent to the optimal fixed-sample size procedure.

PROOF. In view of Theorem 3.1 and the assumption that

$$\lim_{\rho \to 1} \lim_{d \to 0} \{n(d\rho)/n(d)\} = 1,$$

it suffices to verify Conditions A–D of Section 2.

Starting with Condition A, note that the score statistic $\dot{\ell}_n(\gamma_0) = \sum_{k=1}^n (y_k - G(X_k'\gamma_0))w(X_k'\gamma_0)X_k$, where $w(x) = g(x)/\{G(x)(1 - G(x))\}$,is a sum of independent random vectors with mean zero. Since $|y_k - G(X_k'\gamma_0)| \leq 1$ and, by the assumption of compact range, $w(X_k'\gamma_0)$ is uniformly bounded (note that $w(\cdot)$ is a strictly positive function), an elementwise application of the Marcienkiewicz-Zygmund inequality yields

$$E\|\dot{\ell}_n(\gamma_0)\|^2 = O\left(\sum_{k=1}^n \|X_k\|^2\right). \tag{5.12}$$

Now, an application of Chebyshev's inequality with $s = 2$ along with (5.12), (5.8) and the assumption (5.9) implies that Condition A holds.

The verification of Condition B simply reduces to checking the assumptions of Lemma 5.1. By the definition of $\widehat{J}_n(\gamma)$ and (5.6),

$$\widehat{J}_n(\gamma) - \widehat{J}_n(\gamma_0) = \sum_{k=1}^n (T(y_k, X_k'\gamma) - T(y_k, X_k'\gamma_0))\, X_k X_k' \tag{5.13}$$

where

$$T(u, v) = (u - G(v))\frac{d}{dv}\left(\frac{g(v)}{G(v)(1 - G(v))}\right) - \frac{g^2(v)}{G(v)(1 - G(v))}. \tag{5.14}$$

By a Taylor series expansion in (5.13) along with the uniform boundedness of $(y_k - G(X_k'\gamma_0))$ and X_k, $\|\widehat{J}_n(\gamma) - \widehat{J}_n(\gamma_0)\| \leq C_3\|\gamma - \gamma_0\| \sum_{k=1}^n \|X_k\|^2$, where C_3 is a positive constant. Together with (5.8) and (5.9), this proves that the first assumption of Lemma 5.1 is satisfied. Concerning the second assumption, note that $\widehat{J}_n(\gamma_0) - J_n(\gamma_0) = \sum_{k=1}^n \left(T(y_k, X_k'\gamma_0) - ET(y_k, X_k'\gamma_0)\right) X_k X_k'$, where $T(u, v)$ is defined in (5.14), is a sum of independent random matrices. By the Marcienkiewicz-Zygmund inequality, the assumption that the X_k's belong to a compact set and (5.9), we have, for any $s \geq 1$,

$$\begin{aligned} \phi_{\min}^{-2s}(n)E\|\widehat{J}_n(\gamma_0) - J_n(\gamma_0)\|^s &= O\left(\phi_{\min}^{-2s}(n)\sum_{k=1}^n \|X_k\|^{2s}\right) \\ &= O\left(\sum_{n=1}^\infty \phi_{\min}^{1-2s}(n)\right). \end{aligned} \tag{5.15}$$

Applying Chebyshev's inequality with $s = (q + 1)/2$ and making use of (5.15) and (5.10) yields $\sum_{n=1}^\infty P\{\|\widehat{J}_n(\gamma_0) - J_n(\gamma_0)\| \leq \varepsilon\phi_{\min}(J_n(\gamma_0))\}$ $\leq C_4 \sum_{n=1}^\infty \phi_{\min}^{-q}(n) < \infty$, where C_4 is a positive constant. This proves that the second assumption of Lemma 5.1 is also satisfied and hence Condition B holds.

By Lemma 5.2, Condition C will follow if the

$$Z_{kn} = (y_k - G(X_k'\gamma_0))\frac{g(X_k'\gamma_0)}{G(X_k'\gamma_0)(1 - G(X_k'\gamma_0))}u' J_n^{-1/2}(\gamma_0)X_k, \quad 1 \leq k \leq n, \tag{5.16}$$

satisfy the Lindeberg condition with any r-vector $u \neq 0$. Noting that, for any $\delta > 0$, $E(Z_{kn}^2 I\{|Z_{kn}| > \varepsilon\}) \leq \varepsilon^{-\delta} E|Z_{kn}|^{2+\delta}$, it is easy to see from (5.16), the uniform boundedness of $(y_k - G(X_k'\gamma_0))$ and X_k and (5.8) that

$$\sum_{k=1}^{n} E(Z_{kn}^2 I\{|Z_{kn}| > \varepsilon\}) \leq C_5 \phi_{\min}^{-\delta/2}(n) \to 0, \qquad n \to \infty,$$

where C_5 is a positive constant, since, by (5.10), $\phi_{\min}(n) \to \infty$ as $n \to \infty$. Hence Condition C is satisfied.

Finally, Condition D can be verified with the aid of Lemma 5.3. Observe from the definition of the information matrix $J_n(\gamma_0)$ in (5.7) that the assumption of compact range implies that $0 \leq \operatorname{tr}(J_{[n(1+\delta)]}(\gamma_0) - J_n(\gamma_0)) \leq C_6 \sum_{k=n+1}^{n(1+\delta)} \|X_k\|^2$ and $C_7 \phi_{\min}(n) \leq \phi_{\min}(J_n(\gamma_0))$, where C_6 and C_7 are positive constants. For any fixed positive number η, let $\varepsilon = \eta C_7/C_6$. For the selected $\varepsilon > 0$, one can now find a $\delta > 0$ such that (5.11) holds. In view of the inequalities given above, (5.11) yields, for the chosen δ, $\operatorname{tr}\{J_{[n(1+\delta)]}(\gamma_0) - J_n(\gamma_0)\} \leq \eta \phi_{\min}(J_n(\gamma_0))$ and the condition of Lemma 5.3 is satisfied implying that Condition D is satisfied. This completes the proof of Theorem 5.1. □

The conditions imposed in Theorem 5.1 on the binary regression model (5.5) with fixed covariates ensure that the proposed sequential procedure is asymptotically efficient in the sense that the ratio $EN(d)/n(d)$ approaches 1 as the confidence region becomes small. This property of the sequential procedure can be further strengthened and it can be shown that the sequential procedure possesses the property of bounded cost of ignorance.

Sufficient conditions under which the sequential procedure associated with the stopping rule (3.2) possesses the property of bounded cost of ignorance are given below in Theorem 5.2.

THEOREM 5.2. *Under the assumptions of Theorem* 5.1 *and*

$$\liminf_{k \to \infty} \phi_{\min}\left(\sum_{i=k+1}^{k+s} X_i X_i'\right) > 0 \qquad \text{for some integer } s \geq 1, \tag{5.17}$$

the sequential procedure has a bounded cost of ignorance.

PROOF. The assertion of the theorem will follow if one verifies the conditions of Theorem 4.1.

The first condition, namely (4.9), is a straightforward consequence of (5.17). From the definition of $J_k(\gamma)$ and the assumption of compact range we obtain that

$$\begin{aligned} \phi_{\min}(J_{k+s}(\gamma)) - \phi_{\min}(J_k(\gamma)) &\geq \phi_{\min}(J_{k+s}(\gamma) - J_k(\gamma)) \\ &\geq C \phi_{\min}\left(\sum_{i=k+1}^{k+s} X_i X_i'\right), \end{aligned} \tag{5.18}$$

for sufficiently large k, where C is some positive constant.

Next, it was shown in Theorem 5.1 [see display (5.15)] that $E\|\widehat{J}_k(\gamma_0) - J_k(\gamma_0)\|^2$ is of the same order of magnitude as $\sum_{i=1}^k \|X_i\|^2$. Now, recalling that the X_i's are uniformly bounded and tr $\{\text{Cov } \widehat{J}_k(\gamma_0)\} = E\|\widehat{J}_k(\gamma_0) - J_k(\gamma_0)\|^2$, one obtains that (4.10) is satisfied.

In order to verify (4.11) it suffices to show that $k^2 E\|\widehat{\gamma}_k - \gamma_0\|^4 = O(1)$, $k \to \infty$, since, by (5.8) and the assumption of compact range, $\phi_{\max}(J_k(\gamma_0)) \leq \text{tr } J_k(\gamma_0) = O(k)$.

The proof that $k^2 E\|\widehat{\gamma}_k - \gamma_0\|^4 = O(1)$ follow from a result of Ibragimov and Hasminskii (1981). The observations y_i, $i \geq 1$, in the general binary regression model (5.5) are independent and therefore, by Ibragimov and Hasminskii [(1981), Theorem 1.5.8],

$$P\{k^{1/2} E\|\widehat{\gamma}_k - \gamma_0\| > t\} = O(e^{-t^2}), \qquad k \to \infty, \tag{5.19}$$

provided

$$\sup_{i \geq 1} \int_{-\infty}^{\infty} \sup_{\gamma} \frac{\{f_{y_i}^{1/2}(y; X_i, \gamma) - f_{y_i}^{1/2}(y; X_i, \gamma_0)\}^2}{\|\gamma - \gamma_0\|^2} d\nu(y) < \infty \tag{5.20}$$

and, for any large k,

$$\sum_{i=1}^{k} H_i(\gamma, \gamma_0) \geq k c(\gamma_0) \|\gamma - \gamma_0\|^2 / \{1 + \|\gamma - \gamma_0\|^2\}, \tag{5.21}$$

where $f_{y_i}(y; X_i, \gamma)$ is the probability density function of y_i, ν is the supporting measure, $c(\gamma)$ is some function of γ. Further, $H_i(\gamma, \gamma_0)$ is the Hellinger distance for the ith observation given by $H_i(\gamma, \gamma_0) = \int_{-\infty}^{\infty} \{f_{y_i}^{1/2}(y; X_i, \gamma) - f_{y_i}^{1/2}(y; X_i, \gamma_0)\}^2 d\nu(y)$. It can be seen that (5.19) implies that $k^2 E\|\widehat{\gamma}_k - \gamma_0\|^4 = O(1)$ and therefore it remains to verify (5.20) and (5.21) for the binary regression model (5.5) with uniformly bounded covariates.

The supporting measure in (5.20) can be chosen to assign the same weight of one to the two points 0, 1 and zero to any other point on the real line. Then, for any γ, $f_{y_i}(0; X_i, \gamma) = 1 - G(X_i'\gamma)$ and $f_{y_i}(1; X_i, \gamma) = G(X_i'\gamma)$. Since, by Taylor's expansion and the assumption of compact range,

$$\begin{aligned} &\sup_{i \geq 1} \sup_{\gamma} \{f_{y_i}^{1/2}(0; X_i, \gamma) - f_{y_i}^{1/2}(0; X_i, \gamma_0)\}^2 \\ &\quad \leq \sup_{\zeta} \left(\frac{g^2(X_i'\zeta)}{4(1 - G(X_i'\zeta))} \right) \|X_i\|^2 \|\gamma - \gamma_0\|^2 \\ &\quad \leq C_1 \|\gamma - \gamma_0\|^2, \end{aligned}$$

and, similarly, $\sup_{i \geq 1} \sup_{\gamma} \{f_{y_i}^{1/2}(1; X_i, \gamma) - f_{y_i}^{1/2}(1; X_i, \gamma_0)\}^2 \leq C_2 \|\gamma - \gamma_0\|^2$, where C_1 and C_2 are constants, the condition (5.20) is satisfied.

Regarding (5.21), note that, by Ibragimov and Hasminskii [(1981), Theorem 1.7.6], for some positive constant C_3,

$$\sum_{i=1}^{k} H_i(\gamma, \gamma_0) \geq C_3 \phi_{\min}(J_k(\gamma_0)) \|\gamma - \gamma_0\|^2 / (1 + \|\gamma - \gamma_0\|^2).$$

It follows from (5.18) and (5.17) that $\liminf_{k\to\infty} k^{-1}\phi_{\min}(J_k(\gamma_0)) > 0$ and hence (5.21) holds for any sufficiently large k. Together with (5.20), this implies (5.19) and therefore (4.11) follows.

The last two conditions of Theorem 4.1, namely (4.12) and (4.13), can be verified along the lines of Lemma 5.1. Making use of the inequalities utilized in this lemma and (5.15), one can show that

$$\|\widehat{K}_k(\gamma) - \widehat{K}_k(\gamma_0)\| \le \|\gamma - \gamma_0\| a_k, \tag{5.22}$$

where $a_k = rC_4\phi_{\min}^{-1}(J_k(\gamma_0))\sum_{i=1}^{k}\|X_i\|^2$. Note that $\{a_k, k \ge 1\}$ is a nonrandom sequence. By the assumption of compact range, (5.18) and the fact that $\phi_{\max}(J_k(\gamma_0)) = O(k)$, $k^{-1}\phi_{\max}(J_k(\gamma_0))a_k^2 = O(a_k^2) = O(k^2\phi_{\min}^{-2}(J_k(\gamma_0))) = O(1)$ as $k \to \infty$. In view of (5.22), this implies that (4.12) and (4.13) hold and thus completes the proof of Theorem 5.2. □

RETROSPECTIVE STUDIES. In retrospective studies subjects are sampled from a population under consideration to relate response variables to certain demographic variables or exposures to suspected risk factors. These explanatory variables are regarded as random effects since the data are collected by random sampling. Two particular cases of the generalized linear model arise naturally in the context of retrospective studies. These are binary regression and Poisson regression models. In this subsection, we will consider a general model including the binary and Poisson regression models discussed above as special cases. Given n pairs of observations $(y_1, X_1), \ldots, (y_n, X_n)$, assume that each pair satisfies the generalized linear model (5.1) with $\varphi_k = 1$, $k \ge 1$ and the covariates $X_1, \ldots, X_n$ are i.i.d. random vectors.

Chang and Martinsek (1992) and Chang (1995) considered similar problems of ML estimation of the parameters of binary response models when the marginal distribution of random covariates is unknown. It was shown that, under the assumption that the random covariates are i.i.d., the sequential ML estimation procedures for setting up fixed size confidence regions are asymptotically consistent and efficient. Here we will extend these results to the generalized linear model (5) with i.i.d. random covariates. We will also strengthen the results of Chang and Martinsek (1992) by showing that in the case of a logistic regression model the sequential procedure possesses an important property of bounded cost of ignorance.

It will be shown in the following theorem that the sequential procedure described in Section 2 is asymptotically equivalent to the optimal fixed-sample size procedure. Since the best fixed sample size required in this problem is $n(d) = [\chi^2_{r,1-\alpha}/d^2\phi]$, where $\phi = \phi_{\min}(J(\gamma))$ and $J(\gamma)$ is the information matrix based on a single observation (y_1, X_1), the condition (3.4) on $n(d)$ is trivially satisfied.

To simplify the notation, for any parameter point γ, let

$$\begin{aligned} V_k(\gamma) = (&\ddot{b}(h(X_k'\gamma))\{\dot{h}(X_k'\gamma)\}^2 \\ &-\{y_k - \dot{b}(h(X_k'\gamma))\}\ddot{h}(X_k'\gamma))X_kX_k', \qquad k \ge 1. \end{aligned} \tag{5.23}$$

Note that $V_k(\gamma)$ is the observed information matrix of the kth observation. It is easy to see that these matrices are independent and identically distributed.

THEOREM 5.3. *Assume that*

$$\|EV_1(\gamma_0)\| < \infty, \qquad \|\operatorname{Cov} V_1(\gamma_0)\| < \infty, \tag{5.24}$$

$$E\left(\sup_{\|\gamma-\gamma_0\|\le\delta} \|V_1(\gamma) - V_1(\gamma_0)\|/\|\gamma-\gamma_0\|\right)^2 < \infty, \qquad \delta > 0. \tag{5.25}$$

Then the sequential procedure is asymptotically equivalent to the optimal fixed-sample size procedure.

PROOF. The proof is similar to that of Theorem 5.1 in that it is also based on the successive verification of Conditions A–D of Section 2.

Condition A follows from a result of Hsu and Robbins (1947) provided

$$E(\ddot{b}(h(X_1'\gamma_0))\{\dot{h}(X_1'\gamma_0)\}^2\|X_1\|^2) < \infty$$

which is equivalent to the first condition in (5.24). Condition B can be verified by using Lemma 5.1. Concerning the first condition of this theorem, note that, by the definition of $V_k(\gamma)$,

$$\phi_{\min}^{-1}(J_n(\gamma_0)) \sup_{\|\gamma-\gamma_0\|\le\delta} \|\hat{J}_n(\gamma) - \hat{J}_n(\gamma_0)\| \le \delta\phi_{\min}^{-1}(J(\gamma_0))\tau_n,$$

where $\tau_n = n^{-1}\sum_{k=1}^n \sup_{\|\gamma-\gamma_0\|\le\delta} \|V_k(\gamma) - V_k(\gamma_0)\|/\|\gamma-\gamma_0\|$. Since τ_n is an average of i.i.d. random variables with a finite second moment (see (5.25)), it follows from Hsu and Robbins (1947) that τ_n converges completely to a constant and therefore the first condition of Lemma 5.1 is satisfied. The second condition of Lemma 5.1 follows from Hsu and Robbins (1947) since (5.24) holds. Next, in view of (5.24), Condition C is implied by the central limit theorem for i.i.d. random variables. Finally, applying Lemma 5.3 and noting that $\operatorname{tr}(J_{[n(1+\delta)]}(\gamma_0) - J_n(\gamma_0)) = n\delta \operatorname{tr} J(\gamma_0)$ can always be made smaller than $\eta\phi_{\min}(J_n(\gamma_0)) = \eta n\phi_{\min}(J(\gamma_0))$ for any large n by choosing $0 < \delta < \eta\phi_{\min}(J(\gamma_0))/\operatorname{tr} J(\gamma_0)$, one can conclude that Condition D is also satisfied. The proof of Theorem 5.3 is now complete. □

REMARK 5.1. Since $V_k(\gamma)$ is the observed information matrix of the kth pair (y_k, X_k), Condition (5.24) in Theorem 5.3 translates into the assumption that the observed information matrix of each individual observation has two finite moments. Further, (5.25) plays the same role as the classical univariate Cramèr's assumption that, for any parameter point γ, $|\partial^3 \ln f(z;\gamma)/\partial\gamma^3| \le H(z)$ and $EH^m(z) < \infty$ for some $m \ge 1$.

The following example shows that the conditions of Theorem 5.3 are satisfied in most widely used models with random covariates.

EXAMPLE 5.1. Generalized linear models are most frequently used in practice under the assumption of natural link, that is, $h(u) = u$. This implies that the individual observed information matrices $V_k(\gamma)$, $k \geq 1$, defined in (5.23) take the following very simple form: $V_k(\gamma) = \ddot{b}(X_k'\gamma)X_kX_k'$. It is easy to see that

$$\|EV_1(\gamma_0)\| = E(\ddot{b}(X_1'\gamma_0)\|X_1\|^2). \tag{5.26}$$

Next, let $A = \text{Cov } V_1(\gamma_0)$ and $B = E\{V_1(\gamma_0)V_1(\gamma_0)'\}$ (to simplify notation). Then $A = B - (EV_1(\gamma_0))(EV_1(\gamma_0))'$ and $(EV_1(\gamma_0))(EV_1(\gamma_0))'$ is a nonnegative definite matrix. Therefore, the eigenvalues of B dominate those of A. Making use of this fact, it can be seen that $\|A\| \leq \sqrt{r}\phi_{\max}(A) \leq \sqrt{r}\phi_{\max}(B) \leq \sqrt{r}\|B\|$. Hence $\|\text{Cov } V_1(\gamma_0)\| \leq \sqrt{r}\|E\{V_1(\gamma_0)V_1(\gamma_0)'\}\| = \sqrt{r}E(\{\ddot{b}(X_1'\gamma_0)\}^2 \times \|X_1\|^4)$ which together with (5.26) implies that the first two conditions of Theorem 5.3 are satisfied provided

$$E(\{\ddot{b}(X_1'\gamma_0)\}^2\|X_1\|^4) < \infty. \tag{5.27}$$

The verification of the last condition of Theorem 5.3 relies on a Taylor series expansion:

$$\begin{aligned} &E\left(\sup_{\|\gamma-\gamma_0\|\leq\delta}\|V_1(\gamma) - V_1(\gamma_0)\|/\|\gamma-\gamma_0\|\right)^2 \\ &\quad\leq E\left(\|X_1\|^3 \sup_{\|\zeta-\gamma_0\|\leq\delta}\beta(X_1'\zeta)\right)^2, \end{aligned} \tag{5.28}$$

where $\beta(u) = d^3b(u)/du^3$ and ζ is a parameter point lying on the line connecting γ and γ_0. One can now conclude that the assertions of Theorem 5.3 hold in a GLM based on a natural link if (5.27) is satisfied and the right-hand side of (5.28) is finite.

Consider a binary regression model with random covariates. Under the assumption of natural link, $G(x)$ becomes the standard logistic distribution function, that is, $G(x) = e^x/(1+e^x)$. Assume that $E\|X_1\|^6 < \infty$. It follows from (5.6) that $b(\theta) = \ln(1+e^\theta)$ and it is easy to verify that

$$\ddot{b}(\theta) = e^\theta(1+e^\theta)^{-2} \quad \text{and} \quad \beta(\theta) = e^\theta(1-e^\theta)(1+e^\theta)^{-3} \tag{5.29}$$

are bounded over the real line. Therefore (5.27) and (5.28) are finite implying that the sequential procedure for setting up a fixed size confidence region is asymptotically consistent and efficient for the logistic regression model.

Next, assume that the data collected from a retrospective study are modelled using a Poisson distribution. This distribution generates a generalized linear model with $b(\theta) = e^\theta$. Models of this type with the natural link function are known as log-linear models. Since $b(\theta) = e^\theta$ is an increasing function of θ which dominates any power function, observe from the Cauchy inequality that, for some positive constant C_1, $E(\{\ddot{b}(X_1'\gamma_0)\}^2\|X_1\|^4) \leq C_1Ee^{2\|\gamma_0\|\|X_1\|}$ and (5.27) follows from the assumption that $Ee^{C_2\|X_1\|} < \infty$ for any constant C_2.

Similarly, one can see that the same condition implies that the left-hand side of (5.28) is finite since

$$E\left(\|X_1\|^3 \sup_{\|\zeta-\gamma_0\|\le\delta} \beta(X_1'\zeta)\right)^2 \le E\left(\|X_1\|^3 e^{(\|\gamma_0\|+\delta)\|X_1\|}\right)^2$$

and $\|X_1\|^3 \le C_3 e^{(\|\gamma_0\|+\delta)\|X_1\|}$ for some constant C_3. The condition $Ee^{C_2\|X_1\|} < \infty$, $C_2 > 0$, is obviously satisfied if the tails of the distribution of $\|X_1\|$ converge to 0 faster than an exponential rate.

Chang and Martinsek (1992) have established the asymptotic efficiency of the sequential procedure for constructing a fixed size confidence region in the logistic regression case. Their result can be generalized by establishing the property of bounded cost of ignorance, that is, by proving that $EN(d)-n(d)$ remains bounded as the size of the confidence region becomes small. For logistic regression with i.i.d. random covariates, we have the following theorem towards bounded cost of ignorance. The proof of Theorem 5.4 can be carried out using arguments similar to those employed in the proof of Theorem 5.2 and is omitted.

THEOREM 5.4. *Assume that* $E\|X_1\|^8 < \infty$. *Then the sequential procedure has bounded cost of ignorance.*

APPENDIX

This section contains the statements and proofs of some of the results used in this paper that are somewhat technical in nature.

The following lemma is a variation of a result in matrix analysis [see, e.g., Roy (1957), pages 142–143]: If A and B are positive definite $r \times r$ matrices, then, for any r-vector x, $\phi_{\min}(AB^{-1}) \le x'Ax/x'Bx \le \phi_{\max}(AB^{-1})$. Lemma A.1 shows how a double inequality of this type can be utilized in establishing the closeness of $\phi_{\min}(A)/\phi_{\min}(B)$ to 1.

LEMMA A.1. *Let* $\{A_k, k \ge 1\}$ *and* $\{B_k, k \ge 1\}$ *be two sequences of positive definite* $r \times r$ *matrices and let* a_k *and* b_k *denote their minimum eigenvalues, respectively. Also let* $\widetilde{A}_k = D_k^{-1}A_k(D_k')^{-1}$ *and* $\widetilde{B}_k = D_k^{-1}B_k(D_k')^{-1}$, *where* $\{D_k, k \ge 1\}$ *is a sequence of non-singular* $r \times r$ *matrices. Then*

$$|a_k/b_k - 1| \le \delta_k = \|\widetilde{A}_k - \widetilde{B}_k\|/\phi_{\min}(\widetilde{B}_k). \tag{A.1}$$

Consequently, if, for some positive definite $r \times r$ *matrix* D,

$$\widetilde{A}_k \to D, \quad \widetilde{B}_k \to D, \quad k \to \infty, \tag{A.2}$$

then $a_k/b_k \to 1$, $k \to \infty$.

PROOF. First, consider the function $s_k(x) = x'A_kx/x'B_kx$, where x is an r-vector. Let $y = D_k'x$. Since D_k is a non-singular matrix, $s_k(x)$ takes the same set of values as $S_k(y) = y'\widetilde{A}_ky/y'\widetilde{B}_ky$ and therefore $\inf_y S_k(y) \le s_k(x) \le \sup_y S_k(y)$. Now assume without loss of generality that $\|y\| = 1$ and let $\delta_k = \|\widetilde{A}_k - \widetilde{B}_k\|/\phi_{\min}(\widetilde{B}_k)$. Then, utilizing the fact that $\phi_{\min}(\widetilde{B}_k) \le y'\widetilde{B}_ky$, we obtain that $|S_k(y) - 1| = |y'(\widetilde{A}_k - \widetilde{B}_k)y|/y'\widetilde{B}_ky \le \delta_k$. Together with $\inf_y S_k(y) \le s_k(x) \le \sup_y S_k(y)$ this implies, for any r-vector x with $\|x\| = 1$ and $k \ge 1$,

$$1 - \delta_k \le x'A_kx/x'B_kx \le 1 + \delta_k. \tag{A.3}$$

Now fix any $k \ge 1$ and let x_k denote a normalized eigenvector associated with the eigenvalue a_k of the matrix A_k. Then $a_k = x_k'A_kx_k$. By the definition of the minimum eigenvalue, $b_k \le x_k'B_kx_k$. Therefore, in view of (A.3), one obtains that

$$b_k \le x_k'B_kx_k \le (1 - \delta_k)^{-1}x_k'A_kx_k = (1 - \delta_k)^{-1}a_k.$$

On the other hand, letting x_k denote a normalized eigenvector associated with b_k one can easily infer from (A.3) that $a_k \le x_k'A_kx_k \le (1 + \delta_k)x_k'B_kx_k = (1 + \delta_k)b_k$. Thus, $|a_k/b_k - 1| \le \delta_k = \|\widetilde{A}_k - \widetilde{B}_k\|/\phi_{\min}(\widetilde{B}_k)$ which proves (A.1).

If Assumption (A.2) is also satisfied then $\|\widetilde{A}_k - \widetilde{B}_k\| \to 0$ and $\phi_{\min}(\widetilde{B}_k) \to \phi_{\min}(D) > 0$. By (A.1), this readily implies that $a_k/b_k \to 1$, $k \to \infty$. □

LEMMA A.2. *Let $\widehat{\lambda}_n$ denote the ML estimate of λ_0. Assume that Conditions* A *and* B *hold. Then $\widehat{\lambda}_n$ converges completely to λ_0 as $n \to \infty$.*

PROOF. The proof is an extension of the proof of the strong consistency of the ML estimate of the parameter in a generalized linear model given in Fahrmeir and Kaufmann [(1985), Theorem 2; see also Fahrmeir and Kaufmann (1986)].

Fix any $\delta > 0$ and note that the inequality $\ell_k(Z_k; \lambda) - \ell_k(Z_k; \lambda_0) \le 0$ (for any λ such that $\|\lambda - \lambda_0\| = \delta$) implies that there exists a local maximum inside the sphere $\{\lambda \in R^r : \|\lambda - \lambda_0\| \le \delta\}$ and therefore $\|\widehat{\lambda}_k - \lambda_0\| \le \delta$.

For the same δ, consider any λ with $\|\lambda - \lambda_0\| = \delta$ and let $u = (\lambda - \lambda_0)/\delta$. It is clear that $\|u\| = 1$. Now expand the log-likelihood function $\ell_k(Z_k; \lambda)$ in a Taylor series around λ_0:

$$\ell_k(Z_k; \lambda) - \ell_k(Z_k; \lambda_0) = \delta u'\dot{\ell}_k(Z_k; \lambda_0) - (\delta^2/2)u'\widehat{J}_k(\widetilde{\lambda})u, \tag{A.4}$$

where $\widetilde{\lambda}$ lies between λ and λ_0. Define the sets

$$A_k = \left\{ \|\dot{\ell}_k(Z_k; \lambda_0)\| \le (\delta/2) \inf_{\|\lambda - \lambda_0\| \le \delta} \phi_{\min}(\widehat{J}_k(\lambda)) \right\},$$

$$B_k = \left\{ \sup_{\|\lambda - \lambda_0\| \le \delta} \|\widehat{K}_k(\lambda) - K_k(\lambda_0)\| \le \varepsilon\phi_{\min}(K_k(\lambda_0)) \right\}, \qquad 0 < \varepsilon < 1.$$

Note that $u'\widehat{J}_k(\widetilde{\lambda})u \geq \phi_{\min}(\widehat{J}_k(\widetilde{\lambda})) \geq \inf_{\|\lambda-\lambda_0\|\leq\delta} \phi_{\min}(\widehat{J}_k(\lambda))$ and $|u'\dot{\ell}_k(Z_k;\lambda_0)| \leq \|\dot{\ell}_k(Z_k;\lambda_0)\|$ since $\|u\| = 1$. Therefore, by (A.4), $A_k \subset \{\ell_k(Z_k;\lambda) - \ell_k(Z_k;\lambda_0) \leq 0\}$ and

$$\sum_{k=1}^{\infty} P\{\|\widehat{\lambda}_k - \lambda_0\| > \delta\} \leq \sum_{k=1}^{\infty} P\{A_k^c\}. \tag{A.5}$$

In order to complete the proof of the lemma, it suffices to show that $\sum P\{A_k^c\} < \infty$. By Lemma A.1 with $A_k = \widehat{J}_k(\lambda)$, $B_k = J_k(\lambda_0)$ and $D_k = Q_k$, for any λ,

$$\phi_{\min}(\widehat{J}_k(\lambda)) \geq \phi_{\min}(J_k(\lambda_0))\{1 - \|\widehat{K}_k(\lambda) - K_k(\lambda_0)\|/\phi_{\min}(K_k(\lambda_0))\}.$$

Hence, on the set B_k, $\inf_{\|\lambda-\lambda_0\|\leq\delta} \phi_{\min}(\widehat{J}_k(\lambda)) \geq (1-\varepsilon)\phi_{\min}(J_k(\lambda_0))$. Therefore

$$\{A_k^c \cap B_k\} \subset \{\|\dot{\ell}_k(Z_k;\lambda_0)\| > (\delta/2)(1-\varepsilon)\phi_{\min}(J_k(\lambda_0))\}.$$

Thus, using the inequality $P\{A_k^c\} \leq P\{A_k^c \cap B_k\} + P\{B_k^c\}$, we obtain

$$\sum_{k=1}^{\infty} P\{A_k^c\} \leq \sum_{k=1}^{\infty} P\{\|\dot{\ell}_k(Z_k;\lambda_0)\| > (\delta/2)(1-\varepsilon)\phi_{\min}(J_k(\lambda_0))\} + \sum_{k=1}^{\infty} P\{B_k^c\}.$$

By Condition A, $\|\dot{\ell}_k(Z_k;\lambda_0)\|/\phi_{\min}(J_k(\lambda_0)) \to 0$ completely. Also, $\sum P\{B_k^c\} < \infty$ since, by Condition B, $\|\widehat{K}_k(\lambda) - K_k(\lambda_0)\| \to 0$ completely in some neighborhood of λ_0 and $\phi_{\min}(K_k(\lambda_0)) \to \phi_{\min}(K(\lambda_0)) > 0$. Therefore, $\sum P\{A_k^c\} < \infty$ and $\widehat{\lambda}_n$ converges completely to λ_0 in view of (A.5). □

LEMMA A.3. *Let $\{U_k, k \geq 1\}$ be a sequence of positive random variables and $\{m_k, k \geq 1\}$ a sequence of positive real numbers such that $m_1 \leq m_2 \leq m_3 \cdots \to \infty$, $U_k/m_k \to 1$ a.s. as $k \to \infty$. For any $b > 0$, let $T(b) = \inf\{k \geq 1 : U_k \geq b\}$, $t(b) = \inf\{k \geq 1 : m_k \geq b\}$ and assume that*

$$\lim_{\rho\to 1} \lim_{b\to\infty} \{t(b\rho)/t(b)\} = 1. \tag{A.6}$$

Then, as $b \to \infty$, $T(b)/t(b) \to 1$ a.s. and $ET(b)/t(b) \to 1$ if, for some $\delta > 0$, $\sum_{k=1}^{\infty} P\{U_k < \delta m_k\} < \infty$.

PROOF. In order to establish that $T(b)/t(b) \to 1$ a.s., let $A(k_0, \eta) = \{|U_k - m_k| \leq \eta m_k$ for all $k \geq k_0\}$, where $\eta > 0$ and $k_0 \geq 1$. It is easy to see that on this event, $t(b/(1+\eta)) \leq T(b) \leq t(b/(1-\eta))$ for all $b \geq b_0$, where $b_0 \to \infty$ as $k_0 \to \infty$. By (A.6), for any $\eta > 0$, there exists $\eta' > 0$ such that

$$(1-\eta')t(b) \leq t(b/(1+\eta)) \text{ and } t(b/(1-\eta)) \leq (1+\eta')t(b) \text{ for all } b \geq b_0.$$

Therefore, $A(k_0, \eta) \subset \{|T(b)/t(b) - 1| \leq \eta'$ for all $b \geq b_0\}$ and hence $U_k/m_k \to 1$ a.s. implies $P\{A(k_0, \eta)\} \to 1$ as $k_0 \to \infty$ which in turn implies that $T(b)/t(b) \to 1$ a.s.

Next, $T(b)/t(b) \to 1$ a.s. will imply $ET(b)/t(b) \to 1$ if $\{T(b)/t(b),\ b \geq 1\}$ is a uniformly integrable family of random variables. Take the value of δ for which $\sum_{k=1}^{\infty} P\{U_k < \delta m_k\} < \infty$ and note that

$$T(b) \leq I\{T(b) = 1\} + \sum_{k=2}^{\infty} I\{T(b) = k, U_{k-1} \geq \delta m_{k-1}\} + \sum_{k=1}^{\infty} I\{U_k < \delta m_k\}. \tag{A.7}$$

By the definition of the stopping rule $T(b)$, $\{T(b) = k\} \subset \{U_{k-1} < b\}$. Since $\{m_k, k \geq 1\}$ is an increasing sequence of positive numbers,

$$\sum_{k=2}^{\infty} I\{T(b) = k, \quad U_{k-1} \geq \delta m_{k-1}\} \leq \sum_{k=1}^{\infty} I\{m_k < b/\delta\} \leq t(b/\delta).$$

From (A.6), $t(b/\delta) \leq Ct(b)$ for some C as $b \to \infty$. Therefore, it follows from (A.7) that, for sufficiently large b, $T(b)/t(b) \leq 1 + C + \sum_{k=1}^{\infty} I\{U_k < \delta m_k\}$. In view of the assumption $\sum_{k=1}^{\infty} P\{U_k < \delta m_k\} < \infty$, it now implies that $\{T(b)/t(b),\ b \geq 1\}$ is uniformly integrable and therefore $ET(b)/t(b) \to 1$. □

LEMMA A.4. *Suppose that $Z_1, \ldots, Z_n$ are independent random r-vectors with mean zero and $E\|Z_k\|^2 < \infty$. Let $S_n = \sum_{k=1}^{n} Z_k$ and $V_n = Cov\ S_n$. Assume also that $V_n^{-1/2} S_n \xrightarrow{d} N(0, I_r)$ as $n \to \infty$, where $V_n^{1/2}$ is the left Cholesky square root of V_n, and, for any $\eta > 0$, there exists $\delta_\eta > 0$ for which*

$$\operatorname{tr}\left(V_{[n(1+\delta_\eta)]} - V_n\right) \leq \eta \phi_{\min}(V_n). \tag{A.8}$$

Then, for any regular sequence of integer-valued random variables $\{T_n, n \geq 1\}$, $V_{T_n}^{-1/2} S_{T_n}$ is asymptotically normal $(0, I_r)$, $n \to \infty$.

PROOF. It is easy to see that the asymptotic normality of $V_{T_n}^{-1/2} S_{T_n}$ is implied by

$$V_{t_n}^{-1/2} S_{T_n} \xrightarrow{d} N(0, I), \qquad V_{T_n}^{-1/2} V_{t_n}^{1/2} \xrightarrow{p} I_r. \tag{A.9}$$

Concerning the first relation in (A.9), it follows from the assumptions of the lemma that $T_n/t_n \xrightarrow{p} 1$ for some sequence of positive integers $\{t_n, n \geq 1\}$. Therefore it suffices to verify condition (5.4). Consider the cases $0 \leq k \leq t_n\delta$ and $-t_n\delta \leq k \leq 0$ separately. Applying Kolmogorov's inequality elementwise and making use of the inequality

$$\|V_{t_n}^{-1/2}\|^2 = \operatorname{tr}\ V_{t_n}^{-1} \leq r\phi_{\max}(V_{t_n}^{-1}) = r/\phi_{\min}(V_{t_n}), \tag{A.10}$$

it can be seen that

$$P\left\{\max_{0\le k\le t_n\delta}\|V_{t_n}^{-1/2}(S_{t_n+k}-S_{t_n})\|>\varepsilon\right\}\le (r/\varepsilon^2)\|V_{t_n}^{-1/2}\|^2\sum_{i=t_n+1}^{[t_n(1+\delta)]}E\|Z_i\|^2$$

$$=(r/\varepsilon^2)\|V_{t_n}^{-1/2}\|^2\ \mathrm{tr}\ (V_{[t_n(1+\delta)]}-V_{t_n}) \tag{A.11}$$

$$\le (r/\varepsilon)^2\ \mathrm{tr}\ (V_{[t_n(1+\delta)]}-V_{t_n})/\phi_{\min}(V_{t_n}).$$

Now utilizing the same argument when $-t_n\delta\le k\le 0$ and noting that $\phi_{\min}(V_{t_n})$ is an increasing function of n yields

$$P\left\{\max\nolimits_{-t_n\delta\le k\le t_n\delta}\|V_{t_n}^{-1/2}(S_{t_n+k}-S_{t_n})\|>\varepsilon\right\} \tag{A.12}$$

$$\le 2(r/\varepsilon)^2\ \mathrm{tr}\ (V_{[t_n(1+\delta)]}-V_{t_n})/\phi_{\min}(V_{t_n}).$$

Set $\eta=\varepsilon^3/(2r^2)$ and choose $\delta>0$ such that (A.8) holds. It follows from (A.12) that Condition (5.4) is satisfied for the chosen δ and the first relation in (A.9) holds.

The verification of the second relation in (A.9) makes use of properties of Cholesky square roots of positive definite matrices. Since $V_{T_n}^{-1/2}$ and $V_{t_n}^{1/2}$ are the left Cholesky roots of some matrices, they are both lower triangular with positive diagonal elements and hence $V_{T_n}^{-1/2}V_{t_n}^{1/2}$ is the left Cholesky root of $V_{T_n}^{-1/2}V_{t_n}(V_{T_n}^{-1/2})'$. Therefore $V_{T_n}^{-1/2}V_{t_n}(V_{T_n}^{-1/2})'\xrightarrow{p}I_r$ implies that $V_{T_n}^{-1/2}V_{t_n}^{1/2}\xrightarrow{p}I_r$. It is worth noting that this argument would break down if the symmetric square roots of matrices were used because in general $V_{T_n}^{-1/2}V_{t_n}^{1/2}$ is not a symmetric square root of $V_{T_n}^{-1/2}V_{t_n}V_{T_n}^{-1/2}$.

In order to show that $V_{T_n}^{-1/2}V_{t_n}(V_{T_n}^{-1/2})'\xrightarrow{p}I_r$, one can employ an argument similar to that used for proving the first relation in (A.9). Find a sequence of real integers $\{t_n, n\ge 1\}$ such that $t_n\to\infty$ and $T_n/t_n\xrightarrow{p}1$, $n\to\infty$, and assume that, for any $\varepsilon>0$ and some $\delta_\varepsilon>0$,

$$\max_{-t_n\delta_\varepsilon\le k\le t_n\delta_\varepsilon}\|V_{t_n+k}^{-1/2}V_{t_n}(V_{t_n+k}^{-1/2})'-I_r\|\le\varepsilon\qquad\text{for any large } n. \tag{A.13}$$

Then $P\{\|V_{T_n}^{-1/2}V_{t_n}(V_{T_n}^{-1/2})'-I_r\|>\varepsilon\}\le P\{|T_n/t_n-1|>\delta\}$ as $n\to\infty$ and therefore $V_{T_n}^{-1/2}V_{t_n}(V_{T_n}^{-1/2})'\xrightarrow{p}I_r$. Now it suffices to verify that (A.13) holds. Again, considering separately the cases $0\le k\le t_n\delta$ and $-t_n\delta\le k\le 0$, it can be seen from (A.10) that, for any $0\le k\le t_n\delta$,

$$\|V_{t_n+k}^{-1/2}V_{t_n}(V_{t_n+k}^{-1/2})'-I_r\|\le\|V_{t_n+k}^{-1/2}\|\|V_{t_n+k}-V_{t_n}\|\|(V_{t_n+k}^{-1/2})'\|$$

$$\le r\|V_{t_n+k}-V_{t_n}\|/\phi_{\min}(V_{t_n+k}) \tag{A.14}$$

$$\le r\ \mathrm{tr}\ (V_{[t_n(1+\delta)]}-V_{t_n})/\phi_{\min}(V_{t_n})$$

since $\phi_{\min}(V_{t_n+k})\ge\phi_{\min}(V_{t_n})$ and $\|V_{t_n+k}-V_{t_n}\|\le\mathrm{tr}\ (V_{t_n+k}-V_{t_n})\le\mathrm{tr}\ (V_{[t_n(1+\delta)]}-V_{t_n})$ due to the monotonicity of eigenvalues [cf. Bellman (1960),

Chapter 7]. Utilizing the same argument when $-t_n\delta \le k \le 0$ together with (A.14) yields

$$\text{(A.15)}\quad \max_{-t_n\delta \le k \le t_n\delta} \| V_{t_n+k}^{-1/2} V_{t_n} (V_{t_n+k}^{-1/2})' - I_r \| \le 2r \text{ tr } (V_{[t_n(1+\delta)]} - V_{t_n})/\phi_{\min}(V_{t_n}).$$

Set $\eta = \varepsilon/(2r)$ in (A.8) and choose the $\delta > 0$ for which this condition is satisfied. Then the right-hand side of (A.15) is less than ε implying that (A.13) holds and hence both relations in (A.9) hold. This establishes the asymptotic normality of $V_{T_n}^{-1/2} S_{T_n}$ and thus completes the proof of Lemma A.4. □

Acknowledgments. The authors thank Professor David Siegmund for his useful comments. The authors also thank Professor James Berger and an Associate Editor for their encouragement and helpful suggestions and the referees for their critical reading of the manuscript.

REFERENCES

ANSCOMBE, F. J. (1952). Large sample theory of sequential estimation. *Proc. Cambridge Philos. Soc.* **48** 600–607.

BAUM, L. E. and KATZ, M. (1965). Convergence rates in the law of large numbers. *Trans. Amer. Math. Soc.* **120** 108–123.

BELLMAN, R. (1970). *Introduction to Matrix Analysis.* McGraw-Hill, New York.

BERK, R. H. (1972). Consistency and asymptotic normality of maximum likelihood estimates for exponential models. *Ann. Math. Statist.* **43** 193–204.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

CHANG, Y. I. and MARTINSEK, A. (1992). Fixed size confidence regions for parameters of a logistic regression model. *Ann. Statist.* **20** 1953–1969.

CHANG, Y. I. (1995). Estimation in some binary regression models with prescribed accuracy. *J. Statist. Plann. Inference* **44** 313–325.

CHOW, Y. S. and ROBBINS, H. (1965). On the asymptotic theory of fixed-width sequential confidence intervals for the mean. *Ann. Math. Statist.* **36** 457–462.

CHOW, Y. S. and TEICHER, H. (1988). *Probability Theory.* Springer, New York.

CHURCH, J. D. and COBB, E. B. (1973). On the equivalence of the Spearman-Karber and maximum likelihood estimates of the mean. *J. Amer. Statist. Assoc.* **68** 201–202.

CRAMÈR, H. (1946). *Mathematical Methods of Statistics.* Princeton Univ. Press.

FAHRMEIR, L. and KAUFMANN, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.* **13** 342–368. [Correction note (1986) *Ann. Statist.* **14** 1643.]

FAHRMEIR, L. (1987). Asymptotic testing theory for generalized linear models. *Statistics* **18** 65–76.

FELLER, W. (1966). *An Introduction to Probability Theory and Its Applications II*. Wiley, New York.

GLESER, L. (1965). On the asymptotic theory of fixed size sequential confidence bounds for linear regression parameters. *Ann. Math. Statist.* **36** 463–467. [Correction note (1966) *Ann. Math. Statist.* **37** 1053–1055.]

GLESER, L. (1969). On limiting distribution for sums of a random number of independent random vectors. *Ann. Math. Statist.* **40** 935–941.

GOVINDARAJULU, Z. (1987). *The Sequential Statistical Analysis of Hypothesis Testing, Point and Interval Estimation, and Decision Theory.* American Sciences Press, Columbus, OH.

GOVINDARAJULU, Z. (1988). *Statistical Analysis of Bioassay.* Karger, New York.

GOVINDARAJULU, Z. and NANTHAKUMAR, A. (2000). Sequential estimation of the the mean of logistic response function. *Statistics* **33** 309–332.

GRAMBSCH, P. (1983). Sequential sampling based on the observed Fisher information to guarantee the accuracy of the maximum likelihood estimator. *Ann. Statist.* **11** 68–77.

GRAMBSCH, P. (1989). Sequential maximum likelihood estimation with applications to logistic regression in case-control studies. *J. Statist. Plann. Inference* **22** 355–369.

HABERMAN, S. J. (1977). Maximum likelihood estimates in exponential response models. *Ann. Statist.* **5** 815–841.

HORN, R. A. and JOHNSON, C. R. (1985). *Matrix Analysis*. Cambridge Univ. Press.

HSU, P. L. and ROBBINS, H. (1947). Complete convergence and the law of large numbers. *Proc. Nat. Acad. Sci. U.S.A.* **33** 25–31.

IBRAGIMOV, I. A. and HASMINSKII, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer, New York.

KAUFMANN, H. (1987). On the strong law of large numbers for multivariate martingales. *Stochastic Process Appl.* **26** 73–85.

LAI, T. L. and SIEGMUND, D. (1977). A nonlinear renewal theory with applications to sequential analysis, I. *Ann. Statist.* **5** 946–954.

LAI, T. L. and SIEGMUND, D. (1979). A nonlinear renewal theory with applications to sequential analysis, II. *Ann. Statist.* **7** 60–76.

LAI, T. L. and SIEGMUND, D. (1983). Fixed accuracy estimation of an autoregressive parameter. *Ann. Statist.* **11** 478–485.

LAI, T. L. (1996). On uniform integrability and asymptotically risk-efficient sequential estimation. *Sequential Anal.* **15** 237–251.

MILLER, R. G. (1973). Nonparametric estimators of the mean tolerance in bioassay. *Biometrika* **60** 535–542.

NANTHAKUMAR, A. and GOVINDARAJULU, Z. (1994). Risk-efficient estimation of the mean of the logistic response function using the Spearman-Karber estimator. *Statist. Sinica* **4** 305–324.

NANTHAKUMAR, A. and GOVINDARAJULU, Z. (1999). Fixed-width estimation of the mean of logistic response function using the Spearman-Karber estimator. *Biomedical J.* **4** 445–456.

NELDER, J. A. and WEDDERNBURN, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A* **135** 370–384.

ROY, S. N. (1957). *Some Aspects of Multivariate Analysis*. Wiley, New York.

SIEGMUND, D. (1985). *Sequential Analysis*. Springer, New York.

SIEGMUND, D. and SELLKE, T. (1983). Sequential analysis of the proportional hazards model. *Biometrika* **70** 315–326.

SRIVASTAVA, M. S. (1971). On fixed-width confidence bounds for regression parameters. *Ann. Math. Statist.* **42** 1403–1411.

STEIN, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Ann. Math. Statist.* **16** 243–258.

SWEETING, T. J. (1980). Uniform asymptotic normality of the maximum likelihood estimator. *Ann. Statist.* **8** 1375–1381. [Correction note (1982). *Ann. Statist.* **10** 320–321.]

VEXLER, A. A. and KONEV, V. V. (1995). On the mean number of observations under guaranteed estimation of an autoregression parameter. *Automation Remote Control* **56** 844–850.

VEXLER, A. A. and DMITRIENKO, A. A. (1999). Approximations to expected stopping times with applications to sequential estimation. *Sequential Anal.* **18** 165–187.

WEISS, L. (1971). Asymptotic properties of maximum likelihood estimators in some nonstandard cases. *J. Amer. Statist. Assoc.* **66** 345–350.

WEISS, L. (1973). Asymptotic properties of maximum likelihood estimators in some nonstandard cases II. *J. Amer. Statist. Assoc.* **68** 428–430.

WOODROOFE, M. (1977). Second order approximation to sequential point and interval estimation. *Ann. Statist.* **5** 984–995.

WOODROOFE, M. (1982). *Nonlinear Renewal Theory in Sequential Analysis*. SIAM, Philadelphia.

LILLY RESEARCH LABORATORIES
LILLY CORPORATE CENTER
ELI LILLY AND COMPANY
INDIANAPOLIS, INDIANA 46285

DEPARTMENT OF STATISTICS
817 PATTERSON OFFICE TOWER
UNIVERSITY OF KENTUCKY
LEXINGTON, KENTUCKY 40506