Stochastic Systems
2014, Vol. 4, No. 2, 375–403
DOI: 10.1214/12-SSY086

MEAN SQUARE CONVERGENCE RATES FOR MAXIMUM QUASI-LIKELIHOOD ESTIMATORS

By Arnoud V. den Boer[*,‡] and Bert Zwart[†,§]

University of Twente[‡], *Centrum Wiskunde & Informatica (CWI)*[§]

In this note we study the behavior of maximum quasilikelihood estimators (MQLEs) for a class of statistical models, in which only knowledge about the first two moments of the response variable is assumed. This class includes, but is not restricted to, generalized linear models with general link function. Our main results are related to guarantees on existence, strong consistency and mean square convergence rates of MQLEs. The rates are obtained from first principles and are stronger than known a.s. rates. Our results find important application in sequential decision problems with parametric uncertainty arising in dynamic pricing.

1. Introduction.

1.1. *Motivation.* We consider a statistical model of the form

$$E\left[Y(x)\right] = h(x^T \beta^{(0)}), \quad \mathrm{Var}(Y(x)) = v(E\left[Y(x)\right]), \tag{1}$$

where $x \in \mathbb{R}^d$ is a design variable, $Y(x)$ is a random variable whose distribution depends on x, $\beta^{(0)} \in \mathbb{R}^d$ is an unknown parameter, and h and v are known functions on $\mathbb{R}$. Such models arise, for example, from generalized linear models (GLMs), where in addition to (1) one requires that the distribution of $Y(x)$ comes from the exponential family (cf. Nelder and Wedderburn (1972), McCullagh and Nelder (1983), Gill (2001)). We are interested in making inference on the unknown parameter $\beta^{(0)}$.

In GLMs, this is commonly done via maximum-likelihood estimation. Given a sequence of design variables $x_1, \ldots, x_n$ and observed responses $y_1, \ldots, y_n$, where each y_i is a realization of the random variable $Y(x_i)$,

Received November 2012.

[*] Part of this research was done while the first author was with CWI. The authors kindly thank Yuan-chin Ivan Chang (Institute of Statistical Science, Academia Sinica, Taipei, Taiwan) for constructive e-mail contact about this research, and the editorial team for their handling of the paper.

[†] This research is supported by an NWO VIDI grant.

AMS 2000 subject classifications: 62F12, 62J12.

Keywords and phrases: Quasi-likelihood estimation, strong consistency, mean square convergence rates.

the maximum-likelihood estimator (MLE) $\hat{\beta}_n$ is a solution to the equation $l_n(\beta) = 0$, where $l_n(\beta)$ is defined as

$$l_n(\beta) = \sum_{i=1}^{n} \frac{\dot{h}(x_i^T \beta)}{v(h(x_i^T \beta))} x_i (y_i - h(x_i^T \beta)), \tag{2}$$

and where $\dot{h}$ denotes the derivative of h.

As discussed by Wedderburn (1974) and McCullagh (1983), if one drops the requirement that the distribution of $Y(x)$ is a member of the exponential family, and only assumes (1), one can still make inference on β by solving $l_n(\beta) = 0$. The solution $\hat{\beta}_n$ is then called a maximum quasi-likelihood estimator (MQLE) of $\beta^{(0)}$.

In this note, we are interested in the quality of the estimate $\hat{\beta}_n$ for models satisfying (1) by considering the expected value of $||\hat{\beta}_n - \beta^{(0)}||^2$, where $||\cdot||$ denotes the Euclidean norm. An important motivation comes from recent interest in sequential decision problems under uncertainty, in the field of dynamic pricing and revenue management (Besbes and Zeevi, 2009, Araman and Caldentey, 2011, den Boer and Zwart, 2013, den Boer, 2013, Broder and Rusmevichientong, 2012). In such problems, one typically considers a seller of products, with a demand distribution from a parametrized family of distributions. The goal of the seller is twofold: learning the value of the unknown parameters, and choosing selling prices as close as possible to the optimal selling price. The quality of the parameter estimates generally improves in presence of price variation, but that usually has negative effect on short-term revenue. Recently, there has been much interest in designing price-decision rules that optimally balance this so-called exploration-exploitation trade-off. The performance of such decision rules are typically characterized by the regret, which is the expected amount of revenue lost caused by not choosing the optimal selling price. For the design of price-decision rules and evaluation of the regret, knowledge of the behavior of $E[||\hat{\beta}_n - \beta^{(0)}||^2]$ is of vital importance.

1.2. *Literature.* Although much literature is devoted to the (asymptotic) behavior of maximum (quasi-)likelihood estimators for models of the form (1), practically all of them focus on a.s. upper bounds on $||\hat{\beta}_n - \beta^{(0)}||$ instead of mean square bounds. The literature may be classified according to the following criteria:

1. Assumptions on (in)dependence of design variables and error terms. The sequence of vectors $(x_i)_{i\in\mathbb{N}}$ is called the design, and the error terms $(e_i)_{i\in\mathbb{N}}$ are defined as

$$e_i = y_i - h(x_i^T \beta^{(0)}), \quad (i \in \mathbb{N}).$$

Typically, one either assumes a fixed design, with all x_i non-random and the e_i mutually independent, or an adaptive design, where the sequence $(e_i)_{i \in \mathbb{N}}$ forms a martingale difference sequence w.r.t. its natural filtration and where the design variables $(x_i)_{i \in \mathbb{N}}$ are predictable w.r.t. this filtration. This last setting is appropriate for sequential decision problems under uncertainty, where decisions are made based on current parameter-estimates.

2. Assumptions on the dispersion of the design vectors.
Define the design matrix

$$P_n = \sum_{i=1}^{n} x_i x_i^T, \tag{3}$$

and denote by $\lambda_{\min}(P_n)$, $\lambda_{\max}(P_n)$ the smallest and largest eigenvalues of P_n. Bounds on $||\hat{\beta}_n - \beta^{(0)}||$ are typically stated in terms of these two eigenvalues, which in some sense quantify the amount of dispersion in the sequence $(x_i)_{i \in \mathbb{N}}$.

3. Assumptions on the link function.
In GLM terminology, h^{-1} is called the link function. It is called *canonical* or *natural* if $\dot{h} = v \circ h$, otherwise it is called a *general* or *non-canonical* link function. The quasi-likelihood equations (2) for canonical link functions simplify to $l_n(\beta) = \sum_{i=1}^{n} x_i(y_i - h(x_i^T \beta)) = 0$.

To these three sets of assumptions, one usually adds smoothness conditions on h and v, and assumptions on the moments of the error terms.

An early result on the asymptotic behavior of solutions to (2), is from Fahrmeir and Kaufmann (1985). For fixed design and canonical link function, provided $\lambda_{\min}(P_n) = \Omega(\lambda_{\max}(P_n)^{1/2+\delta})$ a.s. for a $\delta > 0$ and some other regularity assumptions, they prove asymptotic existence and strong consistency of $(\hat{\beta}_n)_{n \in \mathbb{N}}$ (their Corollary 1; for the definition of $\Omega(\cdot)$, $O(\cdot)$ and $o(\cdot)$, see the next paragraph on notation). For general link functions, these results are proven assuming $\lambda_{\min}(P_n) = \Omega(\lambda_{\max}(P_n))$ a.s. and some other regularity conditions (their Theorem 5).

Chen et al. (1999) consider only canonical link functions. In the fixed design case, they obtain strong consistency and convergence rates

$$||\hat{\beta}_n - \beta^{(0)}|| = o(\{(\log(\lambda_{\min}(P_n)))^{1+\delta}/\lambda_{\min}(P_n)\}^{1/2}) \text{ a.s.},$$

for any $\delta > 0$; in the adaptive design case, they obtain convergence rates

$$||\hat{\beta}_n - \beta^{(0)}|| = O(\{(\log(\lambda_{\max}(P_n))/\lambda_{\min}(P_n)\}^{1/2}) \text{ a.s.} \tag{4}$$

Their proof however is reported to contain a mistake, see Zhang and Liao (2008, page 1289). These latter authors show for the case of fixed designs and canonical link functions that $||\hat{\beta}_n - \beta^{(0)}|| = O_p(\lambda_{\min}(P_n)^{-1/2})$, provided $\lambda_{\min}(P_n) = \Omega(\lambda_{\max}(P_n)^{1/2})$ a.s. and other regularity assumptions. Zhu and Gao (2013) extend these result to adaptive designs and prove $||\hat{\beta}_n - \beta^{(0)}|| = o_p(\lambda_{\min}(P_n)^{-1/2+\delta})$, for arbitrarily small $\delta > 0$. A.s. bounds on the estimation error in this setting are obtained by Zhang et al. (2011) who show

$$||\hat{\beta}_n - \beta^{(0)}|| = O(\lambda_{\max}(P_n)^{1/2}(\log(\lambda_{\max}(P_n)))^{\delta/2}\lambda_{\min}(P_n)^{-1}) \text{ a.s.,} \tag{5}$$

for arbitrarily small $\delta > 0$.

Chang (1999) extends (4) to a setting with *general* link functions and adaptive designs, under the additional condition $\lambda_{\min}(P_n) = \Omega(n^\alpha)$ a.s. for some $\alpha > 1/2$. His proof however appears to contain a mistake, see Remark 1. In a similar setting, Yue and Chen (2004) derive convergence rates

$$||\hat{\beta}_n - \beta^{(0)}|| = O(\{n \log(\log(\lambda_{\max}(P_n)))\}^{1/2}/n^\delta) \text{ a.s.,} \tag{6}$$

assuming $\lambda_{\min}(P_n) = \Omega(n^{3/4+\delta})$ for some $\delta > 0$. Under weaker conditions on the growth rate of $\lambda_{\min}(P_n)$ and on the moments of the error terms e_i, Yin et al. (2008) extend Yue and Chen (2004) to a setting with adaptive design, general link function, and multivariate response data. They obtain strong consistency and a.s. convergence rates

$$||\hat{\beta}_n - \beta^{(0)}|| = o\left(\frac{\{\lambda_{\max}(P_n)\log(\lambda_{\max}(P_n))\}^{1/2}\{\log(\log(\lambda_{\max}(P_n)))\}^{1/2+\delta}}{\lambda_{\min}(P_n)}\right) \tag{7}$$

for $\delta > 0$, under assumptions on $\lambda_{\min}(P_n), \lambda_{\max}(P_n)$ that ensure that this asymptotic upper bound is $o(1)$ a.s. Note that, since $\lambda_{\max}(P_n) = O(n)$ for uniformly bounded designs, the rates in (7) imply the rates in (6) up to logarithmic terms.

1.3. *Assumptions and contributions.* In contrast with the literature discussed above, we study bounds for the expected value of $||\hat{\beta}_n - \beta^{(0)}||^2$. The design is assumed to be adaptive; i.e. the error terms $(e_i)_{i\in\mathbb{N}}$ form a martingale difference sequence w.r.t. the natural filtration $\{\mathcal{F}_i\}_{i\in\mathbb{N}}$, and the design variables $(x_i)_{i\in\mathbb{N}}$ are predictable w.r.t. this filtration. For applications of our results to sequential decision problems, where each new decision can depend on the most recent parameter estimate, this is the appropriate setting to consider. In addition, we assume $\sup_{i\in\mathbb{N}} E[e_i^2 \mid \mathcal{F}_{i-1}] \leq \sigma^2 < \infty$ a.s. for some $\sigma > 0$, and $\sup_{i\in\mathbb{N}} E[|e_i|^r] < \infty$ for some $r > 2$.

We consider general link functions, and only assume that h and v are thrice continuously differentiable with $\dot{h}(z) > 0$, $v(h(z)) > 0$ for all $z \in \mathbb{R}$. Concerning the design vectors $(x_i)_{i\in\mathbb{N}}$, we assume that they are contained in a bounded subset $X \subset \mathbb{R}^d$. Let $\lambda_1(P_n) \leq \lambda_2(P_n)$ denote the two smallest eigenvalues of the design matrix P_n (if the dimension d of $\beta^{(0)}$ equals 1, write $\lambda_2(P_n) = \lambda_1(P_n)$). We assume that there is a (non-random) $n_0 \in \mathbb{N}$ such that P_{n_0} is invertible, and there are (non-random) functions L_1, L_2 on $\mathbb{N}$ such that for all $n \geq n_0$: $\lambda_1(P_n) \geq L_1(n)$, $\lambda_2(P_n) \geq L_2(n)$, and

$$L_1(n) \geq cn^{\alpha}, \qquad \text{for some } c > 0,\ \frac{1}{2} < \alpha \leq 1 \text{ independent of } n. \tag{8}$$

Based on these assumptions, we obtain three important results concerning the asymptotic existence of $\hat{\beta}_n$ and bounds on $E[||\hat{\beta}_n - \beta^{(0)}||^2]$:

1. First, notice that a solution to (2) need not always exist. Following Chang (1999), we therefore define the last-time that there is no solution in a neighborhood of $\beta^{(0)}$:

$$N_\rho = \sup\left\{ n \geq n_0 : \begin{array}{l} \text{there exists no } \beta \in \mathbb{R}^d \text{ with } l_n(\beta) = 0 \\ \text{and } ||\hat{\beta}_n - \beta^{(0)}|| \leq \rho \end{array} \right\}.$$

   For all sufficiently small $\rho > 0$, we show in Theorem 1 that N_ρ is finite a.s., and provide sufficient conditions such that $E[N_\rho^\eta] < \infty$, for $\eta > 0$.
2. In Theorem 2, we provide the upper bound

$$E\left[\left|\left| \hat{\beta}_n - \beta^{(0)} \right|\right|^2 \mathbf{1}_{n > N_\rho} \right] = O\left(\frac{\log(n)}{L_1(n)} + \frac{n(d-1)^2}{L_1(n)L_2(n)} \right), \tag{9}$$

   where $\mathbf{1}_{n>N_\rho}$ denotes the indicator function of the event $\{n > N_\rho\}$.
3. In case of a canonical link function, Theorem 3 improves these bounds to

$$E\left[\left|\left| \hat{\beta}_n - \beta^{(0)} \right|\right|^2 \mathbf{1}_{n > N_\rho} \right] = O\left(\frac{\log(n)}{L_1(n)} \right). \tag{10}$$

   This improvement clearly is also valid for general link functions provided $d = 1$. It also holds if $d = 2$ and $||x_i||$ is bounded from below by a positive constant (see Remark 2).

Our L^2 bounds (9) are sharper than the (a.s.) bounds derived by Yin et al. (2008). With bounded regressors that are bounded away from zero (a minor condition, since in most applications an intercept term is present in the regressors), the bounds of Yin et al. (2008, Theorem 2.1) reduce to

$$||\hat{\beta}_n - \beta^{(0)}||^2 = o\left(\frac{n \log(n) \log(\log(n))^{1+2\delta}}{\lambda_{\min}(n)^2} \right) \text{ a.s., for some } \delta > 0. \tag{11}$$

For $d = 1$ or $d = 2$, our convergence rates improve the rate (ignoring to logarithmic factors) of Yin et al. (2008) by a factor $n/L_1(n)$. For general $d > 1$, our convergence rates improve (11) (up to logarithmic factors) whenever $L_2(n)/L_1(n) \to \infty$ as $n \to \infty$. And if $L_2(n) \sim L_1(n)$, then our rates still (modestly) improve (11) by removing logarithmic factors. Note that these improvements are not just theoretical constructs, but have practical value. For example, for the case $d = 1$ or 2, Keskin and Zeevi (2013) and den Boer and Zwart (2013) show for certain dynamic pricing problems that a design satisfying $L_1(n) \sim n^{1/2}$ is optimal. Such conclusions can not be obtained from the rates (11).

Our results also differ from Yin et al. (2008) in terms of proof techniques. For general link functions, our starting point is a corollary of the Leray-Schauder theorem to ensure existence of the MQLE; we subsequently bound moments of last-time random variables, use Taylor approximations, apply martingale techniques, and deploy a result (Lemma 7) on the magnitude of solutions to certain quadratic equations. The proof of Yin et al. (2008) starts from a different topological result (a corollary of Brouwer's domain invariant mapping theorem, Dugundji (1966)), and arrives at different convergence rates. Because our L^2 bounds are in general sharper than existing a.s. bounds (Equations (5), (6), (7)), an attempt to derive our results from these bounds (e.g. using an uniform-integrability argument) would lead to weaker results than what we derive from first principles.

An important intermediate result in proving our main theorems is Proposition 2, where we derive

$$E\left|\left|\left(\sum_{i=1}^n x_i x_i^T\right)^{-1}\sum_{i=1}^n x_i e_i\right|\right|^2 = O\left(\frac{\log(n)}{L(n)}\right),$$

for any function L that satisfies $\lambda_{\min}(\sum_{i=1}^n x_i x_i^T) \geq L(n) > 0$ for all sufficiently large n. This actually provides bounds on mean square convergence rates in least-squares linear regression, and forms the counterpart of Lai and Wei (1982) who prove similar bounds in an a.s. setting.

Another auxiliary result derived in this paper is Lemma 4, which shows that the maximum of a martingale $(S_i)_{i\in\mathbb{N}}$ w.r.t. a filtration $\{\mathcal{F}_i\}_{i\in\mathbb{N}}$ satisfies

$$(12) \qquad P\left(\max_{1\leq k\leq n}|S_k| \geq \epsilon\right) \leq 2P\left(|S_n| \geq \epsilon - \sqrt{2\sigma^2 n}\right), \quad (n\in\mathbb{N}, \epsilon > 0),$$

where $\sup_{i\in\mathbb{N}} E[(S_{i+1} - S_i)^2 \mid \mathcal{F}_{i-1}] \leq \sigma^2 < \infty$ a.s. This result extends a similar statement on i.i.d. random variables found in Loève (1977a, Section 18.1C, page 260), and may be of independent interest to the reader.

1.4. *Applications.* Our results find important application in dynamic pricing problems. In these problems a seller tries to estimate from data the revenue-maximizing selling price for a particular product. To this end, the seller estimates unknown parameters $\beta^{(0)}$ of a parametric model that describes customer behavior. Let $r(\beta)$ denote the expected revenue when the seller uses the selling price that is optimal w.r.t. parameter estimate β. In many settings, the expected revenue loss $E[r(\beta^{(0)}) - r(\hat{\beta}_n)]$ caused by estimation errors is quadratic in $||\beta^{(0)} - \hat{\beta}_n||$. Our theorems 1 and 2 can then be used to bound this loss:

$$\begin{aligned}
&E\left[\left|\left|r(\beta^{(0)}) - r(\hat{\beta}_n)\right|\right|\right] \\
&= O\left(E\left[\left|\left|r(\beta^{(0)}) - r(\hat{\beta}_n)\right|\right| \mathbf{1}_{n>N_\rho}\right] + E\left[\left|\left|r(\beta^{(0)}) - r(\hat{\beta}_n)\right|\right| \mathbf{1}_{n\leq N_\rho}\right]\right) \\
&= O\left(E\left[\left|\left|\hat{\beta}_n - \beta^{(0)}\right|\right|^2 \mathbf{1}_{n>N_\rho}\right] + \frac{E\left[N_\rho^\eta\right]}{n^\eta} \max_\beta \left|\left|r(\beta) - r(\beta^{(0)})\right|\right|^2\right) \\
&= O\left(\frac{\log(n)}{L_1(n)} + n^{-\eta}\right).
\end{aligned}$$

In dynamic pricing problems, such arguments are used to design optimal decision policies, cf. den Boer and Zwart (2013). These type of arguments can also be applied to other sequential decision problems with parametric uncertainty, where the objective is to minimize the regret; for example the multiperiod inventory control problem (Anderson and Taylor (1976), Lai and Robbins (1982)) or for parametric variants of bandit problems (cf. Goldenshluger and Zeevi, 2009, Rusmevichientong and Tsitsiklis, 2010).

In his review on experimental design and control problems, Pronzato (2008, page 18, Section 9) mentions that existing consistency results for adaptive design of experiments are usually restricted to models that are linear in the parameters. The class of statistical models that we consider is much larger than only linear models; it includes all models satisfying (1). Our results may therefore also find application in the field of sequential design of experiments.

1.5. *Organization of the paper.* The rest of this paper is organized as follows: Section 2 contains our results concerning the last-time N_ρ and upper bounds on $E[||\hat{\beta}_n - \beta^{(0)}||^2 \mathbf{1}_{n>N_\rho}]$, for general link functions. In Section 3 we derive these bounds in the case of canonical link functions. Section 4 contains the proofs of the assertions in Section 2 and 3. In the appendix, Section 4, we collect and prove several auxiliary results which are used in the proofs of the theorems of Sections 2 and 3.

Notation. For $\rho > 0$, let $B_\rho = \{\beta \in \mathbb{R}^d \mid ||\beta - \beta^{(0)}|| \leq \rho\}$ and $\partial B_\rho = \{\beta \in \mathbb{R}^d \mid ||\beta - \beta^{(0)}|| = \rho\}$. The closure of a set $S \subset \mathbb{R}^d$ is denoted by $\bar{S}$, the boundary by $\partial S = \bar{S} \backslash S$. For $x \in \mathbb{R}$, $\lfloor x \rfloor$ denotes the largest integer that does not exceed x. The Euclidean norm of a vector y is denoted by $||y||$. The norm of a matrix A equals $||A|| = \max_{z:||z||=1} ||Az||$. The 1-norm and ∞-norm of a matrix are denoted by $||A||_1$ and $||A||_\infty$. y^T denotes the transpose of a vector or matrix y. If $f(x), g(x)$ are functions with domain in $\mathbb{R}$ and range in $(0, \infty)$, then $f(x) = O(g(x))$ means there exists a $K > 0$ such that $f(x) \leq Kg(x)$ for all $x \in \mathbb{N}$, $f(x) = \Omega(g(x))$ means $g(x) = O(f(x))$, and $f(x) = o(g(x))$ means $\lim_{x\to\infty} f(x)/g(x) = 0$.

2. Results for general link functions.

In this section we consider the statistical model introduced in Section 1.1 for general link functions h, under all the assumptions listed in Section 1.3. The first main result is Theorem 1, which shows finiteness of moments of N_{ρ_0}. The second main result is Theorem 2, which proves asymptotic existence and strong consistency of the MQLE, and provides bounds on the mean square convergence rates.

Our results on the existence of the quasi-likelihood estimate $\hat{\beta}_n$ are based on the following fact, which is a consequence of the Leray-Schauder theorem (Leray and Schauder, 1934).

Lemma 1 (Ortega and Rheinboldt, 2000, 6.3.4, page 163). *Let C be an open bounded set in $\mathbb{R}^n$, $F : \bar{C} \to \mathbb{R}^n$ a continuous mapping, and $(x - x_0)^T F(x) \geq 0$ for some $x_0 \in C$ and all $x \in \partial C$. Then $F(x) = 0$ has a solution in $\bar{C}$.*

This lemma yields a sufficient condition for the existence of $\hat{\beta}_n$ in the proximity of $\beta^{(0)}$ (recall the definitions $B_\rho = \{\beta \in \mathbb{R}^d \mid ||\beta - \beta^{(0)}|| \leq \rho\}$ and $\partial B_\rho = \{\beta \in \mathbb{R}^d \mid ||\beta - \beta^{(0)}|| = \rho\}$):

Corollary 1. *For all $\rho > 0$, if $\sup_{\beta \in \partial B_\rho} (\beta - \beta^{(0)})^T l_n(\beta) \leq 0$ then there exists a $\beta \in B_\rho$ with $l_n(\beta) = 0$.*

A first step in applying Corollary 1 is to provide an upper bound for $(\beta - \beta^{(0)})^T l_n(\beta)$. To this end, write $g(x) = \frac{\dot{h}(x)}{v(h(x))}$, and choose a $\rho_0 > 0$ such that $(c_2 - c_1 c_3 \rho) \geq c_2/2$ for all $0 < \rho \leq \rho_0$, where

$$c_1 = \sup_{\substack{x \in X, \\ \beta \in B_{\rho_0}}} \frac{1}{2} |\ddot{g}(x^T\beta)| \, ||x|| \, , \quad c_2 = \inf_{\substack{x \in X, \\ \beta, \tilde{\beta} \in B_{\rho_0}}} g(x^T\beta)\dot{h}(x^T\tilde{\beta}),$$
$$c_3 = \sup_{i \in \mathbb{N}} E[|e_i| \mid \mathcal{F}_{i-1}]. \tag{13}$$

The existence of such a ρ_0 follows from the fact that $\dot{h}(x) > 0$ and $g(x) > 0$ for all $x \in \mathbb{R}$, together with the continuity assumptions on h and g.

LEMMA 2. *Let $0 < \rho \leq \rho_0$, $\beta \in B_\rho$, $n \in \mathbb{N}$, and define*

$$A_n = \sum_{i=1}^n g(x_i^T \beta^{(0)}) x_i e_i, \quad B_n = \sum_{i=1}^n \dot{g}(x_i^T \beta^{(0)}) x_i x_i^T e_i,$$
$$J_n = c_1 \sum_{i=1}^n (|e_i| - E[|e_i| \mid \mathcal{F}_{i-1}]) x_i x_i^T.$$

Then $(\beta - \beta^{(0)})^T l_n(\beta) \leq S_n(\beta) - (c_2/2)(\beta - \beta^{(0)})^T P_n(\beta - \beta^{(0)})$, where the martingale $S_n(\beta)$ is defined as

$$\begin{aligned} S_n(\beta) = {} & (\beta - \beta^{(0)})^T A_n + (\beta - \beta^{(0)})^T B_n (\beta - \beta^{(0)}) \\ & + \left\| \beta - \beta^{(0)} \right\| (\beta - \beta^{(0)})^T J_n (\beta - \beta^{(0)}). \end{aligned}$$

Following Chang (1999), define the last-time

$$N_\rho = \sup\{n \geq n_0 \mid \text{there is no } \beta \in B_\rho \text{ s.t. } l_n(\beta) = 0\}.$$

The following theorem shows that the η-th moment of N_ρ is finite, for $0 < \rho \leq \rho_0$ and sufficiently small $\eta > 0$. Recall our assumptions $\sup_{i \in \mathbb{N}} E[|e_i|^r] < \infty$, for some $r > 2$, and $\lambda_{\min}(P_n) \geq L_1(n) \geq cn^\alpha$, for some $c > 0$, $\frac{1}{2} < \alpha \leq 1$ and all $n \geq n_0$.

THEOREM 1. *$N_\rho < \infty$ a.s., and $E[N_\rho^\eta] < \infty$ for all $0 < \rho \leq \rho_0$ and $0 < \eta < r\alpha - 1$.*

REMARK 1. Chang (1999) also approaches existence and strong consistency of $\hat{\beta}_n$ via application of Corollary 1. To this end, he derives an upper bound $A_n + B_n + J_n - n^\alpha \epsilon^*$ for $(\beta - \beta^{(0)})^T l_n(\beta)$, cf. his equation (21). He proceeds to show that for all $\beta \in \partial B_\rho$ the last time that this upper bound is positive, has finite expectation (cf. his equation (22)). However, to deduce existence of $\hat{\beta}_n \in B_\rho$ from Corollary 1, one needs to prove (in Chang's notation)

$$E\left[\sup\{n \geq 1 \mid \exists \beta \in \partial B_\rho : A_n + B_n + J_n - n^\alpha \epsilon^* \geq 0\}\right] < \infty, \tag{14}$$

but Chang proves

$$\forall \beta \in \partial B_\rho : E\left[\sup\{n \geq 1 \mid A_n + B_n + J_n - n^\alpha \epsilon^* \geq 0\}\right] < \infty.$$

(Here the terms A_n, B_n, J_n and ϵ^* depend on β).

Our ideas are also different from Chang in the following sense: to prove (14), we show that T is bounded from above by a sum of last-time random variables, and repeatedly apply the c_r-inequality and Proposition 1, contained in the Appendix. This proposition shows finiteness of moments of last-time random variables, and is based on a Baum-Katz-Nagaev type theorem (Lemma 5) by Stoica (2007), and on bounds on tail probabilities of the maximum of a martingale (Lemma 4, which extends a similar result by Loève (1977a, Section 18.1C, page 260) on sums of i.i.d. random variables).

The following theorem shows asymptotic existence and strong consistency of $\hat{\beta}_n$, and provides mean square convergence rates.

THEOREM 2. *Let $0 < \rho \leq \rho_0$. For all $n > N_\rho$ there exists a solution $\hat{\beta}_n \in B_\rho$ to $l_n(\beta) = 0$, and $\lim_{n\to\infty} \hat{\beta}_n = \beta^{(0)}$ a.s. Moreover,*

$$E\left[\left|\left|\hat{\beta}_n - \beta^{(0)}\right|\right|^2 \mathbf{1}_{n>N_\rho}\right] = O\left(\frac{\log(n)}{L_1(n)} + \frac{n(d-1)^2}{L_1(n)L_2(n)}\right). \tag{15}$$

REMARK 2. If $d = 1$ then the term $\frac{n(d-1)^2}{L_1(n)L_2(n)}$ in (15) vanishes. If $d = 2$, the next to smallest eigenvalue $\lambda_2(P_n)$ of P_n is actually the largest eigenvalue of P_n. If in addition $\inf_{i\in\mathbb{N}} ||x_i|| \geq d_{\min} > 0$ a.s. for some $d_{\min} > 0$, then $\lambda_{\max}(P_n) \geq \frac{1}{2}\text{trace}(P_n) \geq \frac{d_{\min}}{2}n$, and $\frac{n(d-1)^2}{L_1(n)L_2(n)} = O(\frac{1}{L_1(n)})$. The bound in Theorem 2 then reduces to

$$E\left[\left|\left|\hat{\beta}_n - \beta^{(0)}\right|\right|^2 \mathbf{1}_{n>N_\rho}\right] = O\left(\frac{\log(n)}{L_1(n)}\right). \tag{16}$$

REMARK 3. In general, the equation $l_n(\beta) = 0$ may have multiple solutions. Procedures for selecting the “right” root are discussed in Small et al. (2000) and Heyde (1997, Section 13.3). Tzavelas (1998) shows that with probability one there exists not more than one consistent solution.

3. Results for canonical link functions.

In this section we consider again the statistical model introduced in Section 1.1, under all the assumptions listed in Section 1.3. In addition, we restrict to canonical link functions, i.e. functions h that satisfy $\dot{h} = v \circ h$. The quasi-likelihood equations (2) then simplify to

$$l_n(\beta) = \sum_{i=1}^{n} x_i(y_i - h(x_i^T\beta)) = 0. \tag{17}$$

This simplification enables us to improve the bounds from Theorem 2. In particular, the main result of this section is Theorem 3, which shows that the term $O(\frac{n(d-1)^2}{L_1(n)L_2(n)})$ in (15) vanishes, yielding the following upper bound on the mean square convergence rates:

$$E\left[\left|\left|\hat{\beta}_n - \beta^{(0)}\right|\right|^2 \mathbf{1}_{n>N_\rho}\right] = O\left(\frac{\log(n)}{L_1(n)}\right).$$

In the previous section, we invoked a corollary of the Leray-Schauder Theorem to prove existence of $\hat{\beta}_n$ in a proximity of $\beta^{(0)}$. In the case of canonical link function, a similar existence result is derived from the following fact:

Lemma 3 (Chen et al., 1999, Lemma A(i)). *Let $H : \mathbb{R}^d \to \mathbb{R}^d$ be a continuously differentiable injective mapping, $x_0 \in \mathbb{R}^d$, and $\delta > 0$, $r > 0$. If $\inf_{x:||x-x_0||=\delta} ||H(x) - H(x_0)|| \geq r$ then for all $y \in \{y \in \mathbb{R}^d \mid ||y - H(x_0)|| \leq r\}$ there is an $x \in \{x \in \mathbb{R}^d \mid ||x - x_0|| \leq \delta\}$ such that $H(x) = y$.*

Chen et al. (1999) assume that H is smooth, but an inspection of their proof reveals that H being a continuously differentiable injection is sufficient.

We apply Lemma 3 with $H(\beta) = P_n^{-1/2} l_n(\beta)$ and $y = 0$:

Corollary 2. *Let $0 < \rho \leq \rho_0$, $n \geq N_\rho$, $\delta > 0$ and $r > 0$. If both $||H_n(\beta^{(0)})|| \leq r$ and $\inf_{\beta \in \partial B_\delta} ||H_n(\beta) - H_n(\beta^{(0)})|| \geq r$, then there is a $\beta \in B_\delta$ with $P_n^{-1/2} l_n(\beta) = 0$, and thus $l_n(\beta) = 0$.*

Remark 4. The proof of Corollary 2 reveals that $l_n(\beta)$ is injective for all $n \geq n_0$, and thus $\hat{\beta}_n$ is uniquely defined for all $n \geq N_\rho$.

The following theorem improves the mean square convergence rates of Theorem 2 in case of canonical link functions.

Theorem 3. *In case of a canonical link function,*

$$E\left[\left|\left|\hat{\beta}_n - \beta^{(0)}\right|\right|^2 \mathbf{1}_{n\geq N_\rho}\right] = O\left(\frac{\log(n)}{L_1(n)}\right), \quad (0 < \rho \leq \rho_0). \tag{18}$$

Remark 5. Some choices of h, e.g. h the identity or the logit function, have the property that $\inf_{x\in X, \beta\in\mathbb{R}^d} \dot{h}(x^T\beta) > 0$, i.e. c_2 in equation (13) has a positive lower bound independent of ρ_0. Since canonical link functions have $c_1 = 0$ in equation (13), we then can choose $\rho_0 = \infty$ in Lemma 2, Theorem 1 and Theorem 3. Then $N_{\rho_0} = n_0$ and $\hat{\beta}_n$ exists a.s. for all $n \geq n_0$. Moreover, we can drop assumption (8) and obtain

$$E\left[\left|\left|\hat{\beta}_n - \beta^{(0)}\right|\right|^2\right] = O\left(\frac{\log(n)}{L_1(n)}\right), \quad (n \geq n_0). \tag{19}$$

for any positive lower bound $L_1(n)$ on $\lambda_{\min}(P_n)$. Naturally, one needs to assume $\log(n) = o(L_1(n))$ in order to conclude from (19) that $E[||\hat{\beta}_n - \beta^{(0)}||^2]$ converges to zero as $n \to \infty$.

4. Proofs.

Proof of Lemma 2. A Taylor expansion of h and g yields

$$\begin{aligned} y_i - h(x_i^T\beta) &= y_i - h(x_i^T\beta^{(0)}) + h(x_i^T\beta^{(0)}) - h(x_i^T\beta) \\ &= e_i - \dot{h}(x_i^T\tilde{\beta}_{i,\beta}^{(1)})x_i^T(\beta - \beta^{(0)}), \end{aligned} \tag{20}$$

$$\begin{aligned} g(x_i^T\beta) &= g(x_i^T\beta^{(0)}) + \dot{g}(x_i^T\beta^{(0)})x_i^T(\beta - \beta^{(0)}) \\ &\quad + \frac{1}{2}(\beta - \beta^{(0)})^T\ddot{g}(x_i^T\tilde{\beta}_{i,\beta}^{(2)})x_ix_i^T(\beta - \beta^{(0)}), \end{aligned} \tag{21}$$

for some $\tilde{\beta}_{i,\beta}^{(1)}$, $\tilde{\beta}_{i,\beta}^{(2)}$ on the line segment between β and $\beta^{(0)}$. As in Chang (1999, page 241), it follows that

$$\begin{aligned} (\beta - \beta^{(0)})^T l_n(\beta) &= (\beta - \beta^{(0)})^T\sum_{i=1}^n g(x_i^T\beta)x_i(e_i - \dot{h}(x_i^T\tilde{\beta}_{i,\beta}^{(1)})x_i^T(\beta - \beta^{(0)})) \\ &= (\beta - \beta^{(0)})^T\sum_{i=1}^n g(x_i^T\beta^{(0)})x_ie_i \\ &\quad + (\beta - \beta^{(0)})^T\sum_{i=1}^n \dot{g}(x_i^T\beta^{(0)})x_i^T(\beta - \beta^{(0)})x_ie_i \\ &\quad + (\beta - \beta^{(0)})^T\sum_{i=1}^n\left[\frac{1}{2}(\beta - \beta^{(0)})^T\ddot{g}(x_i^T\tilde{\beta}_{i,\beta}^{(2)})x_ix_i^T(\beta - \beta^{(0)})\right]x_ie_i \\ &\quad - (\beta - \beta^{(0)})^T\sum_{i=1}^n g(x_i^T\beta)x_i\dot{h}(x_i^T\tilde{\beta}_{i,\beta}^{(1)})x_i^T(\beta - \beta^{(0)}) \\ &= (\beta - \beta^{(0)})^TA_n + (\beta - \beta^{(0)})^TB_n(\beta - \beta^{(0)}) + (I) - (II), \end{aligned}$$

writing $(I) = (\beta - \beta^{(0)})^T\sum_{i=1}^n[\frac{1}{2}(\beta - \beta^{(0)})^T\ddot{g}(x_i^T\tilde{\beta}_{i,\beta}^{(2)})x_ix_i^T(\beta - \beta^{(0)})]x_ie_i$ and $(II) = (\beta - \beta^{(0)})^T\sum_{i=1}^n g(x_i^T\beta)x_i\dot{h}(x_i^T\tilde{\beta}_{i,\beta}^{(1)})x_i^T(\beta - \beta^{(0)})$. Since

$$(I) = (\beta - \beta^{(0)})^T\sum_{i=1}^n\left[\frac{1}{2}(\beta - \beta^{(0)})^T\ddot{g}(x_i^T\tilde{\beta}_{i,\beta}^{(2)})x_i\right]x_ix_i^T(\beta - \beta^{(0)})e_i$$

$$\begin{aligned}
&\leq (\beta - \beta^{(0)})^T \sum_{i=1}^{n} \left[\frac{1}{2} \left|\left| \beta - \beta^{(0)} \right|\right| |\ddot{g}(x_i^T \tilde{\beta}_{i,\beta}^{(2)})| \, ||x_i|| \right] x_i x_i^T (\beta - \beta^{(0)}) |e_i| \\
&\leq c_1 (\beta - \beta^{(0)})^T \sum_{i=1}^{n} \left|\left| \beta - \beta^{(0)} \right|\right| x_i x_i^T |e_i| (\beta - \beta^{(0)}) \\
&\leq c_1 (\beta - \beta^{(0)})^T \sum_{i=1}^{n} \left|\left| \beta - \beta^{(0)} \right|\right| x_i x_i^T (|e_i| - E\left[|e_i| \mid \mathcal{F}_{i-1}\right])(\beta - \beta^{(0)}) \\
&+ c_1 (\beta - \beta^{(0)})^T \sum_{i=1}^{n} \left|\left| \beta - \beta^{(0)} \right|\right| x_i x_i^T E\left[|e_i| \mid \mathcal{F}_{i-1}\right] (\beta - \beta^{(0)}) \\
&\leq \left|\left| \beta - \beta^{(0)} \right|\right| (\beta - \beta^{(0)})^T J_n (\beta - \beta^{(0)}) \\
&+ c_1 c_3 \left|\left| \beta - \beta^{(0)} \right|\right| (\beta - \beta^{(0)})^T \sum_{i=1}^{n} x_i x_i^T (\beta - \beta^{(0)})
\end{aligned}$$

and

$$(II) \geq c_2 (\beta - \beta^{(0)})^T \sum_{i=1}^{n} x_i x_i^T (\beta - \beta^{(0)}),$$

by combining all relevant inequalities we obtain

$$\begin{aligned}
(\beta - \beta^{(0)})^T l_n(\beta) \leq{}& (\beta - \beta^{(0)})^T A_n + (\beta - \beta^{(0)})^T B_n (\beta - \beta^{(0)}) \\
&+ \left|\left| \beta - \beta^{(0)} \right|\right| (\beta - \beta^{(0)})^T J_n (\beta - \beta^{(0)}) \\
&- (c_2/2)(\beta - \beta^{(0)})^T \sum_{i=1}^{n} x_i x_i^T (\beta - \beta^{(0)}),
\end{aligned}$$

using $(c_1 c_3 ||\beta - \beta^{(0)}|| - c_2) \leq (c_1 c_3 \rho - c_2) \leq -c_2/2$.

Proof of Theorem 1. Fix $\rho \in (0, \rho_0]$ and $0 < \eta < r\alpha - 1$. Let $S_n(\beta)$ be as in Lemma 2. Define the last-time

$$T = \sup\{n \geq n_0 \mid \sup_{\beta \in \partial B_\rho} S_n(\beta) - \rho^2 (c_2/2) L_1(n) > 0\}.$$

By Lemma 2, for all $n > T$,

$$\begin{aligned}
0 &\geq \sup_{\beta \in \partial B_\rho} S_n(\beta) - \rho^2 (c_2/2) L_1(n) \\
&\geq \sup_{\beta \in \partial B_\rho} S_n(\beta) - (c_2/2)(\beta - \beta^{(0)})^T P_n (\beta - \beta^{(0)}) \\
&\geq \sup_{\beta \in \partial B_\rho} (\beta - \beta^{(0)})^T l_n(\beta),
\end{aligned}$$

which by Corollary 1 implies $n > N_\rho$. Then $N_\rho \leq T$ a.s., and thus $E[N_\rho^\eta] \leq E[T^\eta]$ for all $\eta > 0$. The proof is complete if we show the assertions for T.

If we denote the entries of the vector A_n and the matrices B_n, J_n by $A_n[i]$, $B_n[i,j]$, $J_n[i,j]$, then

$$\sup_{\beta \in \partial B_\rho} S_n(\beta) \leq \rho\, ||A_n|| + \rho^2\, ||B_n|| + \rho^3\, ||J_n|| \\ \leq \rho \sum_{1 \leq i \leq d} |A_n[i]| + \rho^2 \sum_{1 \leq i,j \leq d} |B_n[i,j]| + \rho^3 \sum_{1 \leq i,j \leq d} |J_n[i,j]|,$$

using the Cauchy-Schwartz inequality and the fact that $||x|| \leq ||x||_1$, $||A|| \leq \sum_{i,j} |A[i,j]|$ for vectors x and matrices A. (This can be derived from the inequality $||A|| \leq \sqrt{||A||_1 ||A||_\infty}$). We now define $d + 2d^2$ last-times $T_{A[i]}$, $T_{B[i,j]}$, and $T_{J[i,j]}$, for all $1 \leq i,j \leq d$, as follows:

$$T_{A[i]} = \sup\{n \geq n_0 \mid \rho |A_n[i]| - \frac{1}{d+2d^2}\rho^2 (c_2/2) L_1(n) > 0\},$$
$$T_{B[i,j]} = \sup\{n \geq n_0 \mid \rho^2 |B_n[i,j]| - \frac{1}{d+2d^2}\rho^2 (c_2/2) L_1(n) > 0\},$$
$$T_{J[i,j]} = \sup\{n \geq n_0 \mid \rho^3 |J_n[i,j]| - \frac{1}{d+2d^2}\rho^2 (c_2/2) L_1(n) > 0\}.$$

By application of Proposition 1, Section 4, the last-times $T_{A[i]}$ and $T_{B[i,j]}$ are a.s. finite and have finite η-th moment, for all $\eta > 0$ such that $r > \frac{\eta+1}{\alpha} > 2$. Chow and Teicher (2003, page 95, Lemma 3) states that any two nonnegative random variables X_1, X_2 satisfy

$$E\left[(X_1 + X_2)^\eta\right] \leq 2^\eta (E\left[X_1^\eta\right] + E\left[X_2^\eta\right]), \tag{22}$$

for all $\eta > 0$. Consequently

$$\sup_{i \in \mathbb{N}} E\left[\left||e_i| - E\left[|e_i| \mid \mathcal{F}_{i-1}\right]\right|^r\right] \leq \sup_{i \in \mathbb{N}} E\left[\left||e_i| + E\left[|e_i| \mid \mathcal{F}_{i-1}\right]\right|^r\right] \\ \leq \sup_{i \in \mathbb{N}} 2^r (E\left[|e_i|^r\right] + E\left[(E\left[|e_i| \mid \mathcal{F}_{i-1}\right])^r\right]) < \infty,$$

and Proposition 1 implies that the last-times $T_{J[i,j]}$ are also a.s. finite and have finite η-th moment, for all $\eta > 0$ such that $r > \frac{\eta+1}{\alpha} > 2$. Now set $\mathcal{T} = \sum_{1 \leq i \leq d} T_{A[i]} + \sum_{1 \leq i,j \leq d} T_{B[i,j]} + \sum_{1 \leq i,j \leq d} T_{J[i,j]}$. If $n > \mathcal{T}$, then $\sup_{\beta \in \partial B_\rho} S_n(\beta) - \rho^2 (c_2/2) L_1(n) \leq 0$, and thus $T \leq \mathcal{T}$ a.s. and $E[T^\eta] \leq E[\mathcal{T}^\eta]$. $\mathcal{T}$ is finite a.s., since all terms $T_{A[i]}$, $T_{B[i,j]}$ and $T_{J[i,j]}$ are finite a.s. Moreover, by repeated application of (22), for all $\eta > 0$ there is a constant

C_η such that

$$E[\mathcal{T}^\eta] \le C_\eta \left[\sum_{1\le i\le d} E\left[T_{A[i]}\right] + \sum_{1\le i,j\le d} E\left[T^\eta_{B[i,j]}\right] + \sum_{1\le i,j\le d} E\left[T^\eta_{J[i,j]}\right] \right].$$

It follows that $E[\mathcal{T}^\eta] < \infty$ for all $\eta > 0$ such that $r > \frac{\eta+1}{\alpha} > 2$. In particular, this implies $N_\rho < \infty$ a.s., and $E[N_\rho^\eta] < \infty$.

Proof of Theorem 2. The asymptotic existence and strong consistency of $\hat{\beta}_n$ follow directly from Theorem 1 which shows $N_\rho < \infty$ a.s. for all $0 < \rho \le \rho_0$.

To prove the mean square convergence rates, let $0 < \rho \le \rho_0$.

By contraposition of Corollary 1, if there is no solution $\beta \in B_\rho$ to $l_n(\beta) = 0$, then there exists a $\beta' \in \partial B_\rho$ such that $(\beta' - \beta^{(0)})^T l_n(\beta') > 0$, and thus $S_n(\beta') - (c_2/2)(\beta' - \beta^{(0)})^T P_n(\beta' - \beta^{(0)}) > 0$ by Lemma 2. In particular,

$$\begin{aligned}&(\beta' - \beta^{(0)})^T (c_2/2) P_n(\beta' - \beta^{(0)})\\ &-(\beta' - \beta^{(0)})^T \left[A_n + B_n(\beta' - \beta^{(0)}) + \left|\left|\beta' - \beta^{(0)}\right|\right| J_n(\beta' - \beta^{(0)})\right] \le 0,\end{aligned}$$

and, writing

$$(I) = \left|\left|(c_2/2)^{-1} P_n^{-1} \left[A_n + B_n(\beta' - \beta^{(0)}) + \rho J_n(\beta' - \beta^{(0)})\right]\right|\right|^2$$

and

$$(II) = \frac{(d-1)^2 \left|\left|A_n + B_n(\beta' - \beta^{(0)}) + \rho J_n(\beta' - \beta^{(0)})\right|\right|^2}{L_1(n) L_2(n) (c_2/2)^2},$$

Lemma 7, Section 4, implies

$$\rho^2 = \left|\left|\beta' - \beta^{(0)}\right|\right|^2 \le (I) + (II). \tag{23}$$

We now proceed to show

$$(I) + (II) < U_n, \tag{24}$$

for some U_n, independent of β' and ρ, that satisfies

$$E\left[U_n\right] = O\left(\frac{\log(n)}{L_1(n)} + \frac{n(d-1)^2}{L_1(n)L_2(n)}\right).$$

Thus, if there is no solution $\beta \in B_\rho$ of $l_n(\beta) = 0$, then $\rho^2 < U_n$. This implies that there is always a solution $\beta \in B_{U_n^{1/2}}$ to $l_n(\beta) = 0$, and thus $||\hat{\beta}_n - \beta^{(0)}||^2 \mathbf{1}_{n > N_\rho} \le U_n$ a.s., and $E[||\hat{\beta}_n - \beta^{(0)}||^2 \mathbf{1}_{n > N_\rho}] \le E[U_n]$.

To prove (24), we decompose (I) and (II) using the following fact: if M, N are $d \times d$ matrices, and $N(j)$ denotes the j-th column of N, then

$$\begin{aligned} ||MN|| &= \max_{||y||=1} ||MNy|| = \max_{||y||=1} \left|\left| M \sum_{j=1}^{d} y[j]N(j) \right|\right| \\ &\leq \max_{||y||=1} \sum_{j=1}^{d} ||My[j]N(j)|| \leq \sum_{j=1}^{d} ||MN(j)|| \,. \end{aligned}$$

As a result we get

$$\begin{aligned} \left|\left| P_n^{-1} B_n(\beta' - \beta^{(0)}) \right|\right| &\leq \left|\left| P_n^{-1} \sum_{i=1}^{n} \dot{g}(x_i^T \beta^{(0)}) x_i e_i x_i^T \right|\right| \left|\left| \beta' - \beta^{(0)} \right|\right| \\ &\leq \rho \sum_{j=1}^{d} \left|\left| P_n^{-1} \sum_{i=1}^{n} \dot{g}(x_i^T \beta^{(0)}) x_i e_i x_i[j] \right|\right| \end{aligned}$$

and

$$\begin{aligned} \left|\left| P_n^{-1} J_n(\beta' - \beta^{(0)}) \right|\right| &\leq \left|\left| P_n^{-1} \sum_{i=1}^{n} c_1 x_i (|e_i| - E\left[|e_i| \mid \mathcal{F}_{i-1}\right]) x_i^T \right|\right| \left|\left| \beta' - \beta^{(0)} \right|\right| \\ &\leq \rho \sum_{j=1}^{d} \left|\left| P_n^{-1} \sum_{i=1}^{n} c_1 x_i (|e_i| - E\left[|e_i| \mid \mathcal{F}_{i-1}\right]) x_i[j] \right|\right| . \end{aligned}$$

In a similar vein we can derive

$$\left|\left| B_n(\beta' - \beta^{(0)}) \right|\right| \leq \rho \sum_{j=1}^{d} \left|\left| \sum_{i=1}^{n} \dot{g}(x_i^T \beta^{(0)}) x_i e_i x_i[j] \right|\right|$$

and

$$\left|\left| J_n(\beta' - \beta^{(0)}) \right|\right| \leq \rho \sum_{j=1}^{d} \left|\left| \sum_{i=1}^{n} c_1 x_i (|e_i| - E\left[|e_i| \mid \mathcal{F}_{i-1}\right]) x_i[j] \right|\right| .$$

It follows that

$$\begin{aligned} (I) &\leq 2(c_2/2)^{-2} \left(\left|\left| P_n^{-1} A_n \right|\right|^2 + \left|\left| P_n^{-1} B_n(\beta' - \beta^{(0)}) \right|\right|^2 \right) \\ &\quad + 2(c_2/2)^{-2} \rho_0^2 \left|\left| P_n^{-1} J_n(\beta' - \beta^{(0)}) \right|\right|^2 \\ &\quad \leq U_n(1) + U_n(2) + U_n(3), \end{aligned}$$

where we write

$$U_n(1) = 2(c_2/2)^{-2}\left|\left|P_n^{-1}A_n\right|\right|^2,$$
$$U_n(2) = 2(c_2/2)^{-2}\rho_0^2 2\left(\sum_{j=1}^{d}\left|\left|P_n^{-1}\sum_{i=1}^{n}\dot{g}(x_i^T\beta^{(0)})x_ie_ix_i[j]\right|\right|^2\right),$$
$$U_n(3) = 2(c_2/2)^{-2}\rho_0^4 2\left(\sum_{j=1}^{d}\left|\left|P_n^{-1}\sum_{i=1}^{n}c_1x_i(|e_i| - E\left[|e_i| \mid \mathcal{F}_{i-1}\right])x_i[j]\right|\right|^2\right),$$

and

$$(II) \leq U_n(4) + U_n(5) + U_n(6),$$

where we write

$$U_n(4) = \frac{2(d-1)^2\left|\left|A_n\right|\right|^2}{L_1(n)L_2(n)(c_2/2)^2},$$
$$U_n(5) = \frac{2(d-1)^2}{L_1(n)L_2(n)(c_2/2)^2}\left(\rho_0\sum_{j=1}^{d}\left|\left|\sum_{i=1}^{n}\dot{g}(x_i^T\beta^{(0)})x_ie_ix_i[j]\right|\right|\right)^2,$$
$$U_n(6) = \frac{2(d-1)^2\rho_0^4c_1^2}{L_1(n)L_2(n)(c_2/2)^2}\left(\sum_{j=1}^{d}\left|\left|\sum_{i=1}^{n}x_i(|e_i| - E\left[|e_i| \mid \mathcal{F}_{i-1}\right])x_i[j]\right|\right|\right)^2.$$

The desired upper bound U_n for $(I) + (II)$ equals $U_n = \sum_{j=1}^{6}U_n(j)$. For $U_n(1)$, $U_n(2)$, $U_n(3)$, apply Proposition 2 in Section 4 on the martingale difference sequences $(g(x_i^T\beta^{(0)})e_i)_{i\in\mathbb{N}}$, $(\dot{g}(x_i^T\beta^{(0)})x_i[j]e_i)_{i\in\mathbb{N}}$, and $(c_1(|e_i| - E[|e_i| \mid \mathcal{F}_{i-1}]x_i[j]))_{i\in\mathbb{N}}$, respectively. This implies the existence of a constant $K_1 > 0$ such that

$$E[U_n(1) + U_n(2) + U_n(3)] \leq \frac{K_1\log(n)}{L_1(n)}.$$

For $U_n(4)$, $U_n(5)$, $U_n(6)$, the assumption

$$\sup_{i\in\mathbb{N}} E\left[e_i^2 \mid \mathcal{F}_{i-1}\right] \leq \sigma^2 < \infty \text{ a.s.}$$

implies the existence of a constant $K_2 > 0$ such that

$$E\left[U_n(4) + U_n(5) + U_n(6)\right] \leq \frac{K_2n(d-1)^2}{L_1(n)L_2(n)}.$$

Proof of Corollary 2. It is sufficient to show that $H(\beta)$ is injective. Suppose $P_n^{-1/2}l_n(\beta) = P_n^{-1/2}l_n(\beta')$ for some β, β'. Since $n \geq n_0$ this implies $l_n(\beta) = l_n(\beta')$. By a first order Taylor expansion, there are $\tilde{\beta}_i$, $1 \leq i \leq n$, on the line segment between β and β' such that $l_n(\beta) - l_n(\beta') = \sum_{i=1}^n x_i x_i^T \dot{h}(x_i^T \tilde{\beta}_i)(\beta - \beta') = 0$. Since $\inf_{x \in X, \beta \in B_\rho} \dot{h}(x^T\beta) > 0$, Lemma 8 in Section 4 implies that the matrix $\sum_{i=1}^n x_i x_i^T \dot{h}(x_i^T \tilde{\beta}_i)$ is invertible, and thus $\beta = \beta'$.

Proof of Theorem 3. Let $0 < \rho \leq \rho_0$ and $n \geq N_\rho$. A Taylor expansion of $l_n(\beta)$ yields

$$\begin{aligned} l_n(\beta) - l_n(\beta^{(0)}) &= \sum_{i=1}^n x_i (h(x_i^T\beta^{(0)}) - h(x_i^T\beta)) \\ &= \sum_{i=1}^n x_i x_i^T \dot{h}(x_i^T \beta_{in})(\beta^{(0)} - \beta), \end{aligned}$$

for some β_{in}, $1 \leq i \leq n$, on the line segment between $\beta^{(0)}$ and β. Write $T_n(\beta) = \sum_{i=1}^n x_i x_i^T \dot{h}(x_i^T\beta_{in})$, and choose $k_2 > (\inf_{\beta \in B_\rho, x \in X} \dot{h}(x^T\beta))^{-1}$. Then for all $\beta \in B_\rho$,

$$\begin{aligned} \lambda_{\min}(k_2 T_n(\beta) - P_n) &= \lambda_{\min}\left(\sum_{i=1}^n x_i x_i^T (k_2 \dot{h}(x_i^T\beta_{in}) - 1)\right) \\ &\geq \left(\inf_{\beta \in B_{\rho_0}, x \in X} (k_2 \dot{h}(x^T\beta) - 1)\right) \lambda_{\min}(P_n), \end{aligned}$$

by Lemma 8. This implies

$$y^T k_2 T_n(\beta) y \geq y^T P_n y \quad \text{and } y^T k_2^{-1} T_n(\beta)^{-1} y \leq y^T P_n^{-1} y \quad \text{for all } y \in \mathbb{R}^d,$$

cf. Bhatia (2007, page 11, Exercise 1.2.12).

Define $H_n(\beta) = P_n^{-1/2} l_n(\beta)$, $r_n = ||H_n(\beta^{(0)})||$, and $\delta_n = \frac{r_n}{k_2^{-1}\sqrt{L_1(n)}}$. If $\delta_n > \rho$ then it follows immediately that $||\hat{\beta}_n - \beta^{(0)}|| \leq \rho < \frac{||H_n(\beta^{(0)})||}{k_2^{-1}\sqrt{L_1(n)}}$. Suppose $\delta_n \leq \rho$. Then for all $\beta \in \partial B_{\delta_n}$,

$$\begin{aligned} \left|\left|H_n(\beta) - H_n(\beta^{(0)})\right|\right|^2 &= \left|\left|P_n^{-1/2}(l_n(\beta) - l_n(\beta^{(0)}))\right|\right|^2 \\ &= (\beta^{(0)} - \beta)^T T_n(\beta) P_n^{-1} T_n(\beta)(\beta^{(0)} - \beta) \\ &\geq (\beta^{(0)} - \beta)^T T_n(\beta) k_2^{-1} T_n(\beta)^{-1} T_n(\beta)(\beta^{(0)} - \beta) \\ &\geq (\beta^{(0)} - \beta)^T P_n k_2^{-2} (\beta^{(0)} - \beta) \end{aligned}$$

$$\begin{aligned}&\geq k_2^{-2}\left|\left|\beta^{(0)}-\beta\right|\right|^2\lambda_{\min}(P_n)\\&\geq k_2^{-2}\delta_n^2 L_1(n),\end{aligned}$$

and thus we have $\inf_{\beta\in\partial B_{\delta_n}}||H_n(\beta)-H_n(\beta^{(0)})||\geq k_2^{-1}\sqrt{L_1(n)}\delta_n = r_n$ and $||H(\beta^{(0)})||\leq r_n$. By Corollary 2 we conclude that $||\hat{\beta}_n-\beta^{(0)}||\leq\frac{||H_n(\beta^{(0)})||}{k_2^{-1}\sqrt{L_1(n)}}$ a.s.

Now

$$E\left[\left|\left|H_n(\beta^{(0)})\right|\right|^2\right]=E\left[\left(\sum_{i=1}^n x_ie_i\right)^T P_n^{-1}\left(\sum_{i=1}^n x_ie_i\right)\right]=E[Q_n],$$

where Q_n is as in the proof of Proposition 2. There we show $E[Q_n]\leq K\log(n)$, for some $K>0$ and all $n\geq n_0$, and thus we have

$$E\left[\left|\left|\beta-\beta^{(0)}\right|\right|^2\mathbf{1}_{n\geq N_\rho}\right]=O\left(\frac{\log(n)}{L_1(n)}\right).$$

APPENDIX: AUXILIARY RESULTS

In this appendix, we prove and collect several probabilistic results which are used in the preceding sections. Proposition 1 is fundamental to Theorem 1, where we provide sufficient conditions such that the η-th moment of the last-time N_ρ is finite, for $\eta>0$. The proof of the proposition makes use of two auxiliary lemma's. Lemma 4 is a maximum inequality for tail probabilities of martingales; for sums of i.i.d. random variables this statement can be found e.g. in Loève (1977a, Section 18.1C, page 260), and a martingale version was already hinted at in Loève (1977b, Section 32.1, page 51). Lemma 5 contains a so-called Baum-Katz-Nagaev type theorem proven by Stoica (2007). There exists a long tradition of these type of results for sums of independent random variables, see e.g. Spataru (2009) and the references therein. Stoica (2007) makes an extension to martingales. In Proposition 2 we provide L^2 bounds for least-squares linear regression estimates, similar to the a.s. bounds derived by Lai and Wei (1982). The bounds for the quality of maximum quasi-likelihood estimates, Theorem 2 in Section 2 and Theorem 3 in Section 3, are proven by relating them to these bounds from Proposition 2. Lemma 6 is an auxiliary result used in the proof of Proposition 2. Finally, Lemma 7 is used in the proof of Theorem 2, and Lemma 8 in the proof of Theorem 3.

Lemma 4. *Let $(X_i)_{i\in\mathbb{N}}$ be a martingale difference sequence w.r.t. a filtration $\{\mathcal{F}_i\}_{i\in\mathbb{N}}$. Write $S_n=\sum_{i=1}^n X_i$, and suppose* $\sup_{i\in\mathbb{N}}E[X_i^2\mid\mathcal{F}_{i-1}]\leq$

$\sigma^2 < \infty$ *a.s., for some* $\sigma > 0$*. Then for all* $n \in \mathbb{N}$ *and* $\epsilon > 0$*,*

$$P\left(\max_{1\le k\le n} |S_k| \ge \epsilon\right) \le 2P\left(|S_n| \ge \epsilon - \sqrt{2\sigma^2 n}\right). \tag{25}$$

PROOF. We use similar techniques as de la Peña et al. (2009, Theorem 2.21, p.16), where (25) is proven for independent random variables $(X_i)_{i\in\mathbb{N}}$. Define the events $A_1 = \{S_1 \ge \epsilon\}$ and $A_k = \{S_k \ge \epsilon, S_1 < \epsilon, \dots, S_{k-1} < \epsilon\}$, $2 \le k \le n$. Then $A_k (1 \le k \le n)$ are mutually disjoint, and $\{\max_{1\le k\le n} S_k \ge \epsilon\} = \bigcup_{k=1}^n A_k$.

$$\begin{aligned}
&P\left(\max_{1\le k\le n} S_k \ge \epsilon\right)\\
&\le P\left(S_n \ge \epsilon - \sqrt{2\sigma^2 n}\right) + P\left(\max_{1\le k\le n} S_k \ge \epsilon, S_n < \epsilon - \sqrt{2\sigma^2 n}\right)\\
&\le P\left(S_n \ge \epsilon - \sqrt{2\sigma^2 n}\right) + \sum_{k=1}^n P\left(A_k, S_n < \epsilon - \sqrt{2\sigma^2 n}\right)\\
&\le P\left(S_n \ge \epsilon - \sqrt{2\sigma^2 n}\right) + \sum_{k=1}^n P\left(A_k, S_n - S_k < -\sqrt{2\sigma^2 n}\right)\\
&\overset{(1)}{=} P\left(S_n \ge \epsilon - \sqrt{2\sigma^2 n}\right) + \sum_{k=1}^n E\left[\mathbf{1}_{A_k} E\left[\mathbf{1}_{S_n - S_k < -\sqrt{2\sigma^2 n}} \mid \mathcal{F}_k\right]\right]\\
&\overset{(2)}{\le} P\left(S_n \ge \epsilon - \sqrt{2\sigma^2 n}\right) + \sum_{k=1}^n \frac{1}{2} P\left(A_k\right)\\
&= P\left(S_n \ge \epsilon - \sqrt{2\sigma^2 n}\right) + \frac{1}{2} P\left(\max_{1\le k\le n} S_k \ge \epsilon\right),
\end{aligned}$$

where (1) uses $A_k \in \mathcal{F}_k$, and (2) uses $E[\mathbf{1}_{S_n - S_k < -\sqrt{2\sigma^2 n}} \mid \mathcal{F}_k] = P(S_k - S_m > \sqrt{2\sigma^2 n} \mid \mathcal{F}_k) \le E[(S_n - S_k)^2 \mid \mathcal{F}_k]/(2\sigma^2 n) \le 1/2$ a.s. This proves $P(\max_{1\le k\le n} S_k \ge \epsilon) \le 2P(S_n \ge \epsilon - \sqrt{2\sigma^2 n})$. Replacing S_k by $-S_k$ gives $P(\max_{1\le k\le n} -S_k \ge \epsilon) \le 2P(-S_n \ge \epsilon - \sqrt{2\sigma^2 n})$. If $\epsilon - \sqrt{2\sigma^2 n} \le 0$ then (25) is trivial; if $\epsilon > \sqrt{2\sigma^2 n}$ then

$$\begin{aligned}
P\left(\max_{1\le k\le n} |S_k| \ge \epsilon\right) &\le P\left(\max_{1\le k\le n} S_k \ge \epsilon\right) + P\left(\max_{1\le k\le n} -S_k \ge \epsilon\right)\\
&\le 2P\left(S_n \ge \epsilon - \sqrt{2\sigma^2 n}\right) + 2P\left(-S_n \ge \epsilon - \sqrt{2\sigma^2 n}\right)\\
&= 2P\left(|S_n| \ge \epsilon - \sqrt{2\sigma^2 n}\right).
\end{aligned}$$

□

LEMMA 5 (Stoica, 2007). *Let $(X_i)_{i\in\mathbb{N}}$ be a martingale difference sequence w.r.t. a filtration $\{\mathcal{F}_i\}_{i\in\mathbb{N}}$. Write $S_n = \sum_{i=1}^n X_i$ and suppose $\sup_{i\in\mathbb{N}} E[X_i^2 \mid \mathcal{F}_{i-1}] \leq \sigma^2 < \infty$ a.s. for some $\sigma > 0$. Let $c > 0$, $\frac{1}{2} < \alpha \leq 1$, $\eta > 2\alpha - 1$, $r > \frac{\eta+1}{\alpha}$. If $\sup_{i\in\mathbb{N}} E[|X_i|^r] < \infty$, then*

$$\sum_{k\geq 1} k^{\eta-1} P\left(|S_k| \geq ck^\alpha\right) < \infty.$$

PROPOSITION 1. *Let $(X_i)_{i\in\mathbb{N}}$ be a martingale difference sequence w.r.t. a filtration $\{\mathcal{F}_i\}_{i\in\mathbb{N}}$. Write $S_n = \sum_{i=1}^n X_i$ and suppose $\sup_{i\in\mathbb{N}} E[X_i^2 \mid \mathcal{F}_{i-1}] \leq \sigma^2 < \infty$ a.s. for some $\sigma > 0$. Let $c > 0$, $\frac{1}{2} < \alpha \leq 1$, $\eta > 2\alpha - 1$, $r > \frac{\eta+1}{\alpha}$, and define the random variable $T = \sup\{n \in \mathbb{N} \mid |S_n| \geq cn^\alpha\}$, where T takes values in $\mathbb{N} \cup \{\infty\}$. If $\sup_{i\in\mathbb{N}} E[|X_i|^r] < \infty$, then*

$$T < \infty \text{ a.s.}, \quad \text{and } E\left[T^\eta\right] < \infty.$$

PROOF. There exists an $n' \in \mathbb{N}$ such that for all $n > n'$, $c(n/2)^\alpha - \sqrt{2\sigma^2 n} \geq c(n/2)^\alpha/2$. For all $n > n'$,

$$\begin{aligned}
P\left(T > n\right) &= P\left(\exists k > n : |S_k| \geq ck^\alpha\right) \\
&\leq \sum_{j\geq\lfloor\log_2(n)\rfloor} P\left(\exists 2^{j-1} \leq k < 2^j : |S_k| \geq ck^\alpha\right) \\
&\leq \sum_{j\geq\lfloor\log_2(n)\rfloor} P\Big(\sup_{1\leq k\leq 2^j} |S_k| \geq c(2^{j-1})^\alpha\Big) \\
&\overset{(1)}{\leq} 2 \sum_{j\geq\lfloor\log_2(n)\rfloor} P\left(|S_{2^j}| \geq c(2^{j-1})^\alpha - \sqrt{2\sigma^2 2^j}\right) \\
&\overset{(2)}{\leq} 2 \sum_{j\geq\lfloor\log_2(n)\rfloor} P\left(|S_{2^j}| \geq c(2^{j-1})^\alpha/2\right).
\end{aligned}$$

where (1) follows from Lemma 4 and (2) from the definition of n'.

For $t \in \mathbb{R}_+$ write $S_t = S_{\lfloor t\rfloor}$. Then

$$\sum_{j\geq\log_2(n)} P\left(|S_{2^j}| \geq c(2^{j-1})^\alpha/2\right) = \int_{j\geq\log_2(n)} P\left(|S_{2^j}| \geq c(2^{j-1})^\alpha/2\right) dj \tag{26}$$

$$= \int_{k\geq n} \frac{P\left(|S_k| \geq c(k/2)^\alpha/2\right)}{k\log(2)} dk = \sum_{k\geq n} P\left(|S_k| \geq c(k/2)^\alpha/2\right) \frac{1}{k\log(2)}, \tag{27}$$

using a variable substitution $k = 2^j$.

By Chebyshev's inequality,

$$\begin{aligned} P\left(T > n\right) &\leq 2 \sum_{k \geq n} P\left(|S_k| \geq c(k/2)^{\alpha}/2\right) \frac{1}{k \log(2)} \\ &\leq 2 \sum_{k \geq n} \sigma^2 k (c(k/2)^{\alpha}/2)^{-2} \frac{1}{k \log(2)}, \end{aligned}$$

which implies $P(T = \infty) \leq \liminf_{n \to \infty} P(T > n) = 0$. This proves $T < \infty$ a.s.

Since

$$\begin{aligned} E[T^{\eta}] &\leq \eta \left[1 + \sum_{n \geq 1} n^{\eta - 1} P(T > n) \right] \\ &\leq \eta \left[1 + n' \cdot (n')^{\eta-1} + \sum_{n > n'} n^{\eta-1} P(T > n) \right] \\ &\leq M \sum_{n > n'} n^{\eta - 1} \sum_{j \geq \lfloor \log_2(n) \rfloor} P\left(|S_{2^j}| \geq c(2^{j-1})^{\alpha}/2 \right), \end{aligned}$$

for some constant $M > 0$, it follows by (26), (27) that $E[T^{\eta}] < \infty$ if

$$\sum_{n \geq 1} n^{\eta-1} \sum_{k \geq n} P\left(|S_k| \geq c(k/2)^{\alpha}/2\right) k^{-1} < \infty.$$

By interchanging the sums, it suffices to show

$$\sum_{k \geq 1} k^{\eta - 1} P\left(|S_k| \geq 2^{-1-\alpha} c k^{\alpha} \right) < \infty.$$

This last statement follows from Lemma 5. □

Let $(e_i)_{i \in \mathbb{N}}$ be a martingale difference sequence w.r.t. a filtration $\{\mathcal{F}_i\}_{i \in \mathbb{N}}$, such that $\sup_{i \in \mathbb{N}} E[e_i^2 \mid \mathcal{F}_{i-1}] = \sigma^2 < \infty$ a.s., for some $\sigma > 0$. Let $(x_i)_{i \in \mathbb{N}}$ be a sequence of vectors in $\mathbb{R}^d$. Assume that $(x_i)_{i \in \mathbb{N}}$ are predictable w.r.t. the filtration (i.e. $x_i \in \mathcal{F}_{i-1}$ for all $i \in \mathbb{N}$), and $\sup_{i \in \mathbb{N}} ||x||_i \leq M < \infty$ for some (non-random) $M > 0$. Write $P_n = \sum_{i=1}^n x_i x_i^T$. Let $L : \mathbb{N} \to \mathbb{R}_+$ be a (non-random) function and $n_0 \geq 2$ a (non-random) integer such that $\lambda_{\min}(P_n) \geq L(n)$ for all $n \geq n_0$, and $\lim_{n \to \infty} L(n) = \infty$.

Proposition 2. *There is a constant $K > 0$ such that for all $n \geq n_0$,*

$$E \left|\left| \left(\sum_{i=1}^n x_i x_i^T \right)^{-1} \sum_{i=1}^n x_i e_i \right|\right|^2 \leq K \frac{\log(n)}{L(n)}.$$

The proof of Proposition 2 uses the following result:

LEMMA 6. *Let* $(y_n)_{n\in\mathbb{N}}$ *be a nondecreasing sequence with* $y_1 \geq e$. *Write* $R_n = \frac{1}{\log(y_n)}\sum_{i=1}^n \frac{y_i - y_{i-1}}{y_i}$, *where we put* $y_0 = 0$. *Then* $R_n \leq 2$ *for all* $n \in \mathbb{N}$.

PROOF. Induction on n. $R_1 = \frac{1}{\log(y_1)} \leq 1 \leq 2$. Let $n \geq 2$ and define $g(y) = \frac{1}{\log(y)}\frac{y - y_{n-1}}{y} + \frac{\log(y_{n-1})}{\log(y)}R_{n-1}$. If $R_{n-1} \leq 1$, then $R_n = g(y_n) \leq \frac{1}{\log(y_n)} + 1 \leq 2$. Now suppose $R_{n-1} > 1$. Since $z \mapsto (1 + \log(z))/z$ is decreasing in z on $z \geq 1$, and since $y_{n-1} \geq 1$, we have $(1+\log(y))/y \leq (1+\log(y_{n-1}))/y_{n-1}$ for all $y \geq y_{n-1}$. Together with $R_{n-1} > 1$ this implies

$$\frac{\partial g(y)}{\partial y} = \frac{1}{y(\log(y))^2}\left[-1 + \frac{y_{n-1}}{y}(1 + \log(y)) - \log(y_{n-1})R_{n-1}\right] < 0,$$

for all $y \geq y_{n-1}$. This proves $R_n = g(y_n) \leq \max_{y \geq y_{n-1}} g(y) = g(y_{n-1}) = R_{n-1} \leq 2$. □

PROOF OF PROPOSITION 2. Write $q_n = \sum_{i=1}^n x_i e_i$ and $Q_n = q_n P_n^{-1} q_n$. For $n \geq n_0$, P_n is invertible, and

$$\begin{aligned} \left|\left|P_n^{-1}q_n\right|\right|^2 &\leq \left|\left|P_n^{-1/2}\right|\right|^2 \cdot \left|\left|P_n^{-1/2}q_n\right|\right|^2 \leq \lambda_{\min}(P_n)^{-1}q_n P_n^{-1} q_n \\ &\leq L(n)^{-1}Q_n \text{ a.s.,} \end{aligned}$$

where we used $||P_n^{-1/2}|| = \lambda_{\max}(P_n^{-1/2}) = \lambda_{\min}(P_n)^{-1/2}$. We show $E[Q_n] \leq K\log(n)$, for a constant K to be defined further below, and all $n \geq n_0$.

Write $V_n = P_n^{-1}$. Since $P_n = P_{n-1} + x_n x_n^T$, it follows from the Sherman-Morrison formula (Bartlett, 1951) that $V_n = V_{n-1} - \frac{V_{n-1}x_n x_n^T V_{n-1}}{1 + x_n^T V_{n-1} x_n}$, and thus

$$x_n^T V_n = x_n^T V_{n-1} - \frac{(x_n^T V_{n-1} x_n) x_n^T V_{n-1}}{1 + x_n^T V_{n-1} x_n} = x_n^T V_{n-1}/(1 + x_n^T V_{n-1} x_n).$$

As in Lai and Wei (1982), Q_n satisfies

$$\begin{aligned} Q_n &= \left(\sum_{i=1}^n x_i^T e_i\right) V_n \left(\sum_{i=1}^n x_i e_i\right) \\ &= \left(\sum_{i=1}^{n-1} x_i^T e_i\right) V_n \left(\sum_{i=1}^{n-1} x_i e_i\right) + x_n^T V_n x_n e_n^2 + 2x_n^T V_n \left(\sum_{i=1}^{n-1} x_i e_i\right) e_n \\ &= Q_{n-1} + \left(\sum_{i=1}^{n-1} x_i^T e_i\right)\left(-\frac{V_{n-1}x_n x_n^T V_{n-1}}{1 + x_n^T V_{n-1} x_n}\right)\left(\sum_{i=1}^{n-1} x_i e_i\right) \end{aligned}$$

$$\begin{aligned}
&+ x_n^T V_n x_n e_n^2 + 2\frac{x_n^T V_{n-1}}{1+x_n^T V_{n-1} x_n}\left(\sum_{i=1}^{n-1} x_i e_i\right) e_n \\
&= Q_{n-1} - \frac{(x_n^T V_{n-1}\sum_{i=1}^{n-1} x_i e_i)^2}{1+x_n^T V_{n-1} x_n} + x_n^T V_n x_n e_n^2 \\
&+ 2\frac{x_n^T V_{n-1}}{1+x_n^T V_{n-1} x_n}\left(\sum_{i=1}^{n-1} x_i e_i\right) e_n.
\end{aligned}$$

Observe that

$$E\left[\frac{x_n^T V_{n-1}\left(\sum_{i=1}^{n-1} x_i e_i\right)}{1+x_n^T V_{n-1} x_n} e_n\right] = E\left[\frac{x_n^T V_{n-1}\left(\sum_{i=1}^{n-1} x_i e_i\right)}{1+x_n^T V_{n-1} x_n} E\left[e_n \mid \mathcal{F}_{n-1}\right]\right] = 0$$

and

$$E\left[x_n^T V_n x_n e_n^2\right] = E\left[x_n^T V_n x_n E\left[e_n^2 \mid \mathcal{F}_{n-1}\right]\right] \leq E\left[x_n^T V_n x_n\right]\sigma^2.$$

By telescoping the sum we obtain

$$E[Q_n] \leq E[Q_{\min\{n,n_1\}}] + \sigma^2 \sum_{i=n_1+1}^{n} E[x_i^T V_i x_i],$$

where we define $n_1 \in \mathbb{N}$ to be the smallest $n \geq n_0$ such that $L(n) > e^{1/d}$ for all $n \geq n_1$. We have

$$\begin{aligned}
\det(P_{n-1}) &= \det(P_n - x_n x_n^T) \\
&= \det(P_n)\det(I - P_n^{-1} x_n x_n^T) \qquad (28)\\
&= \det(P_n)(1 - x_n^T V_n x_n), \quad (n \geq n_1).
\end{aligned}$$

Here the last equality follows from Sylvester's determinant theorem $\det(I + AB) = \det(I + BA)$, for matrices A, B of appropriate size. We thus have $x_n^T V_n x_n = \frac{\det(P_n)-\det(P_{n-1})}{\det(P_n)}$. For $n \in \mathbb{N}$ let $y_n = \det(P_{n+n_1})$. Then $(y_n)_{n\in\mathbb{N}}$ is a nondecreasing sequence with

$$y_1 \geq \det(P_{n_1+1}) \geq \lambda_{\min}(P_{n_1+1})^d \geq e.$$

Lemma 6 implies

$$\begin{aligned}
\sum_{i=n_1+1}^{n} x_i^T V_i x_i = \sum_{i=n_1+1}^{n} \frac{y_{i-n_1} - y_{i-1-n_1}}{y_{i-n_1}} = \sum_{i=1}^{n-n_1} \frac{y_i - y_{i-1}}{y_i} \\
\leq 2\log(y_{n-n_1}) = 2\log(\det(P_n)), \text{ a.s.}
\end{aligned}$$

Now

$$\begin{aligned}\log(\det(P_n)) &\leq d\log(\lambda_{\max}(P_n)) \leq d\log(\operatorname{tr}(P_n)) \leq d\log(n\sup_{i\in\mathbb{N}}||x_i||^2)\\ &\leq d\log(nM^2).\end{aligned}$$

Furthermore, for all $n_0 \leq n \leq n_1$ we have

$$\begin{aligned}E\left[Q_n\right] &\leq E\left[||q_n||^2\,\lambda_{\max}(P_n^{-1})\right] \leq E\left[\left|\left|\sum_{i=1}^n x_i\epsilon_i\right|\right|^2 L(n_0)^{-1}\right]\\ &\leq L(n_0)^{-1}E\left[2\sum_{i=1}^n \epsilon_i^2\sup_{i\in\mathbb{N}}||x_i||^2\right]\\ &\leq 2L(n_0)^{-1}M^2n_1\sigma^2,\end{aligned}$$

and thus for all $n \geq n_0$,

$$\begin{aligned}E\left[Q_n\right] &\leq E\left[Q_{\min\{n,n_1\}}\right] + \sigma^2\sum_{i=n_1+1}^n E\left[x_i^TV_ix_i\right]\\ &\leq 2L(n_0)^{-1}M^2n_1\sigma^2 + d\log(n) + d\log(M^2)\\ &\leq K\log(n),\end{aligned}$$

where $K = d + [2L(n_0)^{-1}M^2n_1\sigma^2 + d\log(M^2)]/\log(n_0)$. □

LEMMA 7. *Let A be a positive definite $d\times d$ matrix, and b, $x \in \mathbb{R}^d$. If $x^TAx + x^Tb \leq 0$ then $||x||^2 \leq ||A^{-1}b||^2 + (d-1)^2\frac{||b||^2}{\lambda_1\lambda_2}$, where $0 < \lambda_1 \leq \lambda_2$ are the two smallest eigenvalues of A.*

PROOF. Let $0 < \lambda_1 \leq \cdots \leq \lambda_d$ be the eigenvalues of A, and $v_1, \ldots, v_d$ the corresponding eigenvectors. We can assume that these form an orthonormal basis, such that each $x \in \mathbb{R}^d$ can be written as $\sum_{i=1}^d \alpha_iv_i$, for coordinates $(\alpha_1, \ldots, \alpha_d)$, and $b = \sum_{i=1}^d \beta_iv_i$ for some $(\beta_1, \ldots, \beta_d)$. Write

$$S = \left\{(\alpha_1, \ldots, \alpha_d) \mid \sum_{i=1}^d \alpha_i(\lambda_i\alpha_i + \beta_i) \leq 0\right\}.$$

The orthonormality of $(v_i)_{1\leq i\leq d}$ implies $S = \{x \in \mathbb{R}^d \mid x^TAx + x^Tb \leq 0\}$.

Fix $\alpha = (\alpha_1, \ldots, \alpha_d) \in S$ and write $R = \{i \mid \alpha_i(\lambda_i\alpha_i+\beta_i) \leq 0, 1 \leq i \leq d\}$, $R^c = \{1, \ldots, d\}\backslash R$. For all $i \in R$, standard properties of quadratic equations imply $\alpha_i^2 \leq \lambda_i^{-2}\beta_i^2$ and $\alpha_i(\lambda_i\alpha_i + \beta_i) \geq \frac{-\beta_i^2}{4\lambda_i}$. For all $i \in R^c$,

$$\alpha_i(\lambda_i\alpha_i + \beta_i) \leq \sum_{i\in R^c}\alpha_i(\lambda_i\alpha_i + \beta_i) \leq -\sum_{i\in R}\alpha_i(\lambda_i\alpha_i + \beta_i) \leq c,$$

where we define $c = \sum_{i \in R} \frac{\beta_i^2}{4\lambda_i}$. By the quadratic formula, $\alpha_i(\lambda_i \alpha_i + \beta_i) - c \leq 0$ implies

$$\frac{-\beta_i - \sqrt{\beta_i^2 + 4\lambda_i c}}{2\lambda_i} \leq \alpha_i \leq \frac{-\beta_i + \sqrt{\beta_i^2 + 4\lambda_i c}}{2\lambda_i}.$$

(Note that $\lambda_i > 0$ and $c > 0$ implies that the square root is well-defined). It follows that

$$\alpha_i^2 \leq 2\frac{\beta_i^2 + \beta_i^2 + 4\lambda_i c}{4\lambda_i^2} = \frac{\beta_i^2}{\lambda_i^2} + 2c/\lambda_i, \quad (i \in R^c),$$

and thus

$$\begin{aligned} ||x||^2 = \sum_{i=1}^d \alpha_i^2 &\leq \sum_{i \in R} \lambda_i^{-2}\beta_i^2 + \sum_{i \in R^c} \left(\frac{\beta_i^2}{\lambda_i^2} + \frac{2}{\lambda_i} \sum_{j \in R} \frac{\beta_j^2}{4\lambda_j} \right) \\ &\leq \sum_{i=1}^d \lambda_i^{-2}\beta_i^2 + \frac{1}{2} \left(\sum_{i \in R^c} \frac{1}{\lambda_i} \right) \left(\sum_{j \in R} \frac{1}{\lambda_j} \right) \left(\sum_{i=1}^n \beta_i^2 \right) \\ &\leq \left|\left|A^{-1}b\right|\right|^2 + (d-1)^2 \frac{1}{\lambda_1}\frac{1}{\lambda_2} ||b||^2, \end{aligned}$$

where we used $||A^{-1}b||^2 = \sum_{j=1}^d \beta_j^2 \lambda_j^{-2}$ and $(\sum_{i \in R^c} 1)(\sum_{j \in R} 1) \leq 2(d-1)^2$. □

REMARK 6. The dependence on $\lambda_1\lambda_2$ in Lemma 7 is tight in the following sense: for all $d \geq 2$ and all positive definite $d \times d$ matrices A there are $x \in \mathbb{R}^d$, $b \in \mathbb{R}^d$ such that $x^T A x + x^T b \leq 0$ and

$$||x||^2 \geq \frac{1}{8} \left(||A^{-1}b|| + \frac{||b||^2}{\lambda_1 \lambda_2} \right).$$

In particular, choose $\beta_1 = \beta_2 > 0$, $\alpha_1 = -\beta_1/(2\lambda_1)$, and $\alpha_2 = (-\beta_2 - \sqrt{\beta_2^2 + 4\lambda_2\beta_1^2/(4\lambda_1)})/(2\lambda_2)$, and set $b = \beta_1 v_1 + \beta_2 v_2$ and $x = \alpha_1 v_1 + \alpha_2 v_2$, where v_1, v_2 are the eigenvectors of A corresponding to eigenvalues λ_1, λ_2. Then $x^T A x + x^T b = \sum_{i=1}^2 \alpha_i(\lambda_i \alpha_i + \beta_i) = 0$ and

$$\begin{aligned} ||x||^2 = \alpha_1^2 + \alpha_2^2 &\geq \beta_1^2/(4\lambda_1^2) + \beta_2^2/(4\lambda_2^2) + \beta_1^2/(4\lambda_1\lambda_2) \\ &\geq \frac{1}{8}||A^{-1}b||^2 + ||b||^2/(8\lambda_1\lambda_2). \end{aligned}$$

LEMMA 8. *Let $(x_i)_{i \in \mathbb{N}}$ be a sequence of vectors in $\mathbb{R}^d$, and $(w_i)_{i \in \mathbb{N}}$ a sequence of scalars with $0 < \inf_{i \in \mathbb{N}} w_i$. Then for all $n \in \mathbb{N}$,*

$$\lambda_{min}\left(\sum_{i=1}^n x_i x_i^T w_i\right) \geq \lambda_{min}\left(\sum_{i=1}^n x_i x_i^T\right)(\inf_{i \in \mathbb{N}} w_i).$$

PROOF. For all $z \in \mathbb{R}^d$,

$$z^T\left(\sum_{i=1}^n x_i x_i^T w_i\right) z \geq (\inf_{i \in \mathbb{N}} w_i) z^T\left(\sum_{i=1}^n x_i x_i^T\right) z.$$

Let $\tilde{v}$ be a normalized eigenvector corresponding to $\lambda_{\min}(\sum_{i=1}^n x_i x_i^T w_i)$. Then

$$\begin{aligned}\lambda_{\min}\left(\sum_{i=1}^n x_i x_i^T\right) &= \min_{||v||=1} v^T\left(\sum_{i=1}^n x_i x_i^T\right) v \leq \tilde{v}^T\left(\sum_{i=1}^n x_i x_i^T\right)\tilde{v} \\ &\leq \tilde{v}^T\left(\sum_{i=1}^n x_i x_i^T w_i\right)\tilde{v}(\inf_{i \in \mathbb{N}} w_i)^{-1} \\ &= \lambda_{\min}\left(\sum_{i=1}^n x_i x_i^T w_i\right)(\inf_{i \in \mathbb{N}} w_i)^{-1}.\end{aligned}$$

□

REFERENCES

ANDERSON, T. W. and TAYLOR, J. B., Some experimental results on the statistical properties of least squares estimates in control problems. *Econometrica*, 44(6): 1289–1302, 1976.

ARAMAN, V. F. and CALDENTEY, R., Revenue Management with Incomplete Demand Information. In J. J. Cochran, editor, *Encyclopedia of Operations Research*. Wiley, 2011.

BARTLETT, M. S., An inverse matrix adjustment arising in discriminant analysis. *The Annals of Mathematical Statistics*, 22(1): 107–111, 1951. MR0040068

BESBES, O. and ZEEVI, A., Dynamic pricing without knowing the demand function: risk bounds and near-optimal algorithms. *Operations Research*, 57(6): 1407–1420, 2009. MR2597918

BHATIA, R., *Positive Definite Matrices*. Princeton University Press, Princeton, 2007. MR2284176

BRODER, J. and RUSMEVICHIENTONG, P., Dynamic pricing under a general parametric choice model. *Operations Research*, 60(4): 965–980, 2012. MR2979434

CHANG, Y. I., Strong consistency of maximum quasi-likelihood estimate in generalized linear models via a last time. *Statistics & Probability Letters*, 45(3): 237–246, 1999. MR1718035

Chen, K., Hu, I., and Ying, Z., Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *The Annals of Statistics*, 27(4): 1155–1163, 1999. MR1740117

Chow, Y. S. and Teicher, H., *Probability Theory: Independence, Interchangeability, Martingales*. Springer Verlag, New York, third edition, 2003.

de la Peña, V. H., Lai, T. L., and Shao, Q. M., *Self-Normalized Processes: Limit Theory and Statistical Applications*. Springer Series in Probability and its Applications. Springer, New York, first edition, 2009. MR2488094

den Boer, A. V., Dynamic pricing with multiple products and partially specified demand distribution. *Mathematics of Operations Research*, Forthcoming, 2013.

den Boer, A. V. and Zwart, B., Simultaneously learning and optimizing using controlled variance pricing. *Management Science*, Forthcoming, 2013.

Dugundji, J., *Topology*. Allyn and Bacon, Boston, 1966. MR0193606

Fahrmeir, L. and Kaufmann, H., Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13(1): 342–368, 1985. MR0773172

Gill, J., *Generalized Linear Models: A Unified Approach*. Sage Publications, Thousand Oaks, CA, 2001.

Goldenshluger, A. and Zeevi, A., Woodroofe's one-armed bandit problem revisited. *The Annals of Applied Probability*, 19(4): 1603–1633, 2009. MR2538082

Heyde, C. C., *Quasi-Likelihood and Its Application*. Springer Series in Statistics. Springer Verlag, New York, 1997. MR1461808

Keskin, N. B. and Zeevi, A., Dynamic pricing with an unknown linear demand model: asymptotically optimal semi-myopic policies. Working paper, University of Chicago, `http://faculty.chicagobooth.edu/bora.keskin/pdfs/DynamicPricingUnknownDemandModel.pdf`, 2013.

Lai, T. L. and Robbins, H., Iterated least squares in multiperiod control. *Advances in Applied Mathematics*, 3(1): 50–73, 1982. MR0646499

Lai, T. L. and Wei, C. Z., Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1): 154–166, 1982. MR0642726

Leray, J. and Schauder, J., Topologie et equations fonctionelles. *Annales Scientifiques de l'École Normale Supérieure*, 51: 45–78, 1934. MR1509338

Loève, M., *Probability Theory I*. Springer Verlag, New York, Berlin, Heidelberg, 4th edition, 1977a. MR0651017

Loève, M., *Probability Theory II*. Springer Verlag, New York, Berlin, Heidelberg, 4th edition, 1977b. MR0651017

McCullagh, P., Quasi-likelihood functions. *The Annals of Statistics*, 11(1): 59–67, 1983. MR0684863

McCullagh, P. and Nelder, J. A., *Generalized Linear Models*. Chapman & Hall, London, 1983. MR0727836

Nelder, J. A. and Wedderburn, R. W. M., Generalized linear models. *Journal of the Royal Statistical Society, Series A (General)*, 135(3): 370–384, 1972.

Ortega, J. M. and Rheinboldt, W. C., *Iterative Solution of Nonlinear Equations in Several Variables*, volume 30 of *SIAM's Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, 2000. MR1744713

Pronzato, L., Optimal experimental design and some related control problems. *Automatica*, 44(2): 303–325, 2008. MR2530779

Rusmevichientong, P. and Tsitsiklis, J. N., Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2): 395–411, 2010. MR2674726

Small, C. G., Wang, J., and Yang, Z., Eliminating multiple root problems in estimation. *Statistical Science*, 15(4): 313–332, 2000. MR1819708

Spataru, A., Improved convergence rates for tail probabilities. *Bulletin of the Transilvania University of Brasov – Series III: Mathematics, Informatics, Physics*, 2(51): 137–142, 2009. MR2642502

Stoica, G., Baum-Katz-Nagaev type results for martingales. *Journal of Mathematical Analysis and Applications*, 336(2): 1489–1492, 2007. MR2353031

Tzavelas, G., A note on the uniqueness of the quasi-likelihood estimator. *Statistics & Probability Letters*, 38(2): 125–130, 1998. MR1627914

Wedderburn, R. W. M., Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61(3): 439–447, 1974. MR0375592

Yin, C., Zhang, H., and Zhao, L., Rate of strong consistency of maximum quasi-likelihood estimator in multivariate generalized linear models. *Communications in Statistics – Theory and Methods*, 37(19): 3115–3123, 2008. MR2467755

Yue, L. and Chen, X., Rate of strong consistency of quasi maximum likelihood estimate in generalized linear models. *Science in China Series A: Mathematics*, 47(6): 882–893, 2004. MR2127216

Zhang, S. and Liao, Y., On some problems of weak consistency of quasi-maximum likelihood estimates in generalized linear models. *Science in China Series A: Mathematics*, 51(7): 1287–1296, 2008. MR2417495

Zhang, S., Liao, Y., and Ning, W., Asymptotic properties of quasi-maximum likelihood estimates in generalized linear models. *Communications in Statistics – Theory and Methods*, 40(24): 4417–4430, 2011. MR2864166

Zhu, C. and Gao, Q., Asymptotic properties in generalized linear models with natural link function and adaptive designs. *Advances in Mathematics (China)*, 42(1): 121–127, 2013. MR3098890

University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands
E-mail: a.v.denboer@utwente.nl

Centrum Wiskunde & Informatica (CWI)
Science Park 123
1098 XG Amsterdam
The Netherlands
E-mail: bert.zwart@cwi.nl