Mean square convergence rates for maximum quasi-likelihood estimators

Arnoud den Boer, Bert Zwart

1 Centrum Wiskunde & Informatica (CWI)
Science Park 123, 1098 XG Amsterdam, The Netherlands

2 VU University Amsterdam, Department of Mathematics
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands,

September 28, 2012

Abstract

We study the behavior of maximum quasi-likelihood 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. Because the MQLE may not always exist, we consider the last time that the quasi-likelihood equation has no solution in a neighborhood of the true (but unknown) parameter, and provide conditions which guarantee that this last-time has finite moments. We use this to show asymptotic existence and strong consistency of the MQLE, and obtain bounds on the mean square convergence rates. If the dimension of the unknown parameter is at most two, or if the link function is canonical, these bounds coincide with known a.s. bounds on the convergence rates for least-squares linear regression. 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

\[ E[Y(x)] = h(x^T \beta^{(0)}), \quad \text{Var}(Y(x)) = v(E[Y(x)]), \] (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 \( (x_i)_{1 \leq i \leq n} \) and observed responses \( (y_i)_{1 \leq i \leq n} \), where each \( y_i \) is a realization of the
random variable $Y(x_i)$, 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

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

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 $\beta$ 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 paper, 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), den Boer and Zwart (2010), Araman and Caldentey (2011), Broder and Rusmevichientong (2010)). 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
   $$P_n = \sum_{i=1}^{n} x_i x_i^T,$$
and denote by $\lambda_{\min}(P_n)$, $\lambda_{\max}(P_n)$ the smallest and largest eigenvalues of $P_n$. Bounds on $||\hat{\beta}_n - \beta(0)||^2$ 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 $h = v \circ h$, otherwise it is called a general or non-canonical link function. For canonical link functions, the quasi-likelihood equations (2) 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). Chang 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

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

Their proof however is reported to contain a mistake, see Zhang and Liao (2008, page 1289). Chang (1999) extends these convergence rates for adaptive designs to general link functions, under the additional condition $\lambda_{\min}(P_n) = \Omega(n^{\alpha})$ a.s. for some $\alpha > 1/2$. His proof however also appears to contain a mistake, see Remark 1. Yin et al. (2008) extends the setting of Chang (1999), with adaptive design and general link function, to multivariate response data. They obtain strong consistency and convergence rates

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

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. A recent study restricted to fixed designs and canonical link functions is Zhang and Liao (2008), who show $||\hat{\beta}_n - \beta(0)|| = O_P(\lambda_{\min}(P_n)^{-1/2})$, provided $\lambda_{\min}(P_n) = \Omega(\lambda_{\max}(P_n)^{1/2})$ a.s. and other regularity assumptions.

### 1.3 Assumptions and contributions

In contrast with the above-mentioned literature, 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 $\{F_i\}_{i \in \mathbb{N}}$ 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 $\sup_{i \in \mathbb{N}} E\left[ e_i^4 \mid F_{i-1} \right] \leq \sigma^2 < \infty$ a.s. for some $\sigma > 0$, and $\sup_{i \in \mathbb{N}} E\left[ |e_i|^r \right] < \infty$ for some $r > 2$.

We consider general link functions, and only assume that $h$ and $v$ are thrice continuously differentiable with $h(z) > 0$, $v(h(z)) > 0$ 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 \( P_n \) 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

\[
L_1(n) \geq cn^\alpha, \quad \text{for some } c > 0, \frac{1}{2} < \alpha \leq 1 \text{ independent of } n. \tag{4}
\]

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 \( \beta^{(0)} \):

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

For all sufficiently small \( \rho > 0 \), we show in Theorem 1 that \( N_\rho \) is finite a.s., and provide sufficient conditions such that \( E[N_\rho] < \infty \), for \( \eta > 0 \).

2. In Theorem 2, we provide the upper bound

\[
E \left[ \left|\hat{\beta}_n - \beta^{(0)}\right|^2 1_{n > N_\rho} \right] = O \left( \frac{\log(n)}{L(n)} + \frac{n(d-1)^2}{L_1(n)L_2(n)} \right), \tag{5}
\]

where \( 1_{n > N_\rho} \) denotes the indicator function of the event \( \{ n > N_\rho \} \).

3. In case of a canonical link function, Theorem 3 improves these bounds to

\[
E \left[ \left|\hat{\beta}_n - \beta^{(0)}\right|^2 1_{n > N_\rho} \right] = O \left( \frac{\log(n)}{L_1(n)} \right). \tag{6}
\]

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).

An important intermediate result in proving these bounds is Proposition 2, where we derive

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

for any function \( L \) that satisfies \( \lambda_{\min} \left( \sum_{i=1}^n x_i x_i^T \right) \geq L(n) > 0 \) 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.

### 1.4 Applications

A useful application of Theorems 1 and 2 is the derivation of upper bounds of quadratic cost functions in \( \beta \). For example, let \( c(\beta) \) be a non-negative bounded function with \( ||c(\beta) - c(\beta^{(0)})|| \leq K||\beta - \beta^{(0)}||^2 \) for all \( \beta \in \mathbb{R}^d \) and some \( K > 0 \). Application of Theorems 1 and 2 yield the upper
bound

\[
E \left[ \left\| c(\hat{\beta}_n) - c(\beta(0)) \right\| \right] \leq E \left[ \left\| c(\hat{\beta}_n) - c(\beta(0)) \right\| 1_{n>N_p} \right] + E \left[ \left\| c(\hat{\beta}_n) - c(\beta(0)) \right\| 1_{n\leq N_p} \right]
\leq K \cdot E \left[ \left\| \hat{\beta}_n - \beta(0) \right\|^2 1_{n>N_p} \right] + \frac{E \left[ N_p \right]}{N_p} \max_{\beta} \left\| c(\beta) - c(\beta(0)) \right\|^2
= O \left( \frac{\log(n)}{L_1(n)} + n^{-\eta} \right).
\]

In dynamic pricing problems, such arguments are used to design decision rules and derive upper bounds on the regret, cf. den Boer and Zwart (2010). 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 parametric variants of bandit problems (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 \(N_p\) and upper bounds on \(E[\|\hat{\beta}_n - \beta(0)\|^2 1_{n>N_p}]\), 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 5, 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 \(\partial S = \bar{S} \setminus S\). For \(x \in \mathbb{R}_+ \{ x \}\) 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 \(\| A \| = \max_{\| z \| = 1} \| Az \|\). The 1-norm and \(\infty\)-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 \(f(x) \leq K g(x)\) 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 \(N_{\rho}\). 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 $\mathbb{R}^n$, $F: C \to \mathbb{R}^n$ 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 $C$.

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 $g(x) = \frac{\hat{h}(x)}{v(h(x))}$, 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

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

The existence of such a $\rho_0$ follows from the fact that $\hat{h}(x) > 0$ and $g(x) > 0$ for all $x \in \mathbb{R}$.

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

$$
A_n = \sum_{i=1}^n g(x_1^T \beta^{(0)}) x_1 e_i, \quad B_n = \sum_{i=1}^n g(x_i^T \beta^{(0)}) x_i x_i^T e_i, \quad J_n = c_1 \sum_{i=1}^n (|e_i| - E[|e_i| \mid \mathcal{F}_{i-1}]) x_i x_i^T.
$$

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

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

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 $\eta$-th moment of $N_\rho$ 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 $\lambda_{\min}(P_n) \geq L_1(n) \geq \alpha n^\lambda$, for some $c > 0$, $\frac{\alpha}{2} < \lambda \leq 1$ 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)

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

but Chang proves

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

(Here the terms $A_n$, $B_n$, $J_n$, and $\epsilon^*$ depend on $\beta$).

The following theorem shows asymptotic existence and strong consistency of $\hat{\beta}_n$, and provides mean square convergence rates.
Lemma 3 of \( \hat{\beta}_n \) yielding the following upper bound on the mean square convergence rates:

\[
E \left[ \left\| \hat{\beta}_n - \beta(0) \right\|^2 1_{n>N_p} \right] = O \left( \frac{\log(n)}{L_1(n)} + \frac{n(d-1)^2}{L_1(n)L_2(n)} \right).
\] (8)

Remark 2. If \( d = 1 \) then the term \( \frac{n(d-1)^2}{L_1(n)L_2(n)} \) in (8) 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 \( \frac{n(d-1)^2}{L_1(n)L_2(n)} = O \left( \frac{1}{L_1(n)} \right) \).

The bound in Theorem 2 then reduces to

\[
E \left[ \left\| \hat{\beta}_n - \beta(0) \right\|^2 1_{n>N_p} \right] = O \left( \frac{\log(n)}{L_1(n)} \right).
\] (9)

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 \( h = v \circ h \). The quasi-likelihood equations (2) then simplify to

\[
l_n(\beta) = \sum_{i=1}^{n} x_i (y_i - h(x_i^T \beta)) = 0. \tag{10}
\]

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 \( O \left( \frac{n(d-1)^2}{L_1(n)L_2(n)} \right) \) in (8) vanishes, yielding the following upper bound on the mean square convergence rates:

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

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 \( \inf_{x \in \mathbb{R}^d} \|x - x_0\| = \delta \|H(x) - H(x_0)\| \geq r \) 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 \( 0 < \rho \leq \rho_0 \), \( n \geq N_p \), \( \delta > 0 \) and \( r > 0 \). If \( \inf_{\beta \in \partial B_{\delta}} \|H_n(\beta) - H_n(\beta(0))\| \geq r \) and \( \|H_n(\beta(0))\| \leq r \) 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,

\[
E \left[ \left\| \hat{\beta}_n - \beta^{(0)} \right\|^2 1_{n \geq N_\rho} \right] = O \left( \frac{\log(n)}{L_1(n)} \right), \quad (0 < \rho \leq \rho_0).
\] (11)

Remark 5. Some choices of \( h \), e.g. \( h \) the identity or the logit function, have the property that

\[
\inf_{x \in X, \beta \in \mathbb{R}^N} h(x^T \beta) > 0, \quad \text{i.e.} \quad \epsilon_2 \text{ in equation (7) has a positive lower bound independent of } \rho_0.
\]

Since canonical link functions have \( \epsilon_1 = 0 \) in equation (7), 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 (4) and obtain

\[
E \left[ \left\| \hat{\beta}_n - \beta^{(0)} \right\|^2 \right] = O \left( \frac{\log(n)}{L_1(n)} \right), \quad (n \geq n_0).
\] (12)

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 (12) that \( E \left[ \left\| \hat{\beta}_n - \beta^{(0)} \right\|^2 \right] \) converges to zero as \( n \to \infty \).

### 4 Proofs

**Proof of Lemma 2**

A Taylor expansion of \( h \) and \( g \) yields

\[
y_i - h(x_i^T \beta) = y_i - h(x_i^T \beta^{(0)}) + h(x_i^T \beta^{(0)}) - h(x_i^T \beta) = e_i - \hat{h}(x_i^T \hat{\beta}^{(1)}_{i, \beta}) x_i^T \beta^{(0)} - h(x_i^T \beta^{(0)}) - h(x_i^T \beta),
\] (13)

\[
g(x_i^T \beta) = g(x_i^T \beta^{(0)}) + \tilde{g}(x_i^T \beta^{(0)}) x_i^T \beta^{(0)} + \frac{1}{2} (\beta - \beta^{(0)})^T \tilde{g}(x_i^T \beta^{(2)}_{i, \beta}) x_i x_i^T (\beta - \beta^{(0)}),
\] (14)

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

\[
(\beta - \beta^{(0)})^T l_n(\beta) = (\beta - \beta^{(0)})^T \sum_{i=1}^n g(x_i^T \beta) x_i (e_i - \hat{h}(x_i^T \hat{\beta}^{(1)}_{i, \beta}) x_i^T (\beta - \beta^{(0)}))
\]

\[
= (\beta - \beta^{(0)})^T \sum_{i=1}^n g(x_i^T \beta^{(0)}) x_i e_i
+ (\beta - \beta^{(0)})^T \sum_{i=1}^n \tilde{g}(x_i^T \beta^{(0)}) x_i^T (\beta - \beta^{(0)}) x_i e_i
+ (\beta - \beta^{(0)})^T \sum_{i=1}^n \left[ \frac{1}{2} (\beta - \beta^{(0)})^T \tilde{g}(x_i^T \beta^{(2)}_{i, \beta}) x_i x_i^T (\beta - \beta^{(0)}) \right] x_i e_i
- (\beta - \beta^{(0)})^T \sum_{i=1}^n g(x_i^T \beta) x_i \hat{h}(x_i^T \hat{\beta}^{(1)}_{i, \beta}) x_i^T (\beta - \beta^{(0)})
\]

\[
= (\beta - \beta^{(0)})^T A_n + (\beta - \beta^{(0)})^T B_n (\beta - \beta^{(0)}) + (I) - (II),
\]
where we write \((I) = (\beta - \beta(0))^T \sum_{i=1}^n \left[ \frac{1}{2} (\beta - \beta(0))^T \tilde{g}(x_i^T \tilde{\beta}_{i,\beta}^{(2)}) x_i + x_i^T (\beta - \beta(0)) \right] x_i e_i \) and \((II) = (\beta - \beta(0))^T \sum_{i=1}^n g(x_i^T \beta) x_i h(x_i^T \tilde{\beta}_{i,\beta}^{(1)}) x_i^T (\beta - \beta(0)) \). Since

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

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

\[
\leq c_1 (\beta - \beta(0))^T \sum_{i=1}^n \left| \beta - \beta(0) \right| x_i x_i^T (|e_i| - E [ |e_i| | \mathcal{F}_{i-1} ]) (\beta - \beta(0))
\]

\[
+ c_1 (\beta - \beta(0))^T \sum_{i=1}^n \left( \beta - \beta(0) \right) x_i x_i^T E [ |e_i| | \mathcal{F}_{i-1} ] (\beta - \beta(0))
\]

\[
\leq \left| \beta - \beta(0) \right| (\beta - \beta(0))^T J_n(\beta - \beta(0))
\]

\[
+ c_1 c_3 \left( \beta - \beta(0) \right) (\beta - \beta(0))^T \sum_{i=1}^n x_i x_i^T (\beta - \beta(0))
\]

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

\[
(\beta - \beta(0))^T J_n(\beta) \leq (\beta - \beta(0))^T A_n + (\beta - \beta(0))^T B_n(\beta - \beta(0))
\]

\[
+ \left| \beta - \beta(0) \right| (\beta - \beta(0))^T J_n(\beta - \beta(0)) - (c_2/2)(\beta - \beta(0))^T \sum_{i=1}^n x_i x_i^T (\beta - \beta(0)),
\]

using \((c_1 c_3 \left| \beta - \beta(0) \right| - c_2) \leq (c_1 c_3 \rho - c_2) \leq -c_2/2\).

**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 \( n > T \),

\[
0 \geq \sup_{\beta \in \partial B_{\rho}} S_n(\beta) - \rho^2 (c_2/2) L_1(n) \geq \sup_{\beta \in \partial B_{\rho}} S_n(\beta) - (c_2/2)(\beta - \beta(0))^T P_n(\beta - \beta(0))
\]

\[
\geq \sup_{\beta \in \partial B_{\rho}} (\beta - \beta(0))^T J_n(\beta),
\]

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 $A_n[i, i, j, J_n[i, j]$], then
\[
\sup_{\beta \in \partial B_n} S_n(\beta) \leq \rho \|A_n\| + \rho^2 \|B_n\| + \rho^3 \|J_n\|
\]
using the Cauchy-Schwartz inequality and the fact that $||x|| \leq ||x||_1 \leq \sum_{i,j} |A[i,j]|$ 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:
\[
T_{A[i]} = \sup\{ n \geq n_0 \mid \rho |A_n[i]| - \frac{1}{d + 2d^2} \rho^2 (c_2/2) L_1(n) > 0 \}, \quad (1 \leq i \leq d),
\]
\[
T_{B[i,j]} = \sup\{ n \geq n_0 \mid \rho^2 |B_n[i,j]| - \frac{1}{d + 2d^2} \rho^2 (c_2/2) L_1(n) > 0 \}, \quad (1 \leq i, j \leq d),
\]
\[
T_{J[i,j]} = \sup\{ n \geq n_0 \mid \rho^3 |J_n[i,j]| - \frac{1}{d + 2d^2} \rho^2 (c_2/2) L_1(n) > 0 \}, \quad (1 \leq i, j \leq d).
\]
By application of Proposition 1, Section 5, the last-times $T_{A[i]}$ and $T_{B[i,j]}$ are a.s. finite and have finite $\eta$-th moment, for all $\eta > 0$ such that $r > \frac{n+1}{\alpha} > 2$. Chow and Teicher (2003, page 95, Lemma 3) states that any two nonnegative random variables $X_1, X_2$ satisfy
\[
E\left[ (X_1 + X_2)^\eta \right] \leq 2^\eta (E\left[ X_1^\eta \right] + E\left[ X_2^\eta \right]),
\]
for all $\eta > 0$. Consequently
\[
\sup_{i \in \mathbb{N}} E\left[ ||e_i|| - E\left[ ||e_i|| \mid \mathcal{F}_{i-1} \right] ||^\eta \right] \leq \sup_{i \in \mathbb{N}} E\left[ ||e_i|| + E\left[ ||e_i|| \mid \mathcal{F}_{i-1} \right] ||^\eta \right]
\]
\[
\leq \sup_{i \in \mathbb{N}} 2^\eta (E\left[ ||e_i||^\eta \right] + E\left[ \left( E\left[ ||e_i|| \mid \mathcal{F}_{i-1} \right] \right)^\eta \right]) < \infty,
\]
and Proposition 1 implies that the last-times $T_{J[i,j]}$ are also a.s. finite and have finite $\eta$-th moment, for all $\eta > 0$ such that $r > \frac{n+1}{\alpha} > 2$. Now set $T = \sum_{1 \leq i \leq d} T_{A[i]} + \sum_{1 \leq i, j \leq d} T_{B[i,j]} + \sum_{1 \leq i, j \leq d} T_{J[i,j]}$. If $n > T$, then $\sup_{\beta \in \partial B_n} S_n(\beta) - \rho^2 (c_2/2) L_1(n) < 0$, and thus $T \leq T$ a.s. and $E\left[ T^\eta \right] \leq E\left[ T_0^\eta \right]$. $T$ 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 (15), for all $\eta > 0$ there is a constant $C_\eta$ such that
\[
E\left[ T^\eta \right] \leq C_\eta \left[ \sum_{1 \leq i \leq d} E\left[ T_{A[i]}^\eta \right] + \sum_{1 \leq i, j \leq d} E\left[ T_{B[i,j]}^\eta \right] + \sum_{1 \leq i, j \leq d} E\left[ T_{J[i,j]}^\eta \right] \right].
\]
It follows that $E\left[ T^\eta \right] < \infty$ for all $\eta > 0$ such that $r > \frac{n+1}{\alpha} > 2$. In particular, this implies $N_\rho < \infty$ a.s., and $E\left[ N_\rho^\eta \right] < \infty$.

**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 $0 < \rho \leq \rho_0$.

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

By contraposition of Corollary 1, if there is no solution $\beta \in B_n$ to $l_n(\beta) = 0$, then there exists a $\beta' \in \partial B_n$ 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,
\[(\beta' - \beta^{(0)})^T (c_2/2) P_n(\beta' - \beta^{(0)}) - (\beta' - \beta^{(0)})^T \begin{bmatrix} A_n + B_n(\beta' - \beta^{(0)}) + \|\beta' - \beta^{(0)}\| J_n(\beta' - \beta^{(0)}) \end{bmatrix} \leq 0,
\]
and, writing
\[(I) = \left\| (c_2/2)^{-1} P_n^{-1} \left[ A_n + B_n(\beta' - \beta^{(0)}) + \rho J_n(\beta' - \beta^{(0)}) \right] \right\|^2
\]
and
\[(II) = \frac{(d-1)^2 \left\| A_n + B_n(\beta' - \beta^{(0)}) + \rho J_n(\beta' - \beta^{(0)}) \right\|^2}{L_1(n)L_2(n)(c_2/2)^2},
\]
Lemma 7, Section 5, implies
\[\rho^2 = \left\| \beta' - \beta^{(0)} \right\|^2 \leq (I) + (II).\]

We now proceed to show
\[(I) + (II) < U_n,\]
for some $U_n$, independent of $\beta'$ and $\rho$, that satisfies
\[E[U_n] = O \left( \frac{\log(n)}{L_1(n)} + \frac{n(d-1)^2}{L_1(n)L_2(n)} \right).
\]
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 $\beta \in B_{U_1^{1/2}}$ to $l_n(\beta) = 0$, and thus $\|\beta_n - \beta^{(0)}\|^2 1_{n>N_p} \leq U_n$ a.s., and
\[E \left[ \|\beta_n - \beta^{(0)}\|^2 1_{n>N_p} \right] \leq E[U_n].\]

To prove (17), 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
\[\|MN\| = \max_{\|y\|=1} \|MNy\| = \max_{\|y\|=1} \left\| M \sum_{j=1}^d y[j] N(j) \right\| \leq \max_{\|y\|=1} \sum_{j=1}^d \|My[j]N(j)\| \leq \sum_{j=1}^d \|MN(j)\|.
\]
As a result we get
\[\left\| P_n^{-1} B_n(\beta' - \beta^{(0)}) \right\| \leq \left\| P_n^{-1} \sum_{i=1}^n \hat{g}(x_i^T \beta^{(0)}) x_i e_i x_i^T \right\| \|\beta' - \beta^{(0)}\| \leq \rho \sum_{j=1}^d \left\| P_n^{-1} \sum_{i=1}^n \hat{g}(x_i^T \beta^{(0)}) x_i e_i x_i[j] \right\|
\]
and
\[\left\| P_n^{-1} J_n(\beta' - \beta^{(0)}) \right\| \leq \left\| P_n^{-1} \sum_{i=1}^n c_1 x_i (|e_i| - E[|e_i| \mid F_{i-1}]) x_i^T \right\| \|\beta' - \beta^{(0)}\| \leq \rho \sum_{j=1}^d \left\| P_n^{-1} \sum_{i=1}^n c_1 x_i (|e_i| - E[|e_i| \mid F_{i-1}]) x_i[j] \right\|.
\]
In a similar vein we can derive

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

and

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

It follows that

\[ (I) \leq 2(c_2/2)^{-2} \left( \left\| P_n^{-1} A_n \right\|^2 + \left\| P_n^{-1} B_n(\beta' - \beta(0)) \right\|^2 + \rho_0^2 \left\| P_n^{-1} J_n(\beta' - \beta(0)) \right\|^2 \right) \]

\[ \leq U_n(1) + U_n(2) + U_n(3), \]

where we write

\[ U_n(1) = 2(c_2/2)^{-2} \left\| P_n^{-1} A_n \right\|^2, \]

\[ U_n(2) = 2(c_2/2)^{-2} \rho_0^2 \left( \sum_{j=1}^{d} \left\| P_n^{-1} \sum_{i=1}^{n} \hat{g}(x_i^T \beta(0)) x_i e_i[j] \right\|^2 \right), \]

\[ U_n(3) = 2(c_2/2)^{-2} \rho_0^2 \left( \sum_{j=1}^{d} \left\| \sum_{i=1}^{n} c_1 x_i (|e_i| - E [|e_i| \mid \mathcal{F}_{i-1}]) x_i[j] \right\|^2 \right), \]

and

\[ (II) \leq U_n(4) + U_n(5) + U_n(6), \]

where we write

\[ U_n(4) = \frac{2(d-1)^2 \left\| A_n \right\|^2}{L_1(n)L_2(n)(c_2/2)^2}, \]

\[ U_n(5) = \frac{2(d-1)^2}{L_1(n)L_2(n)(c_2/2)^2} \left\| \rho_0 \sum_{j=1}^{d} \left( \sum_{i=1}^{n} \hat{g}(x_i^T \beta(0)) x_i e_i[j] \right) \right\|^2, \]

\[ U_n(6) = \frac{2(d-1)^2}{L_1(n)L_2(n)(c_2/2)^2} \rho_0^2 \left( \sum_{j=1}^{d} \left\| \sum_{i=1}^{n} c_1 x_i (|e_i| - E [|e_i| \mid \mathcal{F}_{i-1}]) x_i[j] \right\| \right)^2. \]

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 5 on the martingale difference sequences \( \{g(x_i^T \beta(0))e_i\}_{i \in \mathbb{N}} \), \( \{\hat{g}(x_i^T \beta(0))x_i[j]e_i\}_{i \in \mathbb{N}} \), and \( \{c_1|x_i|-E[|e_i| \mid \mathcal{F}_{i-1}]x_i[j]\}_{i \in \mathbb{N}} \), 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

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

implies the existence of a constant \( K_2 > 0 \) such that \( E[U_n(4) + U_n(5) + U_n(6)] \leq \frac{K_2 n(d-1)^2}{L_1(n)L_2(n)}. \)
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 $\beta, \beta'$. Since $n \geq n_0$ this implies $l_n(\beta) = l_n(\beta')$. By a first order Taylor expansion, there are $\beta_i, 1 \leq i \leq n$, on the line segment between $\beta$ and $\beta'$ such that

$$l_n(\beta) - l_n(\beta') = \sum_{i=1}^n x_i (x_i^T \hat{h}(x_i^T \beta_i) - x_i^T \hat{h}(x_i^T \beta_i')).$$

Since $\inf_{x \in X, \beta \in \mathbb{B}_\rho} \hat{h}(x^T \beta) > 0$, Lemma 8 in Section 5 implies that the matrix $\sum_{i=1}^n x_i x_i^T \hat{h}(x_i^T \beta_i)$ 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

$$l_n(\beta) - l_n(\beta^{(0)}) = \sum_{i=1}^n x_i (h(x_i^T \beta^{(0)}) - h(x_i^T \beta)) = \sum_{i=1}^n x_i x_i^T \hat{h}(x_i^T \beta) - \beta),$$

for some $\beta_i, 1 \leq i \leq n$, on the line segment between $\beta^{(0)}$ and $\beta$.

Write $T_n(\beta) = \sum_{i=1}^n x_i x_i^T \hat{h}(x_i^T \beta)$, and choose $k_2 > \left( \inf_{\beta \in \mathbb{B}_\rho, x \in X} \hat{h}(x^T \beta) \right)^{-1}$. Then for all $\beta \in \mathbb{B}_\rho$

$$\lambda_{\text{min}} (k_2 T_n(\beta) - P_n) = \lambda_{\text{min}} \left( \sum_{i=1}^n x_i x_i^T (k_2 \hat{h}(x_i^T \beta) - 1) \right) \geq \left( \inf_{\beta \in \mathbb{B}_\rho, x \in X} (k_2 \hat{h}(x^T \beta) - 1) \right) \lambda_{\text{min}} (P_n),$$

by Lemma 8. This implies

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


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

$$\left\| H_n(\beta) - H_n(\beta^{(0)}) \right\|^2 = \left\| P_n^{-1/2}(l_n(\beta) - l_n(\beta^{(0)})) \right\|^2 = (\beta^{(0)} - \beta)^T T_n(\beta) P_n^{-1} T_n(\beta) (\beta^{(0)} - \beta) \geq (\beta^{(0)} - \beta)^T T_n(\beta) k_2^2 T_n(\beta) (\beta^{(0)} - \beta) \geq (\beta^{(0)} - \beta)^T P_n k_2^2 (\beta^{(0)} - \beta) \geq k_2^2 \left\| \beta^{(0)} - \beta \right\|^2 \lambda_{\text{min}} (P_n) \geq k_2^2 \delta_n^2 L_1(n),$$

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

Now $E \left[ \|H_n(\beta^{(0)})\|^2 \right] = E \left[ \sum_{i=1}^n x_i x_i^T P_n^{-1} (\sum_{i=1}^n x_i x_i^T) \right] = E[Q_n]$, 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 $E \left[ ||\beta - \beta^{(0)}||^2 1_{n \geq N_\rho} \right] = O \left( \frac{\log(n)}{L_1(\sigma^2)} \right)$.

5 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 $\eta$-th moment of the last-time $N_\rho$ 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 $\epsilon > 0$,

$$P \left( \max_{1 \leq k \leq n} |S_k| \geq \epsilon \right) \leq 2P \left( |S_n| \geq \epsilon - \sqrt{2\sigma^2n} \right).$$

**(18)**

**Proof.** We use similar techniques as de la Peña et al. (2009, Theorem 2.21, p.16), where (18) is proven for independent random variables $(X_i)_{i \in \mathbb{N}}$. Define the events $A_1 = \{S_1 \geq \epsilon\}$ and $A_k = \{S_k \geq \epsilon, S_1 < \epsilon, \ldots, S_{k-1} < \epsilon\}$, $2 \leq k \leq n$. Then $A_k(1 \leq k \leq n)$ are mutually disjoint, and $
\{\max_{1 \leq k \leq n} S_k \geq \epsilon\} = \bigcup_{k=1}^n A_k$.

$$P \left( \max_{1 \leq k \leq n} S_k \geq \epsilon \right) \leq P \left( S_n \geq \epsilon - \sqrt{2\sigma^2n} \right) + P \left( \max_{1 \leq k \leq n} S_k \geq \epsilon, S_n < \epsilon - \sqrt{2\sigma^2n} \right)$$

$$\leq P \left( S_n \geq \epsilon - \sqrt{2\sigma^2n} \right) + \sum_{k=1}^n P \left( A_k, S_n < \epsilon - \sqrt{2\sigma^2n} \right)$$

$$\leq P \left( S_n \geq \epsilon - \sqrt{2\sigma^2n} \right) + \sum_{k=1}^n P \left( A_k, S_n - S_k < -\sqrt{2\sigma^2n} \right)$$

$$\leq P \left( S_n \geq \epsilon - \sqrt{2\sigma^2n} \right) + \sum_{k=1}^n P \left( A_k, S_n - S_k < -\sqrt{2\sigma^2n} \right)$$

$$\leq P \left( S_n \geq \epsilon - \sqrt{2\sigma^2n} \right) + \sum_{k=1}^n 1/2 P \left( A_k \right)$$

$$= P \left( S_n \geq \epsilon - \sqrt{2\sigma^2n} \right) + P \left( \max_{1 \leq k \leq n} S_k \geq \epsilon \right),$$

where (1) uses $A_k \in \mathcal{F}_k$, and (2) uses $E[1_{S_n - S_k < -\sqrt{2\sigma^2n}} \mid \mathcal{F}_k] = P(S_k - S_n > \sqrt{2\sigma^2n} \mid \mathcal{F}_k) \leq E[(S_n - S_k)^2 \mid \mathcal{F}_k]/(2\sigma^2n) \leq 1/2$ a.s. This proves $P(\max_{1 \leq k \leq n} S_k \geq \epsilon) \leq 2P(S_n \geq \epsilon - \sqrt{2\sigma^2n})$. 

---

14
Replacing \( S_k \) by \(-S_k\) gives \( P(\max_{1 \leq k \leq n} |S_k| > \epsilon) \leq 2P(-S_n \geq \epsilon - \sqrt{2\sigma^2 n}) \). If \( \epsilon - \sqrt{2\sigma^2 n} \leq 0 \) then (18) is trivial; if \( \epsilon > \sqrt{2\sigma^2 n} \) then

\[
P\left( \max_{1 \leq k \leq n} |S_k| \geq \epsilon \right) \leq P\left( \max_{1 \leq k \leq n} S_k \geq \epsilon \right) + P\left( \max_{1 \leq k \leq n} -S_k \geq \epsilon \right) \leq 2P\left( S_n \geq \epsilon - \sqrt{2\sigma^2 n} \right) + 2P\left( -S_n \geq \epsilon - \sqrt{2\sigma^2 n} \right) = 2P\left( |S_n| \geq \epsilon - \sqrt{2\sigma^2 n} \right).
\]

\( \square \)

**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 \( \sup_{i \in \mathbb{N}} E[X_i^2 | \mathcal{F}_{i-1}] \leq \sigma^2 < \infty \) a.s. for some \( \sigma > 0 \). Let \( c > 0 \), \( \frac{1}{2} < \alpha \leq 1 \), \( \eta > 2\alpha - 1 \), \( r > \frac{\alpha + 1}{\alpha} \). If \( \sup_{i \in \mathbb{N}} E |X_i|^r < \infty \), then

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

**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 \( \sup_{i \in \mathbb{N}} E[X_i^2 | \mathcal{F}_{i-1}] \leq \sigma^2 < \infty \) a.s. for some \( \sigma > 0 \). Let \( c > 0 \), \( \frac{1}{2} < \alpha \leq 1 \), \( \eta > 2\alpha - 1 \), \( r > \frac{\alpha + 1}{\alpha} \), 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

\[
T < \infty \text{ a.s., and } E[T^n] < \infty.
\]

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

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

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

\[
\sum_{j \geq \lfloor \log_2(n) \rfloor} P\left( |S_{2^j}| \geq c(2^{j-1})^\alpha / 2 \right) = \int_{j \geq \lfloor \log_2(n) \rfloor} P\left( |S_{2^j}| \geq c(2^{j-1})^\alpha / 2 \right) dj
= \int_{k \geq n} P\left( |S_k| \geq c(k/2)^\alpha / 2 \right) \frac{1}{k \log(2)} dk = \sum_{k \geq n} P\left( |S_k| \geq c(k/2)^\alpha / 2 \right) \frac{1}{k \log(2)}.
\]

using a variable substitution \( k = 2^j \).
By Chebyshev’s inequality,

\[ P(T > n) \leq 2 \sum_{k \geq n} P(|S_k| \geq c(k/2)^\alpha/2) \frac{1}{k \log(2)} \leq 2 \sum_{k \geq n} \sigma^2 k(c(k/2)^\alpha/2)^{-2} \frac{1}{k \log(2)}, \]

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

Since

\[ E[T^n] \leq \eta \left[ 1 + \sum_{n \geq 1} n^{n-1} P(T > n) \right] \]

\[ \leq \eta \left[ 1 + n' \cdot (n')^{n-1} + \sum_{n > n'} n^{n-1} P(T > n) \right] \]

\[ \leq M \sum_{n > n'} n^{n-1} \sum_{j \geq \lceil \log_2(n) \rceil} P(|S_{2j}| \geq c(2^{-j})^{\alpha/2}), \]

for some constant \( M > 0 \), it follows by (19), (20) that \( E[T^n] < \infty \) if

\[ \sum_{n \geq 1} n^{n-1} \sum_{k \geq n} P(|S_k| \geq c(k/2)^\alpha/2) k^{-1} < \infty. \]

By interchanging the sums, it sufﬁces to show

\[ \sum_{k \geq 1} k^{n-1} P(|S_k| \geq 2^{-1-\alpha}ck^\alpha) < \infty. \]

This last statement follows from Lemma 5. \( \square \)

Let \((\epsilon_i)_{i \in \mathbb{N}}\) be a martingale difference sequence w.r.t. a ﬁltration \( \{\mathcal{F}_i\}_{i \in \mathbb{N}} \), such that \( \sup_{i \in \mathbb{N}} E[\epsilon_i^2 \mid \mathcal{F}_{i-1}] = \sigma^2 < \infty \) 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 \( \sup_{i \in \mathbb{N}} \|x_i\|_1 \leq M < \infty \) for some (non-random) \( M > 0 \). Write \( \Pi_n = \sum_{i=1}^n x_i x_i^T \). Let \( L : \mathbb{N} \to \mathbb{R}_+ \) be a (non-random) function and \( n_0 \geq 2 \) a (non-random) integer such that \( \lambda_{\min}(\Pi_n) \geq L(n) \) 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 \( n \geq n_0 \),

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

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 \text{ for all } y \geq y_n-1. \text{ Together with } R_{n-1} > 1 \text{ 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, \text{ 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 \( q_n = \sum_{i=1}^{\infty} x_i e_i \) and \( Q_n = q_n P_n^{-1} q_n. \) For \( n \geq n_0, \) \( P_n \) is invertible, and
\[ ||P_n^{-1} q_n||^2 \leq \left( \sum_{i=1}^{\infty} \frac{1}{\sigma_i^2} \right) \cdot \left( \sum_{i=1}^{\infty} \frac{1}{\sigma_i^2} \right) \leq \lambda_{\min}(P_n)^{-1} q_n P_n^{-1} q_n \leq L(n)^{-1} Q_n \text{ a.s.,} \]
where we used \( \left( \sum_{i=1}^{\infty} \frac{1}{\sigma_i^2} \right) = \lambda_{\max}(P_n)^{-1/2} = \lambda_{\min}(P_n)^{-1/2}. \) 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 \( V_n = V_{n-1} - \frac{x_n x_n^T V_{n-1}}{1 + x_n^T V_{n-1} x_n}, \) 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
\[ Q_n = \left( \sum_{i=1}^{\infty} x_i^T e_i \right) V_n \left( \sum_{i=1}^{\infty} x_i e_i \right) = \left( \sum_{i=1}^{\infty} x_i^T e_i \right) V_n \left( \sum_{i=1}^{\infty} x_i e_i \right) + \frac{x_n^T V_n x_n e_n^2}{1 + x_n^T V_n x_n} + 2x_n^T V_n \left( \sum_{i=1}^{\infty} x_i e_i \right) e_n \]
\[ = Q_{n-1} + \left( \sum_{i=1}^{\infty} x_i^T e_i \right) \left( \frac{V_{n-1} x_n x_n^T V_{n-1}}{1 + x_n^T V_{n-1} x_n} \right) \left( \sum_{i=1}^{\infty} x_i e_i \right) \]
\[ + \frac{x_n^T V_n x_n e_n^2}{1 + x_n^T V_n x_n} + \frac{x_n^T V_{n-1} \left( \sum_{i=1}^{\infty} x_i e_i \right)}{1 + x_n^T V_{n-1} x_n} \left( \sum_{i=1}^{\infty} x_i e_i \right) e_n \]
\[ = Q_{n-1} - \frac{x_n^T V_{n-1} \left( \sum_{i=1}^{\infty} x_i e_i \right)}{1 + x_n^T V_{n-1} x_n} + x_n^T V_n x_n e_n^2 + 2x_n^T V_n \left( \sum_{i=1}^{\infty} x_i e_i \right) e_n. \]
Observe that
\[ E \left[ x_n^T V_{n-1} \left( \sum_{i=1}^{\infty} x_i e_i \right) e_n \right] = E \left[ x_n^T V_{n-1} \left( \sum_{i=1}^{\infty} x_i e_i \right) E[e_n | \mathcal{F}_{n-1}] \right] = 0 \]
and
\[ E \left[ x_n^T V_n x_n e_n^2 \right] = E \left[ x_n^T V_n x_n E[e_n^2 | \mathcal{F}_{n-1}] \right] \leq E \left[ x_n^T V_n x_n \right] \sigma^2. \]
By telescoping the sum we obtain
\[ E[Q_n] \leq E[Q_{\min\{n,n_1\}}] + \sigma^2 \sum_{i=n_1+1}^{n} E[x_i^T V_i x_i], \]

17
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

$$
\det(P_{n-1}) = \det(P_n - x_n x_n^T) = \det(P_n) \det(I - P_n^{-1} x_n x_n^T) = \det(P_n)(1 - x_n^T V_n x_n), \quad (n \geq n_1).
$$

(21)

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 $x_n^T V_n x_n = \frac{\det(P_n)}{\det(P_{n-1})}$. Define the sequence $(y_n)_{n \in \mathbb{N}}$ by $y_n = \det(P_{n+1})$. Then $(y_n)_{n \in \mathbb{N}}$ is a nondecreasing sequence with $y_1 \geq \det(P_{n+1}) \geq \lambda_{\min}(P_{n+1})^d \geq e$. Lemma 6 implies

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

Now $\log(\det(P_n)) \leq d \log(\lambda_{\max}(P_n)) \leq d \log(\text{tr}(P_n)) \leq d \log(n \sup_{i \in \mathbb{N}} ||x_i||^2) \leq d \log(n M^2)$. Furthermore, for all $n_0 \leq n \leq n_1$ we have

$$
E[Q_n] \leq E \left[ ||q_n||^2 \lambda_{\max}(P_n^{-1}) \right] \leq E \left[ \left| \sum_{i=1}^{n} x_i x_i^T L(n_0)^{-1} \right| \right] \leq L(n_0)^{-1} \left( 2 \sum_{i=1}^{n-n_1} \sup_{i \in \mathbb{N}} ||x_i||^2 \right)
$$

$$
\leq 2L(n_0)^{-1}M^2 n_1 \sigma^2,
$$

and thus for all $n \geq n_0$,

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

$$
\leq 2L(n_0)^{-1}M^2 n_1 \sigma^2 + d \log(n) + d \log(M^2)
$$

$$
\leq K \log(n),
$$

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

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

**Proof.** Let $0 < \lambda_1 \leq \ldots \leq \lambda_d$ 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 $\sum_{i=1}^{d} \alpha_i v_i$, for coordinates $(\alpha_1, \ldots, \alpha_d)$, and $b = \sum_{i=1}^{d} \beta_i v_i$ 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 that $S$ equals $\{ x \in \mathbb{R}^d \mid x^T A x + x^T b \leq 0 \}$.

Fix $\alpha = (\alpha_1, \ldots, \alpha_d) \in S$ and write $R = \{ i \mid \alpha_i (\lambda_i \alpha_i + \beta_i) \leq 0, 1 \leq i \leq d \}$, $R^c =\{1, \ldots, d\} \setminus R$. For all $i \in R$, standard properties of quadratic equations imply $\alpha_i^2 \leq \lambda_1^{-2} \beta_i^2$ and $\alpha_i (\lambda_i \alpha_i + \beta_i) \geq \frac{-\beta_i^2}{\lambda_1}$. 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 \( c = \sum_{i \in R} \frac{\beta_i^2}{2 \lambda_i}. \) By the quadratic formula, \( \alpha_i (\lambda_i \alpha_i + \beta_i) - c \leq 0 \) implies

\[
-\beta_i - \sqrt{\beta_i^2 + 4 \lambda_i c} \leq \alpha_i \leq -\beta_i + \sqrt{\beta_i^2 + 4 \lambda_i c}.
\]

(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

\[
||x||^2 = \sum_{i=1}^d \alpha_i^2 \leq \lambda_i^{-2} \beta_i^2 + \sum_{i \in R^c} \left( \frac{\beta_i^2}{\lambda_i^2} + \frac{1}{2} \sum_{j \in R^c} \frac{1}{\lambda_j} \sum_{j \in R^c} \frac{1}{\lambda_j} \right)
\leq \sum_{i=1}^d \lambda_i^{-2} \beta_i^2 + \frac{1}{2} \left( \sum_{i \in R^c} \frac{1}{\lambda_i} \right) \left( \sum_{j \in R^c} \frac{1}{\lambda_j} \right) \left( \sum_{i=1}^n \beta_i^2 \right)
\leq ||A^{-1}b||^2 + (d-1)^2 \frac{1}{\lambda_1 \lambda_2} ||b||^2,
\]

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

**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 \( \beta_1 = \beta_2 > 0, \alpha_1 = -\beta_1/(2 \lambda_1), \alpha_2 = (-\beta_2 - \sqrt{\beta_2^2 + 4 \lambda_2 \beta_1^2/(4 \lambda_1)})/(2 \lambda_2), \) 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 \( \lambda_1, \lambda_2. \) Then \( x^T A x + x^T b = \sum_{i=1}^d \alpha_i (\lambda_i \alpha_i + \beta_i) = 0 \) and

\[
||x||^2 = \alpha_1^2 + \alpha_2^2 \geq \beta_1^2/(4 \lambda_1^2) + \beta_2^2/(4 \lambda_2^2) + \beta_1^2/(4 \lambda_1 \lambda_2) \geq \frac{1}{8} ||A^{-1}b||^2 + ||b||^2/(8 \lambda_1 \lambda_2).
\]

**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 $\lambda_{\min} \left( \sum_{i=1}^{n} x_{i} x_{i}^{T} w_{i} \right)$. Then
\[
\lambda_{\min} \left( \sum_{i=1}^{n} x_{i} x_{i}^{T} \right) = \min_{||v||=1} v^{T} \left( \sum_{i=1}^{n} x_{i} x_{i}^{T} \right) v \leq \tilde{v}^{T} \left( \sum_{i=1}^{n} x_{i} x_{i}^{T} \right) \tilde{v} \leq \tilde{v}^{T} \left( \sum_{i=1}^{n} x_{i} x_{i}^{T} w_{i} \right) \tilde{v} (\inf_{i \in \mathbb{N}} w_{i})^{-1}
\]

\[
= \lambda_{\min} \left( \sum_{i=1}^{n} x_{i} x_{i}^{T} w_{i} \right) (\inf_{i \in \mathbb{N}} w_{i})^{-1}.
\]

\[
\square
\]

References


