CWI

Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

K. Dzhaparidze

On iterative estimators

Department of Operations Research, Statistics, and System Theory Report BS-R9036 December

The Centre for Mathematics and Computer Science is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Research (N.W.O.).

On Iterative Estimators

Kacha Dzhaparidze

Centre for Mathematics and Computer Science
P.O. Box 4079 1009 AB Amsterdam, The Netherlands

Optimal recommendations for drawing statistical inference about unknown parameters are usually based on optimization of some criterion function (e.g. likelihood, least squares, etc.). This optimization is then carried out over all admissible values of parameters, so an unconstrained global optimum is sought. Typically the criterion function is taken asymptotically quadratic, as sample size increases, of the parameter value. We apply to these optimization problems methods developed in numerical analysis (*quasi-Newton* or *conjugate gradient* methods, or rather their appropriate stochastic modifications) which possess the so-called *quadratic termination* property: the minimum of a quadratic function is found in at most d iterations where d is the number of unknowns. As applied to an asymptotically quadratic criterion function, such methods lead to asymptotically efficient estimators for a d-dimentional parameter in at most d iterates.

1980 mathematics Subject Classification: 62F12 65U05

Keywords and phrases: Newton, quasi-Newton and conjugate gradient methods, parameter estimation.

Note: This paper is based on lecture notes presented at the seminar on *Iterative Procedures in Statistics,* organized by G. Celeux and C. Callant, INRIA Rocquencourt, project CLOREC, April 3, 1990.

1. Introduction. Applications to Likelihood Theory

1.1. The application of the basic Newton (- Raphson) method to estimating an unknown parameter involved in the density of an i.i.d. sample has been well-known since Fisher (1925), who has noticed that as applied to maximizing the likelihood function this method, or rather its stochastic modification called the *scoring* method, improves efficiently any rough initial estimator after the very first iterate. As the log-likelihood function is assumed asymptotically quadratic (when the sample size increases), the reason for this effect can be sought in the *quadratic termination* property of Newton's method, which localizes the maximum of a quadratic function in a single iterate. This observation is of immense practical importance, since direct methods for finding maximum likelihood estimators are feasible only in very special cases, and one usually localizes the maximum by applying one of the iterative methods advanced in numerical analysis, which are in fact developments of the classical Newton method. Most important are the so-called *quasi-Newton* and

Report BS-R9036
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

conjugate gradient methods. In order to retain the quadratic termination property, these methods are kept as close to Newton's iterates as possible, only introducing modifications to gain more reliability. As applied to a quadratic objective function, they localize the maximum in fewer then d iterares where d is a number of unknown parameters. In our early papers (see Beinicke and Dzhaparidze (1982), Dzhaparidze and Yaglom (1983) and Dzhaparidze (1983); cf. Paardekooper et. al. (1989)) we have provided for suitable stochastic modifications of the quasi-Newton and conjugate gradient methods which perform on an asymptotically quadratic log-likelihood function like Newton's method improving efficiently any preliminary estimator in fewer then d iterates. In the present paper special attention is paid to the detailed proofs which utilize ideas borrowed from numerical analysis; see section 2 devoted to a short review of iterative methods for unconstrained optimization, much in the spirit of Brodley (1977), Ortega and Rheinboldt (1970) and Scales (1985) where further details can be found.

In the remainder of this section we present a short account of the well-known ideas of asymptotic estimation theory which lie at the basis of our results in section 3; for more details see e.g. LeCam (1960, 1969), Ibragimov and Has'minskii (1981) and Basawa and Scott (1983). We begin in section 1.2 with discussing the classical case of i.i.d. observations, which is an important particular case of the general scheme of LAN experiments. In section 1.3 we present the implication of our results in section 3 as applied to the last scheme; see statement 1.3.1. However, the general setting in section 3 extends beyond the LAN scheme and embraces important situations indicated in sections 1.4 and 1.5.

1.2. In case treated by Fisher of an i.i.d. sample $X_1, ..., X_n$ drawn from a density f_θ which depends on an unknown d-vector valued parameter θ, the logarithm of the likelihood function $f_\theta (X_1)...f_\theta (X_n)$ is usually locally (at each fixed value of a parameter $\theta \in \Theta$ where Θ is an open set in R^d) *asymptotically quadratic* in the sense that (i) it may be developed with the differential $\delta_n = n^{-1/2}$ up to the second order term in the Teylor series, in any direction u for which $\theta + \delta_n u$ is again a parametric value, i.e. $u \in \mathcal{U}_n = \delta_n^{-1} (\Theta - \theta)$, and (ii) a remainder term tends to zero as $n \to \infty$ for each $\theta \in \Theta$ and $u \in \mathcal{U}_n$ in probability $P_{n,\theta}$, where $P_{n,\theta}$ is the distribution of a sample. (Observe that in the course of developing asymptotic theory we will repeatedly come across various remainder terms of this kind which always will be denoted by the same symbol $\eta_n (\theta, u)$, as their actual form is insignificant in this context; cf. e.g. (1.2.1) below).

Furthermore, one can apply here the classical central limit theorem and law of large numbers to the first and second terms in this expansion (since the corresponding gradient vector and Hessian matrix consist of sums of n i.i.d. variables normed by $n^{-1/2}$ and n^{-1} respectively) in order to establish the following so-called *local asymptotic normal* (LAN) property of the likelihood ratio $Z_{n,\theta} (u) = dP_{n,\theta+\delta_n u} / dP_{n,\theta}$:

$$\log Z_{n,\theta} (u) = (u, g_n (\theta)) - (I (\theta) u, u) / 2 + \eta_n (\theta, u) \tag{1.2.1}$$

(here and elsewhere below (., .) means the inner product of elements in R^d) where $\{g_n(\theta) = g_n(X_n, \theta)\}_{n=1,2,...}$ with $X_n = (X_1, ..., X_n)$ is a sequence of asymptotically normal random R^d- valued vectors with zero mean as $n \to \infty$ and positive definite for each $\theta \in \Theta$ covariance matrix I (θ) called *Fisher's information matrix.* We may express the last fact as follows:

(1.2.2) $$\mathcal{L}\{g_n(\theta) \mid P_{n,\theta}\} \Rightarrow N\{0, I(\theta)\}$$

with the distribution law of $g_n(\theta)$ under $P_{n,\theta}$ on the left hand side and the limit Gaussian distribution on the right hand side. Notice that Fisher's information matrix I (θ) appears above in two different places, in (1.2.1) and (1.2.2), since in the sense of the convergence in $P_{n,\theta}$ probability we have for any non-zero $x \in R^d$ that

(1.2.3) $$- ((\partial^2/\partial u^2)\, x, x) \log Z_{n,\theta}(u)\big|_{u=0} \to E_{n,\theta}\{((\partial/\partial u), x) \log Z_{n,\theta}(u)\big|_{u=0}\}^2$$
$$= (I(\theta)\, x, x)$$

and, by the obvious relationship between the gradient vector and Hessian matrix,

(1.2.4) $$(g_n(\theta + \delta_n u) - g_n(\theta), x) \to - (I(\theta)\, u, x).$$

The LAN property (1.2.1)-(1.2.2) plays a key role on treating, for instance, the asymptotic behaviour of the maximum likelihood estimator for θ, which by definition is the rooth of the likelihood equation $g_n(\theta) = 0$ (see, e.g. Ibragimov and Has'minskii (1981)). In particular, one can argue as follows. Suppose that the relationship (1.2.4) admits substituting u by any δ_n-consistent estimator T_n for θ (see Definition 3.2.3 below, where the class of such estimators is denoted by $\mathbb{D}$), that is, the following relationship holds: as n $\to \infty$

(1.2.5) $$\mid g_n(T_n) - g_n(\theta) + I(\theta)\, \delta_n^{-1}(T_n - \theta) \mid \to 0 \quad T_n \in \mathbb{D}$$

in $P_{n,\theta}$ probability. If now a subclass D of $\mathbb{D}$ is singled out consisting of all δ_n-consistent estimators T_n such that $\mid g_n(T_n) \mid \to 0$ as $n \to \infty$ in $P_{n,\theta}$ probability (cf. section 3.4 below), then (1.2.5) yields

(1.2.6) $$\mid g_n(\theta) - I(\theta)\, \delta_n^{-1}(T_n - \theta) \mid \to 0 \quad T_n \in D$$

as $n \to \infty$ in $P_{n,\theta}$ probability. Therefore by (1.2.2) we have

(1.2.7) $$\mathcal{L}\{\delta_n^{-1}(T_n - \theta) \mid P_{n,\theta}\} \Rightarrow N\{0, I^{-1}(\theta)\} \quad T_n \in D.$$

Thus the subclass D, which clearly includes the maximum likelihood estimator provided it is δ_n-consistent, consists of asymptotically efficient estimators in Fisher's sense (see, e.g. Ibragimov and Has'minskii (1981)). The main objective of this paper is to provide for iterative procedures of estimating θ which are well defined (in the sense that at each iterate a δ_n-consistent estimator is constructed; cf. definition 3.4.2) and terminated in fewer then d iterates at an asymptotically efficient estimator belonging to D; see statement 1.3.1 below. Specifically, the basic procedure of estimating the parameter θ, Fisher's *scoring* method, consists of the following stochastic modification of Newton's iterates

(1.2.8) $$\theta_{k+1n} = \theta_{kn} + \delta_n I^{-1}(\theta_{kn})\, g_n(\theta_{kn})$$

with any δ_n-consistent initial estimator θ_{0n} for θ and a consistent estimator I (θ_{kn}) for the information matrix I (θ) which is supposed to be continuous in θ (cf. conditions (i) and (ii) in section 3.4 and remark 3.4.1). As was pointed out by Fisher (1925), the result θ_{1n} of the very first iterate (1.2.8) is asymptotically efficient in the above sense: indeed, this is easily seen in section 3.6.

1.3. The first rigorous treatment of fine asymptotic properties of θ_{1n} in (1.2.8) as an estimator of θ is due to LeCam (1956). Later LeCam (1960, 1969) extended his studies to a more general class of experiments then those generated by i.i.d. observations, as in the previous section. It was assumed that an experiment $\mathcal{E}_n = [\mathcal{X}_n, \mathcal{A}_n, \{P_{n,\theta}\}_{\theta\in\Theta}]$ is n-th in a sequence of experiments, where $\mathcal{X}_n$ is the set of possible outcomes of the n-th experiment, $\mathcal{A}_n$ a σ-algebra defined on $\mathcal{X}_n$, and $\{P_{n,\theta}\}_{\theta\in\Theta}$ a parametric family of distributions on $\mathcal{A}_n$. We observe $X_n \in \mathcal{X}_n$ drawn according to some unknown $\theta \in \Theta$ and wish to make inference concerning θ. As in the previous section, assume the LAN property of the likelihood ratio $Z_{n,\theta}$ (u) (defined as above, with the usual convention $Z_{n,\theta}$ (u) = 0 or ∞ when $P_{n,\theta+\delta_n u}$ is singular with respect to $P_{n,\theta}$ or $P_{n,\theta}$ is singular with respect to $P_{n,\theta+\delta_n u}$) which satisfies (1.2.1) with an asymptotically normal g_n (θ) = g_n (X_n, θ), $X_n \in \mathcal{X}_n$ as in (1.2.2), and a positive definite information matrix I (θ), strictly bounded from below and above (cf. condition II in section 3.3)

As is shown in LeCam (1969), p.68, the conditions stipulated above guarantee relationship (1.2.4), provided g_n (θ) is suitably chosen among asymptotically equivalent candidates. Moreover, with a sufficiently smooth choice of g_n (θ) as a function of θ we have also (1.2.5) for any δ_n-consistent estimator T_n. Otherwise T_n always can be modified as to take on a finite number of possible values, and then (1.2.4) implies (1.2.5) for such a modification; see LeCam (1969), p.81, also LeCam (1974), p.155 for a construction method. We do not enter here in details, assuming relationship (1.2.5) for any δ_n-consistent estimator T_n.

Observe that on deviating from the i.i.d. case one often encounters situations in which components of T_n converge with a rate different from $\delta_n = n^{-1/2}$, they even may have various rates. Therefore $R^d \times R^d$-matrix valued differentials δ_n will be allowed, positive definite and such that $\| \delta_n \| \to 0$ as $n \to \infty$, where $\| \delta_n \| = \sup \{(\delta_n u, u): | u | = 1\}$ is a norm of δ_n. Let θ_{0n} be a δ_n-consistent estimator for θ and G_{0n} a consistent positive definite estimator for I (θ), for instance I (θ_{kn}) when I (θ) is continuous in θ; cf. conditions (i) and (ii) in section 3.4 and Remark 3.4.1. It is easily verified that our results in section 3 imply

Statement 1.3.1. *Under the conditions stipulated in the present section the estimator* θ_{1n} *for* θ *constructed according to the first of Newton's iterates* (1.2.8) *is asymptotically efficient in the sense that it satisfies* (1.2.6) *and* (1.2.7). *Let* θ_{kn} *be the estimator for* θ *constructed as the* k*-th iterate by one of the modified conjugate gradient or quasi-Newton* *methods defined in section* 3.8 *and* 3.9. *Then* θ_{rn} *for some* $r \leq d$ *is asymptotically efficient in the same sense.*

1.4. In Basawa and Pracasa Rao (1980) and Basawa and Scott (1983) one can find plenty of situations in which the convergence in (1.2.3) is violated. In order to include these, so-called *non-ergodic* models into consideration, we assume in section 3 that (1.2.1) and (1.2.5) hold with I (θ) substituted by some $\mathcal{a}_n$-measurable (for each fixed $\theta \in \Theta$) function G_n: $X_n \times \Theta \to R^d \times R^d$ satisfying condition II in section 3.3. Also the function g_n: $X_n \times \Theta \to R^d$ is taken not necessarily asymptotically normal as in (1.2.2) but stochastically bounded satisfying condition I in section 3.3. Therefore we can only conclude in this situation that θ_{1n} of Newton's iterates (1.2.8), for instance, satisfies (1.2.6) with G_n (θ) instead of I (θ).

1.5. Finally, we do not restrict ourself with the special criterion function considered above - the log-likelihood function. The general scheme of experiments described in section 3.1 admits, for instance, partially specified models of regression and time series analysis or partially observable models (see, e.g. Jennrich (1969), Kohn (1978), Robinson (1988), Dzhaparidze and Yaglom (1983), Dzhaparidze (1986), Campillo and Le Grand (1989); see also Celex and Diebolt (1990) and Meilijson (1989) for possible applications to the EM algorithm).

2. Review of non-linear optimization

2.1. Iterative methods. We denote by F = F (x), $x \in R^d$ an objective function F: $R^d \to R^1$ to be minimized, by g = g (x) its gradient vector at x: g (x) = $(\partial/\partial x)$ F (x), and by G = G (x) its Hessian matrix at x: G (x) = $(\partial^2/\partial x^2)$ F (x). Direct methods for solving the minimization problem are usually feasible only for objective functions of very special form. Therefore our attention will be restricted to iterative methods. At the start of the k-th iteration we denote by x_k the current estimate of the minimum. The k-th iteration then consists of computing a *search direction* (vector) p_k and a *steplength* (scalar) α_k from which we obtain the new estimate x_{k+1} according to

(2.1.1) $$x_{k+1} = x_k + \alpha_k p_k.$$

Our main concern is with various methods for determining so-called *descent directions* p_k for which the iterates decrease the function value at each stage

(2.1.2) $$F(x_{k+1}) < F(x_k)$$

at least for sufficiently small positive values of α_k. This inequality is satisfied with $\alpha_k > 0$ when the gradient vector exists and

(2.1.3) $$(g_k, p_k) < 0;$$

here the notation g_k is used to mean g (x_k), similar abbreviations will be used throughout. Indeed, by definition [F ($x_k + \alpha p_k$) - F (x_k)] / $\alpha \to$ (g_k, p_k) as $\alpha \to 0$, so that by (2.1.3) we may choose a $\delta > 0$ to make the left hand side of the last relation negative for all $\alpha \in (0, \delta)$. Hence (2.1.2) takes place for all such α. In accordance with the above considerations, the relationship (2.1.3) is used to define a descent direction p_k at the point x_k: on discussing specific iterative methods below we will always verify the property (2.1.3).

As for a steplength α_k, the maximal possible decrease in (2.1.2) occures for given p_k when it is obtained by *exact linear search* according to the following minimization principle: α is chosen to minimize $F(x_k + \alpha p_k)$ precisely for a given x_k and p_k. If α_k is the value of α that does this, then the chain rule for differentiation shows that $(g(x_k + \alpha_k p_k), (\partial/\partial\alpha)(x_k + \alpha p_k)\,|_{\alpha = \alpha_k}) = (g_{k+1}, p_k) = 0$. Thus after exact linear search the gradient vector g_{k+1} at x_{k+1} becomes orthogonal to p_k:

(2.1.4) $$g_{k+1} \perp p_k \quad \forall\, k.$$

This relationship plays an essential rôle in deriving theoretical properties of iterative methods treated, as we will see below. It should be noted, however, that it is impossible generally to obtain the exact optimum point α_k, as the set of α over which the minimization is taken is too large, and in practice inexact linear searches are performed by terminating according to one or another criteria the search procedure before it has converged finally.

Remark 2.1.1. We will need below the following consequences of exact linear search: (i) by the descency (2.1.3) and orthogonality (2.1.4)

(2.1.5) $$(\Delta g_k, p_k) > 0 \ \text{ where } \ \Delta g_k = g_{k+1} - g_k,$$

or equivalently

(2.1.6) $$(\Delta g_k, \Delta x_k) > 0 \ \text{ where } \ \Delta x_k = x_{k+1} - x_k = \alpha_k p_k$$

by (2.1.1); (ii) the equality in (2.1.6) and orthogonality (2.1.4) yield

(2.1.7) $$g_{k+1} \perp \Delta x_k.$$

2.2. Quadratic termination.

We consider here only methods possessing a property called quadratic termination which means that they minimize a quadratic function exactly in a finite number of iterations. This quadratic objective function can be written in the form

(2.2.1) $$F(x) = \frac{1}{2}(A\,x, x) + (b, x) + c$$

where A is a symmetric positive definite matrix, b a vector and c a scalar. Its gradient function is $g(x) = A\,x + b$ and the Hessian matrix is $G(x) = A$. Thus the gradient vector and Hessian matrix are related here by the following equality

(2.2.2) $$\Delta g_k = A\,\Delta x_k = \alpha_k A\, p_k$$

where Δg_k and Δx_k are such as in (2.1.5) and (2.1.6). Note that in the present case $(g_{k+1}, p_k) = (\Delta g_k + g_k, p_k) = (A\,\Delta x_k + g_k, p_k) = (\alpha_k A\, p_k + g_k, p_k)$ and (2.1.4), valid when exact linear search is performed, is equivalent to

(2.2.3) $$\alpha_k = -(g_k, p_k)\,/\,\lambda_k \ \text{ with } \ \lambda_k = (A\,p_k, p_k) > 0$$

since A is positive definite. Notice also that with iterates (2.1.1) applied to (2.2.1) we have $F(x_{k+1}) - F(x_k) = \alpha_k (g_k, p_k) + \alpha_k^2 (A\,p_k, p_k)\,/\,2$. Hence in case of exact linear search, when α_k is given by (2.2.3), the iterates decrease the function value at each stage: $F(x_{k+1}) - F(x_k) = -(g_k, p_k)^2 / 2\lambda_k$.

2.3. Newton's method. We mention briefly the classical Newton method designed to achieve quadratic termination with a positive definite Hessian matrix in a single iterate given the first and second derivatives of the objective function. It is defined by steplengths $\alpha_k = 1$ and search vectors $p_k = - G_k^{-1} g_k$. Obviously, (2.1.3) is satisfied whenever G_k is positive definite, i.e. the method is descent. As is easily seen, this method indeed achieves quadratic termination in a single iterate: for any initial point x_0 the gradient $g(x) = A x + b$ vanishes at the first iterate $x_1 = x_0 - G_0^{-1} g_0$ as $G_0 = A$ and $g_0 = A x_0 + b$.

Newton's method is the most rapidly convergent method when the Hessian matrix of the objective function is available, however the convergence to the minimum is not guaranteed from an arbitrary starting point. Problems arise when the Hessian matrix is indefinite or singular - in fact few practical problems have a Hessian matrix that is everywhere positive definite. To overcome these difficulties Newton's method is modified by using instead of G_k some other always positive definite matrix which is otherwise close to G_k. We do not dwell here on this kind of modifications; a short review can be found e.g. in Brodlie (1977), section 2 or Scales (1985), section 3.3, see in particular the Gill-Murray algorithm on p. 67.

2.4. Conjugate directions. Quadratic termination. A set of non-zero vectors p_k is said to be mutually *conjugate* with respect to a symmetric positive definite matrix A iff $(A p_k, p_j) = 0$ for $k \neq j$. Using the same symbol $\perp$ as in (2.1.4), we may express this as follows:

(2.4.1) $$A p_k \perp p_j \quad k \neq j.$$

It may be also useful to rewrite (2.4.1) in a matrix form by introducing the matrix Π with columns $p_0, ..., p_{d-1}$: $\Pi = [p_0, ..., p_{d-1}]$. Denote its transpose by Π^T. The relationship (2.4.1) and positive definiteness of A yield

(2.4.2) $$\Pi^T A \Pi = \Lambda = \text{diag}\ \{\lambda_0, ..., \lambda_{d-1}\}$$

where Λ is a diagonal matrix with positive entries $\lambda_0, ..., \lambda_{d-1}$ along the main diagonal (cf. (2.2.3)). As is easily seen, conjugate vectors are linearly independent, that is, it is impossible to find a linear combination such that $\sum_j b_j p_j = 0$ unless all of the coefficients b_j are zero. In matrix notations this means that it is impossible to find a non-zero vector b such that $\Pi\ b = 0$. Indeed, the last equation implies $\Lambda\ b = \Pi^T A\ \Pi\ b = 0$, since A is positive definite. Hence $b = 0$, for Λ is a diagonal matrix with positive entries along the main diagonal; cf. (2.4.2).

Proposition 2.4.1. *Let a quadratic objective function* F *of form* (2.2.1) *be minimized by applying iterates* (2.1.1) *with steplenghts* α_k *obtained by exact linear search and mutually conjugate search vectors* p_k *with respect to the positive definite Hessian matrix*

A. *Then the exact minimum of* F *is located in at most* d *such iterates, thereby the method has quadratic termination.*

Proof. By assumption (2.1.4) and (2.4.1) hold. Due to (2.1.4) the identity $g_k = g_{j+1} + (\Delta g_{j+1} + ... + \Delta g_{k-1})$ with $j < k$, premultiplicated by p_j^T yields

$$(2.4.3) \qquad \begin{aligned}(p_j, g_k) &= (p_j, g_{j+1}) + (p_j, \Delta g_{j+1} + ... + \Delta g_{k-1}) \\ &= (p_j, \Delta g_{j+1} + ... + \Delta g_{k-1}) \quad j < k.\end{aligned}$$

Apply now (2.2.2) to the right hand side of (2.4.3) and then use the conjugacy property (2.4.1). We get

$$(2.4.4) \qquad (p_j, g_k) = (A\, p_j, \alpha_{j+1}\, p_{j+1} + ... + \alpha_{k-1}\, p_{k-1}) = 0 \qquad j < k.$$

When $k = d$ we have $g_d \perp p_j$ with $j < d$. Now, if g_d were non zero it would have to be orthogonal to all the vectors $p_0, ... , p_{d-1}$. But this is impossible as we would then have $d + 1$ linearly independent vectors in a d-dimensional space. Thus $g_d = 0$. Therefore x_d is the minimum of F because A is positive definite. It is always possible that the minimum is located in fewer than d iterations by chance, say at the r-th iterate with $r < d$. In this case $\alpha_r = 0$ (cf. (2.2.3)), and $x_r = x_{r+1} = ... = x_r$. ◊

The above arguments show that g_k must be orthogonal to all $p_0, ... , p_{k-1}$ if it is not already zero. In this case g_k has no component in the subspace spanned by $p_0, ... , p_{k-1}$, and x_k is therefore the minimum in this subspace.

Remark 2.4.2. Under the circumstances of proposition 2.4.1 the conjugacy property (2.4.1) may be expressed as

$$(2.4.5) \qquad (\Delta g_j, p_k) = 0 \quad k \neq j.$$

Indeed, multiply (2.4.1) by α_j and use (2.2.2).

2.5. Conjugate gradient methods. The essential problem in carrying out a conjugate search is to obtain the conjugate directions $p_0, ... , p_{d-1}$. A direct approach to compute $p_0, ... , p_{d-1}$ by Gram - Schmidt orthogonalization is very inefficient. A much more useful procedure can by based on the iterates (2.1.1) where the search vectors p_k are generated simultaneously with x_k according to

$$(2.5.1) \qquad p_0 = -\, g_0, \;\; p_k = -\, g_k + \beta_k\, p_{k-1}, \;\; k = 1, 2, ...$$

where β_k is a positive scalar that distinguishes one particular, so-called conjugate gradient method from another. As is easily seen, these methods used with exact linear search are descent, for (2.5.1) gives descent directions: (2.1.3) is verified by using (2.1.4). Most common choices for β_k are

$$(2.5.2) \qquad \begin{aligned}\beta_k &= (\Delta g_{k-1}, g_k) \,/\, (\Delta g_{k-1}, p_{k-1}), \\ \beta_k &= (\Delta g_{k-1}, g_k) \,/\, |\, g_{k-1}\, |^2, \\ \beta_k &= |\, g_k\, |^2 \,/\, |\, g_{k-1}\, |^2.\end{aligned}$$

If exact linear search is used, the first two of these expressions in fact coincide, for then

$$(2.5.3) \qquad |\, g_k\, |^2 = -\, (g_k, p_k) = (\Delta g_k, p_k).$$

Indeed, the second equality follows from (2.1.4) and the first from definition (2.5.1) as $(g_k, p_k + g_k) = \beta_k\, (g_k, p_{k-1}) = 0$ by applying (2.1.4) again. Observe that by the first of the equalities (2.5.3) the direction p_k is descent; cf. (2.1.3). As for the third expression in (2.5.2), it does not differ from the other two under the circumstances of the next paragraph, for the second of the relations (2.5.6), verified below in the course of proving theorem 2.5.1, implies $(\Delta g_{k-1}, g_k) = | g_k |^2$.

Suppose now that a quadratic function of form (2.2.1) is minimized using exact linear search and one of the equivalent relations (2.5.2). Here exact linear search means that by (2.2.3) and the first of equalities (2.5.3)

(2.5.4) $$\alpha_k = | g_k |^2 / \lambda_k.$$

Note that we can use (2.2.2) to rewrite the first of expressions (2.5.2) as follows: $\beta_k = (A\, p_{k-1}, g_k) / \lambda_{k-1}$. Hence premultiplying p_k in (2.5.1) by $p_{k-1}{}^T A$ we get

(2.5.5) $$(A\, p_{k-1}, p_k) = 0.$$

As is shown in the course of proving the next theorem, (2.5.5) can be extended to (2.4.1): the search vectors (2.5.1) are mutually conjugate with respect to a positive definite matrix A in (2.2.1). Thus by proposition 2.4.1 the conjugate gradient methods achieve quadratic termination.

Theorem 2.5.1. *Let a quadratic objective function* F *of form* (2.2.1) *be minimized by applying iterates* (2.1.1) *with steplengths* α_k *and search vectors* p_k *obtained by exact linear search and one of the conjugate gradient methods* (see (2.5.4) and (2.5.1) with one of the equivalent expressions (2.5.2) for β_k). *Then the exact minimum of* F *is located in at most* d *such iterations, thereby the method has quadratic termination.*

Proof. Note first that α_k and β_k, and hence x_{k+1} and p_{k+1}, are well defined provided p_k is non-zero. If $p_0 = -A\, x_0 - b = 0$ then x_0 is the solution and (2.4.1) holds trivially. Assume $p_k \neq 0$, $k < r$ for some $r \leq d$. Then $g_k \neq 0$, and hence $\alpha_k > 0$, $k < r$ (see (2.5.4)), because if some $g_k = 0$, then $\beta_k = 0$ by (2.5.2). This leads to the contradiction $p_k = g_k = 0$.

We now introduce the induction hypothesis

(2.5.6) $$(A\, p_k, p_j) = 0 \text{ and } (g_k, g_j) = 0,\ j = 0, ..., k - 1 \text{ for some } k \leq r.$$

For k = 1 the relations (2.5.6) clearly hold by (2.5.5) and (2.1.4) respectively. By (2.2.2), (2.5.1) and (2.5.6) we have

$$(g_{k+1}, g_j) = (g_k + \alpha_k A\, p_k, g_j) = (g_k, g_j) + \alpha_k (A\, p_k, -p_j + \beta_j\, p_{j-1}) = 0 \quad j < k,$$

and hence the second relation in (2.5.6) holds for k + 1 since also

$$(g_{k+1}, g_k) = (g_{k+1}, - p_k + \beta_k\, p_{k-1}) = \beta_k (g_{k+1}, p_{k-1}) = \beta_k (g_k + \alpha_k A\, p_k, p_{k-1}) = 0.$$

As $\alpha_j \neq 0$, we have

$$(A\, p_{k+1}, p_j) = (A\, (- g_{k+1} + \beta_{k+1}\, p_k), p_j) = - (A\, g_{k+1}, p_j) = (g_{k+1}, \Delta\, g_j) / \alpha_j = 0$$

for $j < k$ as well as for $j = k$, since $(A\, p_{k+1}, p_k) = 0$ by (2.5.5). The induction now is completed and (2.4.1) is proved.

To complete the proof assume $r < d$ and $p_r = 0$. Then by (2.1.4) and (2.5.1) we have $0 = | p_r |^2 = | g_r |^2 + | p_{r-1} |^2 \geq | g_r |^2$, so that $g_r = 0$ and x_r is the desired minimizer. On the other hand, if $r = d$, then $p_0, ..., p_{d-1}$ form a conjugate basis for A and the result follows from proposition 2.4.1. ◊

2.6. Quasi-Newton methods. The methods presented here are usually more rapidly convergent, robust and economical then conjugate gradient methods when the two can be compared, but they require much more storage and are therefore less suitable for very large problems.

Look at Newton's search directions in section 2.3 and substitute G_k^{-1} with some way an approximation H_k, i.e. write

(2.6.1) $$p_k = - H_k g_k$$

and assume the approximation H_k is "improved" at each step by the following updating formula

(2.6.2) $$H_{k+1} = H_k + Q_k$$

where Q_k is calculated from the values of x_k, x_{k+1}, g_k, g_{k+1} and H_k. The initial approximation H_0 can be any positive definite matrix. We require H_{k+1} to have some of the properties of the inverse Hessian matrix and so, according to (2.2.2), we choose Q_k in such a way that the following, so-called quasi-Newton condition is satisfied (see remark 2.6.2 below):

(2.6.3) $$\Delta x_k = H_{k+1} \Delta g_k.$$

Furthermore, as is show below in the course of proving theorem 2.6.3, all the updates Q_k under consideration satisfy the following extension of (2.6.3), called the heredity condition:

(2.6.4) $$\Delta x_j = H_k \Delta g_j \quad j < k.$$

It is shown also that these updates Q_k determine via (2.6.1) and (2.6.2) search directions p_k with conjugacy property (2.4.1), provided the iterates (2.1.1) are applied with exact linear search to minimize a quadratic function (2.2.1). The methods then have quadratic termination by proposition 2.4.1.

Updates Q_k have to inherit also the symmetry and positive definiteness of G_k^{-1}, and then H_{k+1} is symmetric and positive definite if H_k is; see lemma 2.6.1 below.

The first updating formula widely applied to function minimization was the Davidon-Fletcher-Powell (DFP) formula:

(2.6.5) $$H_{k+1} = H_k + (\Delta x_k, \Delta g_k)^{-1} \Delta x_k \Delta x_k^T - (H_k \Delta g_k, \Delta g_k)^{-1} H_k \Delta g_k \Delta g_k^T H_k.$$

It is a member of Broyden's family of updating formulas given up to a scalar parameter π_k by the following equation

(2.6.6) $$H_{k+1} = H_k + Q_k^{(1)} + Q_k^{(2)}$$

where

$$Q_k^{(1)} = (\Delta x_k, \Delta g_k)^{-1} \Delta x_k \Delta x_k^T$$

and

$$Q_k^{(2)} = - (H_k \Delta g_k, \Delta g_k)^{-1} H_k \Delta g_k \Delta g_k^T H_k + \pi_k (H_k \Delta g_k, \Delta g_k) w_k w_k^T$$

with

$$w_k = (\Delta x_k, \Delta g_k)^{-1} \Delta x_k - (H_k \Delta g_k, \Delta g_k)^{-1} H_k \Delta g_k.$$

Indeed, as $\pi_k = 0$, (2.6.6) coincides with (2.6.5), while $\pi_k = 1$ gives a new formula which is known as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) formula. It can be tidied up into the form

(2.6.7) $\quad H_{k+1} = Q_k^{(1)} + W_k H_k W_k^T$ with $W_k = I - (\Delta x_k, \Delta g_k)^{-1} \Delta x_k \Delta g_k^T$.

Broyden's family of updating formulas (2.6.6) involves only one parameter to distinguish individual quasi-Newton methods. It should be stressed however that according to Dixon (1972) each member of Broyden's family generates the same sequence of points x_k when minimizing a general objective function F, though it is observed in practice that the choice of π_k could make a substantial difference to the performance, and that the DFGS formula seems to be superior. As Dixon's result puts any difference down to inaccuracy in line search (and rounding errors), the reason for the supremacy of the DFGS is sought in the context of inexact line search; cf. Brodlie (1977).

It is easily checked that the sequence of approximation matrices H_k remains positive definite for a wide range, including the points $\pi_k = 0$ and 1, of parameter values (see, e.g. Scales (1985), section 3.5.6). As the arguments used in the simplest special case of $\pi_k = 0$ are typical, we restrict our attention to this case; see lemma 2.6.1 below. Suppose, meanwhile, that H_k is positive definite. Obviously, the direction (2.6.1) is then descent, as (2.1.3) is verified.

Lemma 2.6.1. *If* H_k *is positive definite and a steplength* α_k *is chosen by exact linear search, then* H_{k+1} *defined by the updating formula* (2.6.5) *is positive definite.*

Proof. For an arbitrary non-zero vector x we have by (2.6.5) that

$$(H_{k+1}x, x) = (H_k x, x) - (H_k \Delta g_k, \Delta g_k)^{-1} (H_k \Delta g_k, x)^2 + (\Delta x_k, \Delta g_k)^{-1} (\Delta x_k, x)^2.$$

The denominators in the last two terms are positive: the first by assumption and the second by (2.1.6) which holds because exact linear search is performed and the direction is descent; cf. remark 2.1.1. Due to Schwartz' inequality $(H_k \Delta g_k, x)^2 \leq (H_k x, x)(H_k \Delta g_k, \Delta g_k)$ the combination of the first two terms is non-negative, with equality if $x = c\, \Delta g_k$ for some non-zero constant c. But then the third term is strongly positive. Thus $(H_{k+1}x, x) > 0$. ◊

Remark 2.6.2. By comparing (2.6.2) and (2.6.6) we see that the update is splitted in two terms $Q_k = Q_k^{(1)} + Q_k^{(2)}$. To get the idea of this splitting observe that $Q_k^{(1)} \Delta g_k = \Delta x_k$ and $Q_k^{(2)} \Delta g_k = - H_k \Delta g_k$, since $w_k \perp \Delta g_k$. Therefore $H_{k+1} \Delta g_k = Q_k^{(1)} \Delta g_k = \Delta x_k$ which yields the quasi-Newton property (2.6.3) of Broyden's family of updates (2.6.6). This observation is related in a certain way to the assertion of corollary 2.6.4 to theorem 2.6.3 below.

Suppose a quadratic objective function F of form (2.2.1) is minimized by applying iterates (2.1.1) with exact linear search and search vectors (2.6.1). Then the steplength is positive and given by the following formula (cf. (2.2.3)):

(2.6.8) $\qquad \alpha_k = (H_k g_k, g_k) / \lambda_k$ with $\lambda_k = (A\, p_k, p_k) > 0$.

Theorem 2.6.3. *Let a quadratic objective function* F *of form* (2.2.1) *is minimized by applying iterates* (2.1.1) *with steplengths* α_k *and search vectors* p_k *obtained by exact linear search and one of the quasi-Newton methods* (see (2.6.8) and (2.6.1) with updates

(2.6.6)). *Then the exact minimum of* F *is located in at most* d *such iterations, thereby the method has quadratic termination.*

Proof. We consider only the special case of the DFP updates (2.6.5), for the extension to include the extra term of Broiden's family follows a very similar line. Assume that H_k is well defined (symmetric and positive definite) and g_k is non-zero. Then $\alpha_k > 0$ by (2.6.8) and H_{k+1} is well defined, according to lemma 2.6.1. This means that x_{k+1} is also well defined. Thus, by induction, all $x_0, ..., x_k$ are well defined provided $g_0, ..., g_{k-1}$ were non-zero. Assume therefore that $g_0, ..., g_{d-1}$ are non-zero, or else the result is proved.

We prove by induction that (2.4.5) and (2.6.4) hold. Note that by remark 2.4.2 relationship (2.4.5) expresses the conjugacy of search vectors with respect to the Hessian matrix G (x) = A of the quadratic objective function F under consideration. First assume k = 1. Then (2.6.4) turns into the quasi-Newton condition (2.6.3); cf. remark 2.6.2. To get (2.4.5) with k = 1 apply successively definition (2.6.1), heredity (2.6.4) and finally the consequence (2.1.7) of exact linear search: $(\Delta g_0, p_1) = -(H_1 \Delta g_0, g_1) = -(\Delta x_0, g_1) = 0$.

Assume (2.4.5) and (2.6.4) hold for some k < d. We handle first conjugacy proving (2.4.5) with k substituted by k + 1. Taking into consideration the consequence (2.1.7) of exact linear search we get by definitions (2.6.1) and (2.6.5) that

$$\begin{aligned}(\Delta g_j, p_{k+1}) &= -(H_{k+1} \Delta g_j, g_{k+1})\\ &= -(H_k \Delta g_j, g_{k+1}) + (H_k \Delta g_k, g_{k+1})(H_k \Delta g_j, g_k) / (H_k \Delta g_k, g_k) \quad j \leq k.\end{aligned}$$

Obviously, the right hand side vanishes as j = k. In case j < k we see that both terms on the right hand side vanish: apply the induction hypothesis and use heredity (2.6.4), then the equality in (2.1.6) and finally (2.4.4) and (2.4.5). We get

$$(H_k \Delta g_j, g_{k+1}) = (\Delta x_j, g_k + \Delta g_k) = \alpha_j (p_j, g_k + \Delta g_k) = 0 \quad j < k$$

and

$$(H_k \Delta g_j, \Delta g_j) = (\Delta x_j, \Delta g_k) = \alpha_j (p_j, \Delta g_k) = 0 \quad j < k. \tag{2.6.9}$$

Thus the induction concerning conjugacy is complete.

As for heredity, look at (2.6.4) with k substituted by k+1, i.e. with the right hand side of form $H_{k+1} \Delta g_j = [H_k + Q_k^{(1)} + Q_k^{(2)}] \Delta g_j$. For k = j the result is clear: see remark 2.6.2 where quasi-Newton property (2.6.3) is shown. For j < k the second and third terms vanish due to (2.4.5) and (2.6.9) respectively, while the first term turns into desired Δx_j by the induction hypothesis. Thus the proof of heredity (2.6.4) is complete. By assumption $g_0, ..., g_{d-1}$ are non-zero, hence it follows from proposition 2.4.1 that $g_d = 0$. ◊

We assume in the remainder of this section that the exact minimum of a quadratic objective function F is located in exactly d iterations, i.e. $g_0, ..., g_{d-1}$ are non-zero. Then in view of (2.2.2) the heredity property (2.6.4) means that

$$H_d A \Delta x_k = \Delta x_k \quad k < d, \tag{2.6.10}$$

that is $H_d A$ has d linearly independent eigenvectors $\Delta x_0, ..., \Delta x_{d-1}$ with eigenvalues unity. Hence $H_d A = I$. This observation yields the following corollary to proposition 2.6.3.

Corollary 2.6.4. *Assume that under the circumstances of theorem 2.6.3 the exact minimum of* F *is located in exactly d iterations. Then* $H_d = A^{-1}$. *Moreover ,*

(2.6.11) $\quad Q_0^{(1)} + \ldots + Q_{d-1}^{(1)} = A^{-1}$ *and* $Q_0^{(2)} + \ldots + Q_{d-1}^{(2)} = - H_0$
where $Q_k^{(1)}$ *and* $Q_k^{(2)}$ *are such as in* (2.6.6).

Proof. We have already shown that $H_d = A^{-1}$. Since $H_d = H_0 + (Q_0^{(1)} + \ldots + Q_{d-1}^{(1)}) + (Q_0^{(2)} + \ldots + Q_{d-1}^{(2)})$, it suffices to prove the first of the relationships in (2.6.11). In view of the equality in (2.1.6), conjugacy (2.4.1) means that $(A\, \Delta x_k, \Delta x_j) = 0$ with $k \neq j$. Hence by definition of $Q_k^{(1)}$ we have

(2.6.12) $\quad (Q_0^{(1)} + \ldots + Q_{d-1}^{(1)})\, A\, \Delta x_k = \Delta x_k \; k < d.$

To complete the proof of the first of relationships (2.6.11), compare (2.6.10) and (2.6.12) and use the same considerations as in the course of verifying the equality $H_d = A^{-1}$. ◊

3. Statistical estimation of parameters

3.1. Experiment. Consider an experiment $\mathcal{E} = \{\mathfrak{X}, \mathfrak{A}, P \in \mathbb{P}\}$ where $\mathfrak{X}$ is the set of possible outcomes, $\mathfrak{A}$ a σ-algebra defined on $\mathfrak{X}$, and $\mathbb{P}$ a family of distributions on $\mathfrak{A}$. We observe $X \in \mathfrak{X}$ drawn according to some unknown $P \in \mathbb{P}$ and wish to make inference concerning P, or rather its certain characteristics. Suppose we have at our disposal a certain $\mathfrak{A}$-measurable non-negative criterion function $F: X \to R_+$ which depends on P, i.e. F = F (X, P), exclusively through these characteristics. These dependence is supposed to allow the following finite dimensional parametrization. For each fixed $P \in \mathbb{P}$ define a subset [P] of $\mathbb{P}$ by $[P] = \{P': F(X, P) = F(X, P') \text{ for all } X \in \mathfrak{X}\}$. Suppose now that $[\mathbb{P}] = \{[P]: P \in \mathbb{P}\}$ allows a d-dimensional parametrization: there exists a one to one mapping $\vartheta: [\mathbb{P}] \to \Theta$ where Θ is an open set in R^d. Thus by definition of [P] this mapping induces only a partial parametrization upon the model: only the characteristics involved in F are parametrized, and apart from X, the criterion function F depends on a parameter value $\theta \in \Theta$ only. Therefore the index P will be substituted by θ, and the criterion function will be written in form F (X, θ) whenever $P \in [P] = \vartheta^{-1}(\theta)$ for some $\theta \in \Theta$.

3.2. A sequence of parametric experiments. Suppose that the experiment considered above is n-th in a certain sequence of experiments $\mathcal{E}_1, \mathcal{E}_2, \ldots$ and index all quantities introduced in 3.1 by n, except the parameter $\theta \in \Theta$. Our inference concerning $\theta \in \Theta$ is of asymptotic nature: it is valid for n large enough. We will often deal with sequences of vector valued random variables $\{x_n\}_{n=1,2,\ldots}$ depending sometimes on a parameter value i.e. $x_n = x_n(\theta)$, which converges in norm as $n \to \infty$ to zero in probability $P_n \in \vartheta^{-1}(\theta)$, i.e. $P_n (| x_n(\theta) | > \varepsilon) \to 0$ for each $\theta \in \Theta$ and $\varepsilon > 0$.

Definition 3.2.1. If sequences of vector valued random variables $\{x_n(\theta)\}_{n=1,2,\ldots}$ and $\{y_n(\theta)\}_{n=1,2,\ldots}$ are such that $| (x_n(\theta), y_n(\theta)) |$ converges as $n \to \infty$ to zero in probability $P_n \in \vartheta^{-1}(\theta)$ for each $\theta \in \Theta$, then we say that $x_n(\theta)$ and $y_n(\theta)$ are *asymptotically orthogonal* in probability $P_n \in \vartheta^{-1}(\theta)$, expressing this relationship symbolically as follows (analogously to (2.1.4)):

(3.2.1) $$x_n(\theta) \perp^{(P_n)} y_n(\theta).$$

A sequence of vector valued random variables $\{y_n(\theta)\}_{n=1,2,...}$ is called *stochastically bounded* if it is asymptotically orthogonal to any sequence $\{x_n(\theta)\}_{n=1,2,...}$ converging in norm as $n \to \infty$ to zero in probability.

Remark 3.2.2. In particular, a sequence $\{y_n(\theta)\}_{n=1,2,...}$ is stochastically bounded if for every $\varepsilon > 0$ there exists $b = b(\varepsilon) > 0$ such that for all n large enough

(3.2.2) $$P_n(|\, y_n(\theta)\,| > b) < \varepsilon \quad \forall\, P_n \in \vartheta^{-1}(\theta).$$

For simplicity of exposition we will assume below that for a stochastically bounded sequence $\{y_n(\theta)\}_{n=1,2,...}$ the relation (3.2.2) is valid for all n large enough.

On treating problems of estimating the parameter $\theta \in \Theta$, we consider only estimators of θ which are δ_n-consistent in the sense of definition 3.2.3 below, where $\{\delta_n\}_{n=1,2,...}$ is a sequence of positive definite matrices in $R^d \times R^d$ such that $\|\, \delta_n \,\| \to 0$ as $n \to \infty$.

Definition 3.2.3. An $\mathcal{Q}_n$-measurable function T_n: $X_n \to R^d$ is called a *δ_n-consistent estimator* of the parameter $\theta \in \Theta$ if for each $\theta \in \Theta$ the sequence $\{\delta_n^{-1}(T_n - \theta)\}_{n=1,2,...}$ is stochastically bounded; cf. definition 3.2.1 and remark 3.2.2. For convenience, the class of all δ_n-consistent estimators of the parameter $\theta \in \Theta$ will be denoted by $\mathbb{D}$.

Remark 3.2.4. Clearly, the event $\{T_n \notin \Theta\}$ is not excluded. Since, however, Θ is an open set in R^d, the probability $P_n \in \vartheta^{-1}(\theta)$ of this event tends to zero as $n \to \infty$. Therefore, in order to avoid troubles, for instance, in evaluating at T_n functions defined only on $\theta \in \Theta$, we may use without affecting asymptotic inference the following convention: on evaluating at T_n a function $g(\theta)$ defined on $\theta \in \Theta$ substitute T_n if $T_n \in \Theta$ and an arbitrary value of θ from Θ if $T_n \notin \Theta$. Thus writing below $g(T_n)$ we always assume without further comments this convention.

3.3. Basic conditions. We suppose that our criterion function $F_n(X_n, \theta)$ is asymptotically quadratic in the sense of the following

Definition 3.3.1. For a fixed parameter value $\theta \in \Theta$ consider a perturbation $\theta + \delta_n u$ in direction $u \in \mathcal{U}_n = \delta_n^{-1}(\Theta - \theta)$. We say that a criterion function $F_n(X_n, \theta)$ is *asymptotically quadratic* if for each $\theta \in \Theta$ there exist $\mathcal{Q}_n$-measurable (for each fixed $\theta \in \Theta$) functions g_n: $X_n \times \Theta \to R^d$ and G_n: $X_n \times \Theta \to R^d \times R^d$ such that

(3.3.1) $$F_n(X_n, \theta + \delta_n u) - F_n(X_n, \theta) = (u, g_n(X_n, \theta)) + \frac{1}{2}(G_n(\theta)u, u) + \eta_n(\theta, u)$$

where g_n, G_n and η_n satisfy the following conditions:

I. The sequence $\{g_n(X_n, \theta)\}_{n=1,2,...}$ is stochastically bounded; cf. definition 3.2.1 and remark 3.2.2.

II. The symmetric matrix $G_n(\theta)$ is stochastically bounded from below and above in the sense that for every $\varepsilon > 0$ there exist $a = a(\varepsilon)$ and $b = b(\varepsilon)$, $0 < a \le b < \infty$ such that for all n large enough the event $\{a < \inf_{\theta \in \Theta} \inf_{|u|=1} (G_n (\theta)u, u) \leq \sup_{\theta \in \Theta} \sup_{|u|=1} (G_n (\theta)u, u) < b\}$ occurs with probability P_n exceeding $1 - \varepsilon$, for all $P_n \in \vartheta^{-1}(\theta)$; cf. remark 3.2.2.

III. For each $\theta \in \Theta$ and $u \in \mathcal{U}_n$, a remainder term $\eta_n (\theta, u)$ tends to zero as $n \to \infty$ in probability $P_n \in \vartheta^{-1}(\theta)$, i.e. $P_n (| \eta_n (\theta, u) | > \varepsilon) \to 0$ for every $\varepsilon > 0$.

Note that (3.3.1) can be viewed as a kind of stochastic version of Teylor's expansion, valid in many applications. Typically, g_n can be taken as δ_n times the gradient vector of F_n with respect to θ, or rather its suitable (e.g. as smooth in θ as possible) approximation up to terms which can be absorbed in the remainder η_n, and G_n as the stochastic limit as $n \to \infty$ of the Hessian matrix premultiplicated from both sides by δ_n. Hence, apart from (3.3.1), one can expect the following relationship between g_n and G_n: for each $\theta \in \Theta$ and $u \in \mathcal{U}_n$ and every $\varepsilon > 0$

(3.3.2) $\qquad P_n (| g_n (X_n, \theta + \delta_n u) - g_n (X_n, \theta) + G_n (\theta)u | > \varepsilon) \to 0 \text{ as } n \to \infty$

for all $P_n \in \vartheta^{-1}(\theta)$. Furthermore in many cases u in (3.3.2) can be substituted by the stochastically bounded random vector $\delta_n^{-1} (T_n - \theta)$, $T_n \in \mathbb{D}$ (see definition 3.2.3). Therefore the following condition can be verified (cf. comments on relationships (1.2.4) and (1.2.5) in section 1.3):

IV. For each $\theta \in \Theta$ and $\varepsilon > 0$

$$P_n (| g_n (X_n, T_n) - g_n (X_n, \theta) - G_n (\theta) \delta_n^{-1} (T_n - \theta) | > \varepsilon) \to 0 \text{ as } n \to \infty$$

with $P_n \in \vartheta^{-1}(\theta)$, provided $g_n (X_n, T_n)$ is $\mathcal{A}_n$-measurable and condition I extends to it.

3.4. Iterative procedure. Under condition IV we may single out a subclass D of $\mathbb{D}$ consisting of all δ_n-consistent estimators T_n such that for each $\theta \in \Theta$ and $\varepsilon > 0$

(3.4.1) $\qquad P_n (| g_n (X_n, T_n) | > \varepsilon) \to 0 \text{ as } n \to \infty$

with $P_n \in \vartheta^{-1}(\theta)$. Thus, by condition IV every estimator $T_n \in D$ satisfies the following relation: for each $\theta \in \Theta$ and $\varepsilon > 0$

(3.4.2) $\qquad P_n (| g_n (X_n, \theta) + G_n (\theta) \delta_n^{-1} (T_n - \theta) | > \varepsilon) \to 0 \text{ as } n \to \infty$

with $P_n \in \vartheta^{-1}(\theta)$.

The special estimator defined as the minimizer of the criterion function F_n over Θ, or as the root of the corresponding gradient equation $g_n (X_n, \theta) = 0$, belongs to D, provided it exists (this is so, for instance, when F_n is continuous over a compact set Θ) and it is $\mathcal{A}_n$-measurable and δ_n-consistent (for appropriate conditions see Sieders and Dzhaparidze (1987)). Typically, however, the above minimization problem admits only some "approximate solution", so that the actual problem is to prove that this approximate solution, used as an estimator for θ, belongs to D. The present paper is aimed at proving that by applying iterative methods of section 2, or rather their stochastic modification, we arrive at a desired approximate solution in finitely many steps, provided we are given

(i) a δ_n-consistent estimator of θ used as the initial point θ_{0n}, and

(ii) a "consistent estimator" G_{0n} of $G_n (\theta)$, a $\mathcal{A}_n$-measurable symmetric matrix valued function G_n: $X_n \rightarrow R^d \times R^d$ such that for every $\varepsilon > 0$ we have $P_n(| G_{0n} - G_n(\theta) | > \varepsilon) \rightarrow 0$ as $n \rightarrow \infty$, with $P_n \in \vartheta^{-1}(\theta)$.

Remark 3.4.1. Note that if $G_n(\theta)$ is continuous in θ, then $G_n(T_n)$ with any $T_n \in \mathbb{D}$, for instance $\theta_{0n} \in \mathbb{D}$, may be used as a consistent estimator G_{0n} of $G_n(\theta)$. We assume that not only θ_{0n} is corrected according to remark 3.2.4 but G_{0n} as well, to possess the properties of $G_n(\theta)$ stipulated in section 3.3, condition II.

Thus we suppose that an asymptotically quadratic criterion function $F_n(X_n, \theta)$ is given which can be expanded as in (3.3.1) with $g_n(X_n, \theta)$ and $G_n(\theta)$ satisfying conditions I-IV in section 3.3, as well as an initial point θ_{0n} and a "consistent estimator" G_{0n} corrected according to remarks 3.2.4 and 3.4.1. We will form then the following iterative procedure of estimating the parameter θ:

(3.4.3) $$\theta_{k+1n} = \theta_{kn} + \delta_n \alpha_{kn} p_{kn}$$

with a steplength α_{kn} and search vector p_{kn}, the choice of which distinguishes one particular method from another. Of course, they ought to be calculable from given observations, i.e. $\mathcal{Q}_n$-measurable functions α_{kn}: $X_n \rightarrow R_+$ and p_{kn}: $X_n \rightarrow R^d$ respectively. Besides, both of the sequences $\{\alpha_{kn}\}_{n=1,2,...}$ and $\{p_{kn}\}_{n=1,2,...}$ ought to be chosen stochastically bounded in the sense of definition 3.2.1, in order to guarantee that the iterates (3.4.3) are well defined in the following sense:

Definition 3.4.2. We will say that the iterates (3.4.3) are *well defined* if starting with any $\theta_{0n} \in \mathbb{D}$ we stay in $\mathbb{D}$ untill the iterates are terminated: if a procedure is terminated after r iterates then $\theta_{kn} \in \mathbb{D}$ for $k \leq r$.

Indeed, with the above choice of a steplength α_{kn} and search vector p_{kn} the sequence $\{\Delta\theta_{kn}\}_{n=1,2,...}$ with

(3.4.4) $$\Delta\theta_{kn} = \delta_n^{-1}(\theta_{k+1n} - \theta_{kn}) = \alpha_{kn} p_{kn}$$

is stochastically bounded, as well. Since $\delta_n^{-1}(\theta_{k+1n} - \theta) = \delta_n^{-1}(\theta_{kn} - \theta) + \Delta\theta_{kn}$ and therefore $P_n(| \delta_n^{-1}(\theta_{k+1n} - \theta) | > b) \leq P_n(| \delta_n^{-1}(\theta_{kn} - \theta) | > b) + P_n(| \Delta\theta_{kn} | > b)$, one can choose here b large enough to render both terms on the right hand side arbitrarily small: the first by the assumption $\theta_{kn} \in \mathbb{D}$ and the second by the assumption that the sequence $\{\Delta\theta_{kn}\}_{n=1,2,...}$ is stochastically bounded; cf. remark 3.2.2. Hence the iterates (3.4.3) are well defined: $\theta_{k+1n} \in \mathbb{D}$. Denote $g_{kn} = g_n(X_n, \theta_{kn})$ and $\Delta g_{kn} = g_{k+1n} - g_{kn}$ where $g_n(X_n, \theta)$ is such as in definition 3.3.1 of an asymptotically quadratic criterion function. Then by condition IV in section 3.3 we have that for each $\theta \in \Theta$ and $\varepsilon > 0$

(3.4.5) $$P_n(| \Delta g_{kn} - G_n(\theta)\Delta\theta_{kn} | > \varepsilon) \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

for all $P_n \in \vartheta^{-1}(\theta)$. Therefore by condition II in section 3.3 the sequence $\{\Delta g_{kn}\}_{n=1,2,...}$ is also stochastically bounded.

As was said above, the procedure is aimed at finding an estimator $\theta_{rn} \in D$ for which the corresponding g_{rn} possesses the following property: $P_n(| g_{rn} | > \varepsilon) \rightarrow 0$ as $n \rightarrow \infty$ for each $\theta \in \Theta$ and $\varepsilon > 0$, where $P_n \in \vartheta^{-1}(\theta)$. Therefore the procedure is terminated after r iterates if $|\ g_{kn}\ | > \varepsilon$ for $k < r$, while $|\ g_{rn}\ | \leq \varepsilon$ where ε is some positive tolerance value. Now, if the tolerance value ε is small enough and the sample size n large enough, then with the special choice of steplengths α_{kn} and search vectors p_{kn} discussed in sections 3.7-3.9 below the procedure (3.4.3) terminated this way will lead to $\theta_{rn} \in D$ with probability arbitrarily close to 1.

3.5. Asymptotic descency. Starting with some $\theta_{0n} \in \mathbb{D}$, consider an iterative procedure (3.4.3) of estimating the parameter θ which is well defined in the sense of definition 3.4.2 as steplengths α_{kn} and search vectors p_{kn} are supposed stochastically bounded. Similarly to section 2 (cf. relationship (2.1.3)), we always assume that all search directions p_{kn} before termination are chosen asymptotically descent in the following sense:
Definition 3.5.1. We say that search directions p_{kn} are *asymptotically descent* if for each $\theta \in \Theta$ and $\varepsilon > 0$

(3.5.1) $$P_n\,((g_{kn}, p_{kn}) \geq 0) \to 0 \ \text{ as } n \to \infty$$

for all $P_n \in \vartheta^{-1}(\theta)$, where $g_{kn} = g_n\,(X_n, \theta_{kn})$ as above.

Observe that by definition all g_{kn} before terminaton possess the following property: for every $\varepsilon > 0$ there exists $a = a\,(\varepsilon) > 0$ such that for all n large enough the event $\{|\ g_{kn}\ | > a\}$ occurs with probability $P_n \in \vartheta^{-1}(\theta)$ exceeding $1 - \varepsilon$. We suppose throughout that all search directions p_{kn} before termination possess the same property. As a consequence, we have that by condition II in section 3.3 and remark 3.4.1 for every $\varepsilon > 0$ and all n large enough there exist $a = a\,(\varepsilon)$ and $b = b\,(\varepsilon)$, $0 < a \leq b < \infty$ such that the event $\{a < \lambda_{kn} < b\}$ with

(3.5.2) $$\lambda_{kn} = (G_{0n}\, p_{kn}, p_{kn})$$

occurs with probability $P_n \in \vartheta^{-1}(\theta)$ exceeding $1 - \varepsilon$, where. This allows us to choose a steplength similarly to (2.2.3):

(3.5.3) $$\alpha_{kn} = -\,(g_{kn}, p_{kn})\,/\,\lambda_{kn}$$

which is well defined: for each $\theta \in \Theta$

(3.5.4) $$P_n\,(\alpha_{kn} \leq 0) \to 0 \ \text{ as } n \to \infty$$

where $P_n \in \vartheta^{-1}(\theta)$.

Lemma 3.5.2. *The choice of a steplength* α_{kn} *and search vector* p_{kn} *made above provides for exact linear search since for each* $\theta \in \Theta$ *and* $\varepsilon > 0$ *we have* $P_n\,(|\,(g_{k+1n}, p_{kn})\,| > \varepsilon) \to 0$ *as* $n \to \infty$ *where* $P_n \in \vartheta^{-1}(\theta)$.

Remark 3.5.3. According to definition 3.2.1 the last relationship means that g_{k+1n} is asymptotically orthogonal to p_{kn}. Thus analogously to (2.1.4) this property may be expressed by means of the symbol introduced in (3.2.2) as follows:

(3.5.5) $$g_{k+1n} \perp^{(P_n)} p_{kn}.$$

Proof of lemma 3.5.2. By condition (ii) in section 3.4 and remark 3.4.1 we can substitute $G_n\,(\theta)$ in (3.4.5) by its consistent estimator G_{0n}, and then use (3.4.4). We get

(3.5.6) $$P_n\,(|\ \Delta g_{kn} - \alpha_{kn}\, G_{0n}\, p_{kn}\ | > \varepsilon) \to 0 \text{ as } n \to \infty.$$

Since by assumption the sequence $\{p_{kn}\}_{n=1,2,...}$ is stochastically bounded, the last relationship yields in view of definition 3.2.1 that

(3.5.7) $$\Delta g_{kn} - \alpha_{kn} G_{0n} p_{kn} \perp^{(P_n)} p_{kn}.$$

Substitute now in (3.5.7) the expression for α_{kn} given by (3.5.3). We get (3.5.5). ◊

Remark 3.5.4. Similarly to remark 2.1.1 we have the following consequences of asymptotic descency and asymptotic orthogonality (3.5.5). Under the conditions of lemma 3.5.2: (i) by (3.5.1) and (3.5.5) we have for each $\theta \in \Theta$ that

(3.5.8) $$P_n((\Delta g_{kn}, p_{kn}) \leq 0) \to 0 \quad \text{as } n \to \infty$$

with $P_n \in \vartheta^{-1}(\theta)$, or equivalently

(3.5.9) $$P_n((\Delta g_{kn}, \Delta\theta_{kn}) \leq 0) \to 0 \quad \text{as } n \to \infty$$

by (3.4.4) and (3.5.4); (ii) by (3.4.4), (3.5.4) and (3.5.5)

(3.5.10) $$g_{k+1n} \perp^{(P_n)} \Delta\theta_{kn}.$$

3.6. Newton's method. Similarly to section 2.3 we choose the search direction $p_{kn} = -G_{0n}^{-1} g_{kn}$ which by (3.5.3) corresponds to choosing the steplenght $\alpha_{kn} = 1$. As is known, starting with any $\theta_{0n} \in \mathbb{D}$ we get here the desired estimator in just one step: $\theta_{1n} = \theta_{0n} + \delta_n G_{0n}^{-1} g_{0n} \in D$, therefore (3.4.1) and (3.4.2) hold with $T_n = \theta_{1n}$. Indeed, with the present choice of steplengths and search directions, we have by (3.4.5) with $k = 0$ that $P_n(|g_{1n}| > \varepsilon) \to 0$ as $n \to \infty$ for all $P_n \in \vartheta^{-1}(\theta)$. Thus by definition of the set D in section 3.4 the inclusion $\theta_{1n} \in D$ is true, as well as its consequence - (3.4.2) with $T_n = \theta_{1n}$.

3.7. Asymptotically conjugate directions. Consider again an iterative procedure of estimation (3.4.3) with iterates as in section 3.5. The matrix $G_n(\theta)$ in definition 3.7.1 below is such as in definition 3.3.1 of an asymptotically quadratic criterion function.

Definition 3.7.1. We say that a set of directions p_{kn} is *asymptotically conjugate* with respect to $G_n(\theta)$ (cf. (2.4.1)) if for each $\theta \in \Theta$ and $\varepsilon > 0$ we have as $k \neq j$ and $n \to \infty$ that $P_n(|(G_n(\theta)\, p_{kn}, p_{jn})| > \varepsilon) \to 0$ for all $P_n \in \vartheta^{-1}(\theta)$, i.e.

(3.7.1) $$G_n(\theta)\, p_{kn} \perp^{(P_n)} p_{jn} \qquad k \neq j.$$

Remark 3.7.2. If directions p_{kn} are asymptotically conjugate with respect to $G_n(\theta)$, then they are asymptotically conjugate with respect to G_{0n} too, that is, $G_n(\theta)$ in (3.7.1) can be substituted by G_{0n}. Indeed, this is guaranteed by condition II in section 3.3 and by the consistency of G_{0n}; see condition (ii) in section 3.4 and remark 3.4.1.

Similarly to section 2.4 introduce the matrix $\Pi_n = [p_{0n}, ..., p_{d-1n}]$ and consider $\Lambda_n = \Pi_n^T G_{0n} \Pi_n$ which is *asymptotically diagonal* in the sense that the diagonal entries are given by (3.5.2), while all non-diagonal entries vanish in probability $P_n \in \vartheta^{-1}(\theta)$ as $n \to \infty$, for by remark 3.7.2 property (3.7.1) holds with G_{0n} substituted in place of $G_n(\theta)$.

Lemma 3.7.3. *Under the conditions of lemma* 3.5.2 *asymptotically conjugate directions* $p_{0n}, ..., p_{d-1n}$ *are asymptotically linearly independent in the sense that for any sequence of* *random vectors* $\{b_n\}_{n=1,2,...}$ *and all* $P_n \in \vartheta^{-1}(\theta)$ *the relation* $P_n(|\Pi_n b_n| > \varepsilon) \to 0$ *as* n $\to \infty$ *implies* $P_n(|b_n| > \varepsilon) \to 0$.

Proof. Under the required conditions we have $P_n(|\Lambda_n b_n| > \varepsilon) \to 0$, which in turn implies the desired result, since Λ_n is asymptotically diagonal. ◊

Proposition 3.7.4. *Let a criterion function* $F_n(X_n, \theta)$ *be asymptotically quadratic in the sense of definition* 3.3.1. *Suppose we are given a* δ_n*-consistent estimator for* θ *used as the initial point* θ_{0n} *and a consistent estimator* G_{0n} *for* $G_n(\theta)$ (cf. conditions (i) and (ii) in section 3.4). *Consider an iterative procedure* (3.4.3) *of estimating the parameter* θ *which satisfies the conditions of lemma* 3.5.2. *Besides, suppose that steplengths* α_{kn} *are defined by* (3.5.3) *and search vectors* p_{kn} *are asymptotically mutually conjugate with respect to* $G_n(\theta)$. *Then the procedure is terminated in fewer then* d *iterates, say* $r \leq d$, *and* $\theta_{rn} \in D$.

Proof. We will show first that all p_{jn} with $j < k$ are asymptotically orthogonal to g_{kn} in probability $P_n \in \vartheta^{-1}(\theta)$, that is

$$p_{jn} \perp^{(P_n)} g_{kn} \quad j < k. \tag{3.7.2}$$

Since the conditions of lemma 3.5.2 are satisfied, we have (3.5.5). Hence, by the identity (2.4.3) adapted to the present case it suffices to prove that $p_{jn} \perp^{(P_n)} (\Delta g_{j+1n} + ... + \Delta g_{k-1n})$ for all $j < k$ or, by (3.4.5) and condition II in section 3.3, that $p_{jn} \perp^{(P_n)} G_n(\theta)(\Delta\theta_{j+1n} + ... + \Delta\theta_{k-1n})$. In view of definition (3.4.4) we can rewrite the last relationship as $p_{jn} \perp^{(P_n)} G_n(\theta)(\alpha_{j+1n} p_{j+1n} + ... + \alpha_{k-1n} p_{k-1n})$ for $j < k$, and verify its validity by using (3.7.1). The proof of (3.7.2) is complete.

When $k = d$ we have $g_{dn} \perp^{(P_n)} p_{jn}$ for all $j < d$. This means that the conditions of Lemma 3.7.3 are satisfied with $b_n = g_{dn}$, therefore

$$P_n(|g_{dn}| > \varepsilon) \to 0 \quad \text{as } n \to \infty \tag{3.7.3}$$

for all $P_n \in \vartheta^{-1}(\theta)$. In view of the convention in section 3.4 the procedure is terminated at $\theta_{dn} \in D$: if the tolerance value ε is small enough and the sample size n large enough, then according to (3.7.3) the procedure terminated this way leads to $\theta_{dn} \in D$ with probability arbitrarily close to 1. ◊

It is always possible that we get $|g_{rn}| \leq \varepsilon$ in fewer than d iterations by chance with r < d. The above arguments show that if the tolerance value ε is small enough and the sample size n large enough, then the procedure terminated this way at r-th iterate leads to $\theta_{rn} \in D$ with probability arbitrarily close to 1. In this case α_{rn} can be made arbitrarily small due to definition (3.5.3) with probability close to 1, so that all the consequent $\theta_{r+1n}, ..., \theta_{dn}$ belong to D.

Remark 3.7.5. In view of (3.5.4) and (3.5.8) the asymptotic conjugacy (3.7.1) is equivalent to (cf. remark 2.4.2)

$$p_{kn} \perp^{(P_n)} \Delta g_{jn} \quad k \neq j. \tag{3.7.4}$$

3.8. Conjugate gradient methods. Consider an asymptotically quadratic criterion function $F_n(X_n, \theta)$ in the sense of definition 3.3.1, and suppose a δ_n-consistent estimator for θ used as the initial point θ_{0n}, and a consistent estimator G_{0n} for $G_n(\theta)$ are given (cf. conditions (i) and (ii) in section 3.4). Similarly to section 2.5 we define here a special iterative procedure of type (3.4.3) for estimating the parameter θ. Namely we define search vectors p_{kn} as follows:

(3.8.1) $$p_{0n} = -\, g_{0n}, \quad p_{kn} = -\, g_{kn} + \beta_{kn}\, p_{k-1n} \quad k \geq 1$$

with $g_{kn} = g_n(X_n, \theta_{kn})$ as usual and

(3.8.2) $$\beta_{kn} = |\, g_{kn}\,|^2 \,/\, |\, g_{k-1n}\,|^2;$$

see (3.8.7) below for asymptotically equivalent alternatives. As for a steplength α_{kn} defined again by (3.5.3), it provides for exact linear search since (3.5.5) is satisfied according to lemma 3.5.2. Under the condition that β_{kn} is stochastically bounded it also provides for asymptotic descency of iterates, as premultiplying (3.8.1) by $g_{kn}{}^T$ and applying (3.5.5) we get

(3.8.3) $$P_n(|\, (g_{kn}, p_{kn}) + |\, g_{kn}\,|^2\, | > \varepsilon) \to 0 \quad \text{as } n \to \infty$$

for all $P_n \in \vartheta^{-1}(\theta)$ which yields (3.5.1). By (3.8.3) we have the following asymptotically equivalent alternative to (3.5.3):

(3.8.4) $$\alpha_{kn} = |\, g_{kn}\,|^2 \,/\, \lambda_{kn}$$

which obviously satisfies (3.5.4).

Suppose all g_{jn} and β_{jn}, hence all p_{jn}, were stochastically bounded for $j \leq k$. Then the steplength α_{kn} is stochastically bounded as well. In order to guarantee this we shall show that under conditions I - IV stipulated in section 3.3 the present iterates are well defined in the sense of definition 3.4.2. As usual, the iterates are terminated if $|\, g_{rn}\,| \leq \varepsilon$ for some r, so that only β_{kn} and p_{kn} with $k < r$ should be well defined.

Obviously, if θ_{0n} is such that $|\, g_{0n}\,| \leq \varepsilon$ for a tolerance value ε , then there is no need in further iterations as θ_{0n} is already a desired estimator. Suppose $|\, g_{0n}\,| > \varepsilon$. Then $p_{0n} = -\, g_{0n}$ by definition, so that according to (3.8.4) and remark 3.4.1 on properties of G_{0n} the steplength $\alpha_{0n} = |\, g_{0n}\,|^2 \,/\, \lambda_{0n}$ is positive. If a tolerance value ε is small enough and a sample size n large enough, as we always suppose, then the probability of the event $\{|\, g_{0n}\,| > \varepsilon\}$ is close to 1, as well as the probability that α_{0n} is positive. Obviously, (3.5.1) with $k = 0$ is satisfied and the search direction $p_{0n} = -\, g_{0n}$ is asymptotically descent. Since the search direction p_{0n} and steplength α_{0n} are stochastically bounded, by assumptions (i) and (ii) on θ_{0n} and G_{0n} and conditions I and IV in section 3.3, lemma 3.5.2 is applicable: not only θ_{0n} is a δ_n-consistent estimator of θ, but also θ_{1n}.

Suppose $|\, g_{1n}\,| > \varepsilon$, for otherwise θ_{1n} is already a desired estimator. By choosing appropriately ε and n, we can render the probability of the event $\{|\, g_{1n}\,| > \varepsilon\}$ close to 1, as well as the probability that $\beta_{1n} = |\, g_{1n}\,|^2 \,/\, |\, g_{0n}\,|^2$ is positive. As g_{1n} is stochastically bounded (this is verified similarly to g_{0n} by taking into consideration assumptions (i) and

(ii) on θ_{0n} and G_{0n} and conditions I and IV in section 3.3), then β_{1n} is stochastically bounded too. Therefore $p_{1n} = - g_{1n} + \beta_{1n} p_{0n}$ and $\alpha_{1n} = - (g_{1n}, p_{1n}) / \lambda_{1n}$ satisfy the conditions of lemma 3.5.2 and θ_{2n} is a δ_n-consistent estimator of θ.

Applying repeatedly the above arguments we can verify step by step the conditions of lemma 3.5.2. This leads to the following

Statement 3.8.1. *Let a criterion function* $F_n (X_n, \theta)$ *be asymptotically quadratic in the sense of definition* 3.3.1. *Suppose we are given a* δ_n*-consistent estimator for* θ *used as the initial point* θ_{0n} *and a consistent estimator* G_{0n} *for* $G_n (\theta)$ (cf. conditions (i) and (ii) in section 3.4). *Then the iterative procedure* (3.4.3) *of estimating the parameter* θ *with steplengths* α_{kn} *and search vectors* p_{kn} *given by* (3.8.4) *and* (3.8.1) *respectively, is well defined and asymptotically descent in the sense of definitions* 3.4.2 *and* 3.5.1 *respectively.*

By (3.5.5) and (3.8.3) we have

(3.8.5) $$P_n (| \, | g_{kn} |^2 - (\Delta g_{kn}, p_{kn}) | > \varepsilon) \to 0 \text{ as } n \to \infty,$$

for all $P_n \in \vartheta^{-1}(\theta)$. Besides, taking into consideration relationship (3.8.9) which is verified below in the course of proving theorem 3.8.2, we have for all $P_n \in \vartheta^{-1}(\theta)$

(3.8.6) $$P_n (| \, | g_{kn} |^2 - (\Delta g_{k-1n}, g_{kn}) | > \varepsilon) \to 0 \text{ as } n \to \infty.$$

Due to (3.8.5) and (3.8.6) we get two asymptotically equivalent alternatives to (3.8.2):

(3.8.7) $$\beta_{kn} = (\Delta g_{k-1n}, g_{kn}) / (\Delta g_{k-1n}, p_{k-1n}) \text{ and } \beta_{kn} = (\Delta g_{k-1n}, g_{kn}) / | g_{k-1n} |^2.$$

Note that due to (3.5.8) we may consider yet another asymptotically equivalent alternative $\beta_{kn} = (G_{0n} p_{k-1n}, g_{kn}) / \lambda_{k-1n}$ and hence premultiplying p_{kn} in (3.8.1) by $p_{k-1n}{}^T G_{0n}$ we get

(3.8.8) $$G_{0n} p_{k-1n} \perp^{(P_n)} p_{kn}$$

irrespectively which of the indicated versions of β_{kn} were used.

Theorem 3.8.2. *Under the same circumstances as indicated in statement* 3.8.1 *the stochastic modifications of conjugate gradient methods described by iterates* (3.8.1) *together with one of the formulas for* β_{kn} *given by* (3.8.2) *or* (3.8.7), *are well defined and terminated in fewer then* d *iterates, say* $r \le d$, *and* $\theta_{rn} \in D$.

Proof. Statement 3.8.1 tells us that the iterates are well defined. Therefore we can use proposition 3.7.4 to draw the desired conclusion, provided the asymptotic mutual conjugacy is verified of the present search vectors p_{kn} (with respect to $G_n (\theta)$ or, equivalently, to G_{0n}; cf. remark 3.7.2). To this end we introduce the induction hypothesis

(3.8.9) $$G_{0n} p_{kn} \perp^{(P_n)} p_{jn} \text{ and } g_{kn} \perp^{(P_n)} g_{jn}, \; j = 0, ..., k - 1 \text{ for some } k \le r;$$

cf. (2.5.6). For k = 1 the relations (3.8.9) clearly hold by (3.8.8) and (3.5.5) respectively. By definition (3.8.1) we have $(g_{kn} + \alpha_{kn} G_{0n} p_{kn}, g_{jn}) = (g_{kn}, g_{jn}) + \alpha_{kn} (G_{0n} p_{kn}, - p_{jn} + \beta_{jn} p_{j-1n})$. Therefore (3.8.9) implies

(3.8.10) $$(g_{kn} + \alpha_{kn} G_{0n} p_{kn}) \perp^{(P_n)} g_{jn} \quad j < k.$$

By (3.5.6) and (3.8.10) we have $g_{k+1n} \perp^{(P_n)} g_{jn}$ for $j < k$, and hence the second relationship in (3.8.9) holds for k + 1, since it is easily seen that $g_{k+1n} \perp^{(P_n)} g_{kn}$. Indeed, by definition (3.8.1) we have $(g_{k+1n}, g_{kn}) = (g_{k+1n}, - p_{kn} + \beta_{kn}\, p_{k-1n})$, so that in view of (3.5.5) it suffices to show that $g_{k+1n} \perp^{(P_n)} p_{k-1n}$ or, by (3.5.7), that $(g_{kn} + \alpha_{kn}\, G_{0n}\, p_{kn}) \perp^{(P_n)} p_{k-1n}$. But the last relationship is obvious by (3.5.5) and (3.8.9).

Regarding the first of relationships (3.8.9), it suffices to show that $G_{0n}\, p_{k+1n} \perp^{(P_n)} p_{jn}$ for $j < k$, because for $j = k$ we already have (3.8.8). By definition (3.8.1) this is equivalent to $G_{0n}\,(- g_{k+1n} + \beta_{k+1n}\, p_{kn}) \perp^{(P_n)} p_{jn}$ for $j < k$. By (3.8.9) we only need to verify that $g_{k+1n} \perp^{(P_n)} G_{0n}\, p_{jn}$ for $j < k$. In view of (3.5.6) we can replace $G_{0n}\, p_{jn}$ in the last relationship by $\Delta g_{jn} / \alpha_{jn}$, $j = 0, \ldots, k - 1$ (note that this is allowed since the procedure is well defined, therefore all $\alpha_{0n}, \ldots, \alpha_{k-1n}$ possess the property (3.5.4)) to see that it holds due to the following relationship valid in virtue of (3.8.9): $g_{k+1n} \perp^{(P_n)} \Delta g_{jn}$ for $j < k$. The induction now is complete and mutual asymptotic conjugacy of the present search vectors p_{kn} with respect to G_{0n} is proved.

To complete the proof assume that $r < d$ and p_{rn} converges in norm to zero as $n \to \infty$ in probability $P_n \in \vartheta^{-1}(\theta)$. Then g_{rn} also possesses this property. Indeed, by (3.8.1) and (3.5.5) the convergence $P_n(|\,p_{rn}|^2 > \varepsilon) \to 0$ assumed above implies $P_n(|\,g_{rn}|^2 + |\,p_{r-1n}|^2 > \varepsilon) \to 0$, and this in turn implies $P_n(|\,g_{rn}|^2 > \varepsilon) \to 0$. Thus $\theta_{rn} \in D$ is the desired estimator. On the other hand, if $r = d$, then $p_{0n}, \ldots, p_{d-1n}$ are asymptotically linearly independent in the sense indicated in lemma 3.7.3. In exactly the same way as at the end of proving proposition 3.7.4, this property of search directions implies (3.7.3). ◊

3.9. Quasi-Newton methods. As before assume an asymptotically quadratic criterion function $F_n(X_n, \theta)$ is given, as well as an initial point θ_{0n} and a "consistent estimator" G_{0n} to start with iterations (3.4.3) where steplengths α_{kn} and search vectors p_{kn} are defined similarly to (2.6.1) and (2.6.8) respectively:

$$(3.9.1) \qquad p_{kn} = - H_{kn}\, g_{kn} \text{ and } \alpha_{kn} = (H_{kn}\, g_{kn}, g_{kn}) / \lambda_{kn}$$

with λ_{kn} given by (3.5.2). As for updating formulas, starting with any symmetric positive definite matrix $H_0 = H_{0n}$, independent of n, we consider here stochastic modifications of quasi-Newton methods defined in section 2.6 via Broyden's family of updating formulas (2.6.6) which are modified by adding to all entries the index n and substituting Δx_k by $\Delta\theta_{kn}$. We restrict our attention to the most important particular cases - DFP and BFGS formulas which correspond to chosing the parameter π_k equal to 0 and 1 respectively. This leads to

$$(3.9.2) \qquad H_{k+1n} = H_{kn} + Q_{kn}^{(1)} - (H_{kn}\, \Delta g_{kn}, \Delta g_{kn})^{-1}\, H_{kn}\, \Delta g_{kn}\, \Delta g_{kn}^{T}\, H_{kn}$$

and

$$H_{k+1n} = Q_{kn}^{(1)} + W_{kn}\, H_{kn}\, W_{kn}^{T} \text{ with } W_{kn} = I - (\Delta\theta_{kn}, \Delta g_{kn})^{-1}\, \Delta\theta_{kn}\, \Delta g_{kn}^{T}$$

respectively; cf. (2.6.5) and (2.6.7).

The following quasi-Newton property of the updates

(3.9.3) $$\Delta\theta_{kn} = H_{k+1n}\,\Delta g_{kn}$$

is verified as in section 2 (cf. (2.6.3) and remark 2.6.2) by verifying that $Q_{kn}^{(1)}\,\Delta g_{kn} = \Delta\theta_{kn}$ and $Q_{kn}^{(2)}\,\Delta g_{kn} = -\,H_{kn}\,\Delta g_{kn}$.

Next, the above updating formulas involve in denominators the expressions

(3.9.4) $$(H_{kn}\,\Delta g_{kn}, \Delta g_{kn}) \text{ and } (\Delta\theta_{kn}, \Delta g_{kn})$$

(note, by the way, that due to (3.9.3) the second of these expressions is of the same type as the first but with H_{k+1n} instead of H_{kn}). In order to guarantee that the first of these expressions is well defined we need to show that the sequence of H's constructed step by step according to the above formulas preserve the symmetry and positive definiteness of the initial matrix H_0. In view of the equality $(H_{kn}\,g_{kn}, g_{kn}) = -\,(p_{kn}, g_{kn})$ and (3.5.1), the last fact entails the asymptotic descency of the iterates. Besides, a steplength chosen according to (3.9.1) provides for exact linear search as (3.5.5) is satisfied; see lemma 3.5.2. Therefore we can verify, similarly to remark 3.5.4 (i), that the second of expressions (3.9.4) is also well defined; cf. (3.5.9). As in section 2.6 we will deal with the simplest DFP updates (3.9.2) only, for the arguments used are in fact typical.

Assume the basic conditions stipulated in section 3.3. In the beginning the situation is similar to the previous section: suppose $|\,g_{0n}\,| > \varepsilon$ for some tolerance value ε, for otherwise θ_{0n} is already a desired estimator. Since a tolerance value ε is chosen small enough and a sample size n large enough to render the probability of the event $\{|\,g_{0n}\,| > \varepsilon\}$ close to 1, then by choosing H_0 positive definite we get the asymptotically descent direction p_{0n} defined by (3.9.1) with $k = 0$, as $(H_{0n}\,g_{0n}, g_{0n}) = -\,(p_{0n}, g_{0n})$ does satisfy (3.5.1). As for the steplength α_{0n}, it is well defined in the sense of (3.5.4) and stochastically bounded, for g_{0n} is stochastically bounded (this is verified in the same way as in the previous section), as well as p_{0n}. These considerations allow us to apply lemma 3.5.2 to conclude that (i) θ_{1n} is a δ_n-consistent estimator of θ, like θ_{0n}, and (ii) the relationship (3.5.5) holds for $k = 0$. According to remark 3.5.4 (i) the last fact and asymptotic descency of the direction p_{0n} means that (3.5.8) holds for $k = 0$.

Next, suppose $|\,g_{1n}\,| > \varepsilon$ for some tolerance value ε, for otherwise θ_{1n} is already a desired estimator, and verify that g_{1n} is stochastically bounded since θ_{1n} is a δ_n-consistent estimator (in exactly the same manner as was done for g_{0n}). As was said at the end of the previous paragraph the relationship (3.5.8) holds for $k = 0$; hence as a sample size n is large enough, the probability that the second of the expressions (3.9.4) is strongly positive for $k = 0$ is close to one. Since H_0 is positive definite, the same claim is true concerning the first of these expressions. These considerations, together with those used in the course of proving lemma 2.6.1, lead to the following conclusion: H_{1n} constructed by means of the DFP formula (3.9.2) is positive definite with probability close to one. Since

$$\|\,H_{k+1n}\,\| = \|\,H_{kn}\,\| + |\,\Delta\theta_{kn}\,|^2\,(\Delta\theta_{kn}, \Delta g_{kn})^{-1} + |\,H_{kn}\,\Delta g_{kn}\,|^2\,(H_{kn}\,\Delta g_{kn}, \Delta g_{kn})^{-1}$$

for the DFP updates (3.9.2), then not only g_{1n}, but also $p_{1n} = -\,H_{1n}\,g_{1n}$ is stochastically bounded. As for the steplength α_{1n} given by (3.9.1) with $k = 1$, the above claims allow us to verify that it is well defined in the sense of (3.5.4) and stochastically bounded. Thus we can apply lemma 3.5.2 to conclude that (i) θ_{2n} is a δ_n-consistent estimator of θ, like θ_{0n} and θ_{1n}, and (ii) the relationship (3.5.5) holds for $k = 1$.

Applying repeatedly the above arguments we can verify step by step the conditions of lemma 3.5.2. This leads to the following

Statement 3.9.1. *Let a criterion function* $F_n(X_n, \theta)$ *be asymptotically quadratic in the sense of definition* 3.3.1. *Suppose we are given a* δ_n*-consistent estimator of* θ *used as the initial point* θ_{0n} *and a consistent estimator* G_{0n} *of* $G_n(\theta)$ (cf. conditions (i) and (ii) in section 3.4). *Then the iterative procedure* (3.4.3) *of estimating the parameter* θ, *where steplengths* α_{kn} *and search vectors* p_{kn} *are given by* (3.9.1), *is well defined and asymptotically descent in the sense of definitions* 3.4.2 *and* 3.5.1 *respectively.*

Based on this statement we will prove now the main result of this section.

Theorem 3.9.2. *Under the same circumstances as indicated in statement* 3.9.1 *the stochastic modifications defined above of quasi-Newton methods, are well defined and terminated in fewer then* d *iterates, say* $r \leq d$, *and* $\theta_{rn} \in D$.

Proof. Statement 3.9.1 allows us to use the same arguments as in the course of proving theorem 3.8.2 which reduce the problem to verifying asymptotic conjugacy of the present search directions in order to draw the desired conclusion by applying proposition 3.7.4. To avoid unnecessary repetitions, we skip here these details and provide only for the proof of asymptotic conjugacy. Namely, we prove by induction that

(i) the relationship (3.7.4) holds, i.e. the present directions p_{kn} are asymptotically conjugate in the sense of definition 3.7.1, and

(ii) they possess the following property called *asymptotic heredity*: for each $\theta \in \Theta$ and $\varepsilon > 0$

(3.9.5) $$P_n(|\Delta\theta_{jn} - H_{kn}\Delta g_{jn}| > \varepsilon) \to 0 \quad \text{as } n \to \infty \quad j < k$$

with $P_n \in \vartheta^{-1}(\theta)$.

Assume first $k = 1$. Then (3.9.5) holds due to the quasi-Newton condition (3.9.3). To get (3.7.4) with $k = 1$ note first that by definition (3.9.1) and quasi-Newton property (3.9.3) we have $(\Delta g_{0n}, p_{1n}) = -(H_{1n}\Delta g_{0n}, g_{1n}) = -(\Delta\theta_{0n}, g_{1n})$. Hence the consequence (3.5.8) of exact linear search yields (3.7.4) with $k = 1$.

Assume (3.7.4) and (3.9.5) hold for some $k < d$. We handle first conjugacy proving (3.7.4) with k substituted by $k + 1$. Since by definition (3.9.1) we have

$$(\Delta g_{jn}, p_{k+1n}) = -(H_{k+1n}\Delta g_{jn}, g_{k+1n}) = -([H_{kn} + Q_{kn}^{(1)} + Q_{kn}^{(2)}]\Delta g_{jn}, g_{k+1n}),$$

and by the consequence (3.5.10) of exact linear search we have $Q_{kn}^{(1)}\Delta g_{jn} \perp^{(P_n)} g_{k+1n}$ with

(3.9.6) $$Q_{kn}^{(1)}\Delta g_{jn} = (\Delta\theta_{kn}, \Delta g_{kn})^{-1}(\Delta\theta_{kn}, \Delta g_{jn})\Delta\theta_{kn},$$

then it suffices to prove that

(3.9.7) $$[H_{kn} + Q_{kn}^{(2)}]\Delta g_{jn} \perp^{(P_n)} g_{k+1n} \quad j \leq k.$$

Observe that since

(3.9.8) $$Q_{kn}^{(2)}\Delta g_{jn} = -(H_{kn}\Delta g_{kn}, \Delta g_{kn})^{-1}(H_{kn}\Delta g_{kn}, \Delta g_{jn})H_{kn}\Delta g_{kn},$$

the left hand side of (3.9.7) equals to zero in case $j = k$.

In case $j < k$ it is not hard to verify that

(3.9.9) $$H_{kn}\, \Delta g_{jn} \perp^{(P_n)} g_{k+1n}$$

by applying first asymptotic heredity (3.9.5), then (3.4.4) and finally (3.7.2) and (3.7.4). By the same arguments we have $H_{kn}\, \Delta g_{jn} \perp^{(P_n)} \Delta g_{kn}$ for $j < k$, which in view of (3.9.8) shows that $Q_{kn}{}^{(2)}\, \Delta g_{jn}$ tends to zero in probability. This fact and (3.9.9) yield (3.9.7). Thus the induction concerning conjugacy is complete.

As for asymptotic heredity, substitute k in (3.9.5) by k+1 and look at

(3.9.10) $$H_{k+1n}\, \Delta g_{jn} = [H_{kn} + Q_{kn}{}^{(1)} + Q_{kn}{}^{(2)}]\, \Delta g_{jn}.$$

For $k = j$ the result is clear, as we have already verified the quasi-Newton property (3.9.3) which shows that in this case $H_{k+1n}\, \Delta g_{kn}$ coincides with $\Delta\theta_{kn}$. For $j < k$ we use the induction hypothesis which tells us that it suffices to show that the second and third terms on the right hand side of (3.9.10), given by (3.9.6) and (3.9.8) respectively, tend to zero in probability. Regarding the third term this is already shown in the course of proving (3.9.7). As for the second term, verify the claim by applying (3.7.4) to (3.9.6) by taking into consideration (3.4.4). Thus the proof of the asymptotic heredity property (3.9.5) is complete. ◊

We assume below that the quasi-Newton procedures of estimating the parameter θ treated in the present section are terminated in exactly d iterations, i.e. $\theta_{kn} \in \mathbb{D}$ for $k \leq d$, but $\theta_{kn} \notin D$ for $k < d$ while $\theta_{dn} \in D$. In view of (3.4.5) the asymptotic heredity (3.9.5) means that

(3.9.11) $$P_n\, (|\, [I - H_{dn}\, G_n\, (\theta)]\, \Delta\theta_{kn}\, | > \varepsilon) \to 0 \quad k < d.$$

Furthermore, all the stochastically bounded $\alpha_{0n}, \ldots, \alpha_{d-1n}$ satisfy (3.5.4), and in virtue of (3.4.4), definitions of Π_n and Λ_n in section 3.7 and remark 3.7.2, the relationship (3.9.11) and condition II in section 3.3 yield $|[G_n\, (\theta)^{-1} - H_{dn}\,]\, \Lambda_n\, | \to 0$ for $k < d$. Since Λ_n is asymptotically diagonal with the stochastically bounded from below and above diagonal entries (3.5.2) , we also have

(3.9.12) $$P_n\, (|\, G_n\, (\theta)^{-1} - H_{dn}\, | > \varepsilon) \to 0 \quad k < d,$$

that is, H_{dn} is a "consistent estimator" of $G_n\, (\theta)^{-1}$ in the same sense as in section 3.4, condition (ii). This observation implies the following corollary to proposition 2.6.3.

Corollary 3.9.3. *Assume that under the circumstances of theorem* 3.9.2 *a quasi-Newton procedure of estimating the parameter* θ *is terminated in exactly* d *iterates. Then* H_{dn} *is a consistent estimator of* $G_n\, (\theta)^{-1}$. *Moreover,* $Q_{0n}{}^{(1)} + \ldots + Q_{d-1n}{}^{(1)}$ *and* $Q_0{}^{(2)} + \ldots + Q_{d-1}{}^{(2)}$ *are consistent estimators of* $G_n\, (\theta)^{-1}$ *and* $- H_0$ *respectively.*

Proof. We have already shown (3.9.12). Since $H_{dn} = H_0 + (Q_{0n}{}^{(1)} + \ldots + Q_{d-1n}{}^{(1)}) + (Q_{0n}{}^{(2)} + \ldots + Q_{d-1n}{}^{(2)})$, it suffices to prove the consistency of the second term as an estimator for $G_n\, (\theta)^{-1}$. In view of (3.4.4), conjugacy (3.7.1) means that $G_n\, (\theta)\, \Delta\theta_{kn} \perp^{(P_n)} \Delta\theta_{jn}$ for $k \neq j$. Hence by definition of $Q_{kn}{}^{(1)}$ we have that $P_n\, (|\, [I - (Q_{0n}{}^{(1)} + \ldots + Q_{d-1n}{}^{(1)})\, G_n\, (\theta)]\, \Delta\theta_{kn}\, | > \varepsilon) \to 0$ for $k < d$. To complete the proof of consistency of $Q_{0n}{}^{(1)} + \ldots +$

$Q_{d\text{-}1n}^{(1)}$, compare the last relationship and (3.9.11) and use the same arguments as above. ◊

References

I.V. Basawa and B.L.S. Pracasa Rao (1980). *Statistical Inference for Stochastic Processes.* London: Academic Press.

I.V. Basawa and D.J. Scott (1983). *Asymptotically Optimal Inference for Non-ergodic Models.* Lecture Notes in Statistics 17. New York: Springer

G. Beinicke and K. Dzhaparidze (1982). On parameter estimation by the Davidon-Fletcher-Powell method. *Theory Probab. Appl. 27,* 396-402.

K.W. Brodley (1977). Unconstrained minimization, in *The State of the Art in Numerical Analysis* (D. Jacobs, ed.), 229-268. London: Academic Press.

F. Campillo and F. Le Grand (1989). MLE for partially observed diffusions: direct maximization vs. the EM algorithm. *Stochastic Processes and their Applications 33,* 245-274.

G. Celex and J. Diebolt (1990). A simulated annealing type EM algorithm. *Rapports de Recherche,* INRIA, Centre de Rocquencourt.

L.C.W. Dixon (1972). Quasi-Newton algorithms generate identical points. *Math. Prog. 2,* 383-387.

K. Dzhaparidze (1983). On iterative procedures of asymptotic inference. *Statistica Neerlandica 37,* 181-189.

K. Dzhaparidze (1986). *Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series.* New York: Springer.

K. Dzhaparidze and A.M. Yaglom (1983). Spectrum parameter estimation in time series analysis, in *Developments in Statistics, vol. 4* (P.R. Krishnaiah, ed.), 1-96. New York: Academic Press.

R.A. Fisher (1925). Theory of statistical estimation. *Proc. Cambridge Philos. Soc. 22,* 700-725.

I.A. Ibragimov and R. Z. Has'minskii (1981). *Statistical Estimation, Asymptotic Theory.* New York: Springer.

R.I. Jennrich (1969). Asymptotic properties of non-linear leas squares estimators. *Ann. Mathem. Statist. 40,* 633-643.

R. Kohn (1978). Asymptotic properties of time domein Gaussian estimators. *Adv. Appl. Probab. 10,* 339-359.

L. LeCam (1956). On the asymptotic theory of estimation and testing hypotheses, in *Proceedings in the third Berkeley Symposium on Mathematical Statistics and Probability* (J. Neyman, ed.). Berkeley: California Univ. Press.

L. LeCam (1960). Locally asymptotically normal families of distributions. *Univ. California Publ. in Statist., vol. 3,* no. 2, 129-156. Berkeley: California Univ. Press.

L. LeCam (1969). *Theorie asymptotique de la decision statistique.* Monreal: Presses Univ. Monreal.

L. LeCam (1974). *Notes on Asymptotic Methods in Statistical Decision Theory.* Centre de

Recherches Mathematiques, Monreal: Presses Univ. Monreal.

I. Meilijson (1989). A fast improvement to the EM algorithm on its own terms. *J. R. Statist. Soc. B 51,* 127-138.

J.M. Ortega and W.C. Rheinboldt (1970). *Iterative Solution of Nonlinear Equations in Several Variables.* New York: Academic Press.

H.C.H. Paardekooper, H.B.A. Steens and G. van der Hoek (1989). A note on properties of iterative procedures of asymptotic inference. *Statistica Neerlandica 43,* 245-253.

P.M. Robinson (1988). The stochastic difference between econometric statistics. *Econometrica 56,* 531-548.

L.E. Scales (1985). *Introduction to Non-Linear Optimization.* London: MacMillan Publ. LTD.

A. Sieders and K. Dzhaparidze (1987). A large deviation result for parameter estimators and its application to nonlinear regression analysis. *Ann. Statist.* 15, 1031-1049.