March 1994 LIDS-P-2237

INCREMENTAL LEAST SQUARES METHODS[1] AND THE EXTENDED KALMAN FILTER

by

Dimitri P. Bertsekas[2]

Abstract

In this paper we propose and analyze nonlinear least squares methods, which process the data incrementally, one data block at a time. Such methods are well suited for large data sets and real time operation, and have received much attention in the context of neural network training problems. We focus on the Extended Kalman Filter, which may be viewed as an incremental version of the Gauss-Newton method. We provide a nonstochastic analysis of its convergence properties, and we discuss variants aimed at accelerating its convergence.

[1] Research supported by NSF under Grant 9300494-DMI.

[2] Department of Electrical Engineering and Computer Science, M.I.T., Cambridge, Mass., 02139.

1. INTRODUCTION

We consider least squares problems of the form

$$\begin{aligned} &\text{minimize} \quad f(x) = \|g(x)\|^2 = \sum_{i=1}^{m} \|g_i(x)\|^2 \\ &\text{subject to} \quad x \in \Re^n, \end{aligned} \tag{1}$$

where g is a continuously differentiable function with component functions $g_1, \ldots, g_m$, where $g_i : \Re^n \to \Re^{r_i}$. Here we write $\|z\|$ for the usual Euclidean norm of a vector z, that is, $\|z\| = \sqrt{z'z}$, where prime denotes transposition. We also write ∇g_i for the $n \times r_i$ gradient matrix of g_i, and ∇g for the $n \times (r_1 + \cdots + r_m)$ gradient matrix of g. Each component g_i is referred to as a *data block*, and the entire function $g = (g_1, \ldots, g_m)$ is referred to as the *data set*.

One of the most common iterative methods for solving this problem is the Gauss-Newton method, given by

$$x^{k+1} = x^k - \alpha^k \big(\nabla g(x^k) \nabla g(x^k)'\big)^{-1} \nabla g(x^k) g(x^k), \tag{2}$$

where α^k is a positive stepsize, and we assume that the $n \times n$ matrix $\nabla g(x^k) \nabla g(x^k)'$ is invertible. The case $\alpha^k = 1$ corresponds to the pure form of the method, where x^{k+1} is obtained by linearizing g at the current iterate x^k, and minimizing the norm of the linearized function, that is,

$$x^{k+1} = \arg\min_{x \in \Re^n} \|g(x^k) + \nabla g(x^k)'(x - x^k)\|^2, \qquad \text{if } \alpha^k = 1. \tag{3}$$

In problems where there are many data blocks, the Gauss-Newton method may be ineffective, because the size of the data set makes each iteration very costly. For such problems it may be better to use an *incremental method* that does not wait to process the entire data set before updating x; instead, the method cycles through the data blocks in sequence and updates the estimate of x after each data block is processed. A further advantage is that estimates of x become available as data is accumulated, making the approach suitable for real-time operation. Such methods include the Widrow-Hoff LMS algorithm [WiH60], [WiS85] for the case where the data blocks are linear, and other steepest-descent like methods for nonlinear data blocks that have been used extensively for the training of neural networks under the generic name of *backpropagation methods*. A cycle through the data set of a typical example of such a method starts with a vector x^k and generates x^{k+1} according to

$$x^{k+1} = \psi_m,$$

where ψ_m is obtained at the last step of the recursion

$$\psi_i = \psi_{i-1} - \alpha^k \nabla g_i(\psi_{i-1}) g_i(\psi_{i-1}), \qquad i = 1, \ldots, m,$$

α^k is a positive stepsize, and $\psi_0 = x^k$. Backpropagation methods are often effective, and they are supported by stochastic [PoT73], [Lju77], [KuC78], [Pol87], [BeT89], [Whi89a], [Gai93], as well as deterministic convergence analyses [Luo91], [Gri93], [LuT93], [MaS93], [Man93]. There are also parallel asynchronous versions of backpropagation methods, and corresponding stochastic [Tsi84], [TBA86], [BeT89], [Gai93], as well as deterministic convergence results [Tsi84], [TBA86], [BeT89], [MaS93]. However, backpropagation methods typically have a slow convergence rate not only because they are first order steepest descent-like methods, but also because they require a diminishing stepsize $\alpha^k = O(1/k)$ for convergence. If α^k is instead taken to be a small constant, an oscillation within each data cycle typically arises, as shown by [Luo91].

In this paper we focus on methods that combine the advantages of backpropagation methods for large data sets with the often superior convergence rate of the Gauss-Newton method. We thus consider an incremental version of the Gauss-Newton method, which operates in cycles through the data blocks. The $(k+1)$st cycle starts with a vector x^k and a positive semidefinite matrix H^k to be defined later, then updates x via a Gauss-Newton-like iteration aimed at minimizing

$$\lambda(x - x^k)'H^k(x - x^k) + \|g_1(x)\|^2,$$

where λ is a scalar with

$$0 < \lambda \leq 1,$$

then updates x via a Gauss-Newton-like iteration aimed at minimizing

$$\lambda^2(x - x^k)'H^k(x - x^k) + \lambda\|g_1(x)\|^2 + \|g_2(x)\|^2,$$

and similarly continues, with the ith step consisting of a Gauss-Newton-like iteration aimed at minimizing the weighted partial sum

$$\lambda^i(x - x^k)'H^k(x - x^k) + \sum_{j=1}^{i} \lambda^{i-j}\|g_j(x)\|^2.$$

In particular, given x^k, the $(k+1)$st cycle sequentially generates the vectors

$$\psi_i = \arg\min_{x \in \Re^n} \left\{ \lambda^i(x - x^k)'H^k(x - x^k) + \sum_{j=1}^{i} \lambda^{i-j}\|\tilde{g}_j(x, \psi_{j-1})\|^2 \right\}, \qquad i = 1, \ldots, m, \tag{4}$$

and sets

$$x^{k+1} = \psi_m, \tag{5}$$

where $\tilde{g}_j(x, \psi_{j-1})$ are the linearized functions

$$\tilde{g}_j(x, \psi_{j-1}) = g_j(\psi_{j-1}) + \nabla g_j(\psi_{j-1})'(x - \psi_{j-1}), \tag{6}$$

and ψ_0 is the estimate of x at the end of the kth cycle

$$x^k = \psi_0. \tag{7}$$

As will be seen later, the quadratic minimizations above can be efficiently implemented.

The most common version of the preceding algorithm is obtained when the matrices H^k are updated by the recursion

$$H^{k+1} = \lambda^m H^k + \sum_{j=1}^{m} \lambda^{m-j} \nabla g_j(\psi_{j-1}) \nabla g_j(\psi_{j-1})'. \tag{8}$$

Then for $\lambda = 1$ and $H^0 = 0$, the method becomes the well-known Extended Kalman Filter (EKF for short) specialized to the case where the state of the underlying dynamical system stays constant and the measurement equation is nonlinear. The matrix H^k has the meaning of the inverse of an approximate error covariance of the estimate x^k. In the case $\lambda < 1$, the effect of old data blocks is discounted, and successive estimates produced by the method tend to change more rapidly. In this way one may obtain a faster rate of progress of the method, and this is the main motivation for considering $\lambda < 1$.

The EKF has been used extensively in a variety of control and estimation applications (see e.g. [AWB68], [Jaz70], [Meh71], [THS77], [AnM79], [WeM80]), and has also been suggested for the training of neural networks (see e.g. [WaT90] and [RRK92]). The version of the algorithm (4)-(8) with $\lambda < 1$ has also been proposed by Davidon [Dav76] who, unaware of the earlier work in the control and estimation literature, described the qualitative behavior of the method together with favorable computational experience, but gave no convergence analysis. The first convergence analysis of the EKF was given by Ljung [Lju79], who assuming $\lambda = 1$, used a stochastic formulation and the ODE approach of [Lju77] to prove satisfactory convergence properties for a version of the EKF that is closely related to the one considered here (Theorem 6.1 of [Lju79], which assumes a stationary measurement equation and additive noise). Ljung also showed that the EKF, when applied to more complex models where the underlying dynamic system is linear but its dynamics depend on x, exhibits complex behavior, including the possible convergence to biased estimates. For such models he suggested the use of a different formulation of the least squares problem involving the innovations process (see also [Urs80]). The algorithms and analysis of the present paper apply to any type of deterministic least squares problem, and thus also apply to Ljung's innovations-based formulation.

A deterministic analysis of the EKF method (4)-(8) where $\lambda < 1$ was given in the MS thesis by Pappas [Pap82], written under the author's supervision. Pappas considered only the special case where $\min_x \|g(x)\|^2 = 0$, and showed that the EKF converges locally to a nonsingular solution of the system $g(x) = 0$ at a rate which is linear with convergence ratio λ^m. He also argued by example that

when $\lambda < 1$ and $\min_x \|g(x)\|^2 > 0$, the iterates ψ_i produced by the EKF within each cycle generally oscillate with a "size" of oscillation that diminishes as λ approaches 1.

The purpose of this paper is to provide a deterministic analysis of the convergence properties of the EKF. Our analysis is complicated by the lack of an explicit stepsize in the algorithm. In the case where $\lambda = 1$ we show that the limit points of the generated sequence $\{x^k\}$ by the EKF are stationary points of the least squares problem. To improve the rate of convergence of the method, which is sublinear and typically slow, we suggest a convergent and empirically faster variant where λ is initially less than 1 and is progressively increased towards 1.

One nice aspect of the deterministic analysis is that it decouples the stochastic modeling of the data generation process from the algorithmic solution of the least squares problem. Otherwise expressed, the EKF discussed here will typically find a least squares solution even if the least squares formulation is inappropriate for the real parameter estimation problem. This is a valuable insight because it is sometimes thought that convergence of the EKF depends on the validity of the underlying stochastic model assumptions.

2. THE EXTENDED KALMAN FILTER

When the data blocks are linear functions, it takes a single pure Gauss-Newton iteration to find the least squares estimate. This iteration can be implemented as an incremental algorithm, the *Kalman filter*, which we now describe. Assume that the functions g_i are linear of the form

$$g_i(x) = z_i - C_i x, \tag{9}$$

where $z_i \in \Re^{r_i}$ are given vectors and C_i are given $r_i \times n$ matrices. Let us consider the incremental method that generates the vectors

$$\psi_i = \arg\min_{x \in \Re^n} \sum_{j=1}^{i} \lambda^{i-j} \|z_j - C_j x\|^2, \qquad i = 1, \ldots, m. \tag{10}$$

Then the method can be recursively implemented as shown by the following well-known proposition:

Proposition 1. (*Kalman Filter*) Assuming that the matrix $C_1'C_1$ is positive definite, the least squares estimates

$$\psi_i = \arg\min_{x \in \Re^n} \sum_{j=1}^{i} \lambda^{i-j} \|z_j - C_j x\|^2, \qquad i = 1, \ldots, m,$$

can be generated by the algorithm

$$\psi_i = \psi_{i-1} + H_i^{-1} C_i'(z_i - C_i\psi_{i-1}), \qquad i = 1, \ldots, m, \tag{11}$$

where ψ_0 is an arbitrary vector, and the positive definite matrices H_i are generated by

$$H_i = \lambda H_{i-1} + C_i' C_i, \qquad i = 1, \ldots, m, \tag{12}$$

with

$$H_0 = 0.$$

More generally, for all $\bar{i} < i$ we have

$$\psi_i = \psi_{\bar{i}} + H_i^{-1} \sum_{j=\bar{i}+1}^{i} \lambda^{i-j} C_j'(z_j - C_j\psi_{\bar{i}}), \qquad i = 1, \ldots, m. \tag{13}$$

The proof is obtained by using the following lemma for the case of two data blocks, the straightforward proof of which is ommited:

Lemma 1. Let ζ_1, ζ_2 be given vectors, and Γ_1, Γ_2 be given matrices such that $\Gamma_1'\Gamma_1$ is positive definite. Then the vectors

$$\psi_1 = \arg\min_{x \in \Re^n} \|\zeta_1 - \Gamma_1 x\|^2, \tag{14}$$

and

$$\psi_2 = \arg\min_{x \in \Re^n} \left\{ \|\zeta_1 - \Gamma_1 x\|^2 + \|\zeta_2 - \Gamma_2 x\|^2 \right\}, \tag{15}$$

are also given by

$$\psi_1 = \psi_0 + (\Gamma_1'\Gamma_1)^{-1}\Gamma_1'(\zeta_1 - \Gamma_1\psi_0), \tag{16}$$

and

$$\psi_2 = \psi_1 + (\Gamma_1'\Gamma_1 + \Gamma_2'\Gamma_2)^{-1}\Gamma_2'(\zeta_2 - \Gamma_2\psi_1), \tag{17}$$

where ψ_0 is an arbitrary vector.

The proof of Eqs. (12) and (13) of Prop. 1 follows by applying Lemma 1 with the correspondences $\psi_0 \sim \psi_0$, $\psi_1 \sim \psi_{\bar{i}}$, $\psi_2 \sim \psi_i$, and

$$\zeta_1 \sim \begin{pmatrix} \sqrt{\lambda^{i-1}} z_1 \\ \vdots \\ \sqrt{\lambda^{i-\bar{i}}} z_{\bar{i}} \end{pmatrix}, \qquad \Gamma_1 \sim \begin{pmatrix} \sqrt{\lambda^{i-1}} C_1 \\ \vdots \\ \sqrt{\lambda^{i-\bar{i}}} C_{\bar{i}} \end{pmatrix},$$

$$\zeta_2 \sim \begin{pmatrix} \sqrt{\lambda^{i-\bar{i}-1}} z_{\bar{i}+1} \\ \vdots \\ z_i \end{pmatrix}, \qquad \Gamma_2 \sim \begin{pmatrix} \sqrt{\lambda^{i-\bar{i}-1}} C_{\bar{i}+1} \\ \vdots \\ C_i \end{pmatrix},$$

and by carrying out the straightforward algebra.

Note that the positive definiteness assumption on $C_1'C_1$ in Prop. 1 is needed to guarantee that the first matrix H_1 is positive definite and hence invertible; then the positive definiteness of the subsequent matrices $H_2, \ldots, H_m$ follows from Eq. (12). As a practical matter, it is possible to guarantee the positive definiteness of $C_1'C_1$ by lumping a sufficient number of measurements into the first data block (C_1 should contain n linearly independent columns). An alternative is to redefine ψ_i as

$$\psi_i = \arg\min_{x \in \Re^n} \left\{ \delta\lambda^i \|x - \psi_0\|^2 + \sum_{j=1}^{i} \lambda^{i-j} \|z_j - C_j x\|^2 \right\}, \qquad i = 1, \ldots, m,$$

where δ is a small positive scalar. Then it can be seen that ψ_i is generated by the same equations (11) and (12), except that the initial condition $H_0 = 0$ is replaced by

$$H_0 = \delta I,$$

so that $H_1 = \delta I + C_1'C_1$ is positive definite even if $C_1'C_1$ is not. Note, however, that in this case, the last estimate ψ_m is only approximately equal to the least squares estimate x^*, even if $\lambda = 1$ (the approximation error depends on the size of δ).

Consider now the general case where the data blocks g_i are nonlinear. Then the EKF can be used, and its first cycle can be implemented by means of the Kalman filter equations of Prop. 1. Using the formulas (11) and (12) with the identifications

$$z_i = g_i(\psi_{i-1}) - \nabla g_i(\psi_{i-1})'\psi_{i-1}, \qquad C_i = -\nabla g_i(\psi_{i-1})',$$

the kth cycle of the EKF can be written in the incremental form

$$\psi_i = \psi_{i-1} - H_i^{-1} \nabla g_i(\psi_{i-1}) g_i(\psi_{i-1}), \qquad i = 1, \ldots, m, \tag{21}$$

where the matrices H_i are generated by

$$H_i = \lambda H_{i-1} + \nabla g_i(\psi_{i-1}) \nabla g_i(\psi_{i-1})', \qquad i = 1, \ldots, m, \tag{22}$$

with

$$H_0 = 0. \tag{23}$$

To contrast the EKF with the pure form of the Gauss-Newton method (unit stepsize), note that a single iteration of the latter can be written as

$$x^{k+1} = \arg\min_{x \in \Re^n} \sum_{i=1}^{m} \|\tilde{g}_i(x, x^k)\|^2. \tag{24}$$

Using the formulas of Prop. 1 with the identifications

$$z_i = g_i(x^k) - \nabla g_i(x^k)'x^k, \qquad C_i = -\nabla g_i(x^k)',$$

we can generate x^{k+1} by an incremental algorithm as

$$x^{k+1} = \overline{\psi}_m,$$

where

$$\overline{\psi}_i = \overline{\psi}_{i-1} - \overline{H}_i^{-1}\nabla g_i(x^k)\big(g_i(x^k) + \nabla g_i(x^k)'(\overline{\psi}_{i-1} - x^k)\big), \qquad i = 1, \ldots, m, \tag{25}$$

$\overline{\psi}_0 = x^k$, and the matrices $\overline{H}_i$ are generated by

$$\overline{H}_i = \overline{H}_{i-1} + \nabla g_i(x^k)\nabla g_i(x^k)', \qquad i = 1, \ldots, m, \tag{26}$$

with

$$\overline{H}_0 = 0. \tag{27}$$

Thus, by comparing Eqs. (21)-(23) with Eqs. (25)-(27), we see that, if $\lambda = 1$, a cycle of the EKF through the data set differs from a pure Gauss-Newton iteration only in that the linearization of the data blocks g_i is done at the corresponding current estimates ψ_{i-1} rather than at the estimate x^k available at the start of the cycle.

3. CONVERGENCE OF THE EXTENDED KALMAN FILTER

We have considered so far a single cycle of the EKF. To obtain an algorithm that cycles through the data set multiple times, there are two basic approaches. The first approach is to reset the matrix H to some fixed matrix H_0 at the start of each cycle. Unfortunately, the convergence properties of the resulting algorithm are questionable, and one can construct examples where the method diverges, basically because the increments $\psi_i - \psi_{i-1}$ produced by the method [cf. Eq. (21)] may be too large. One may attempt to correct this behavior by selecting H_0 to be a sufficiently large multiple of the identity matrix, but this leads to large asymptotic convergence errors (biased estimates) as can be seen through simple examples where the data blocks are linear.

The second approach, which is followed in this paper, is to create a larger data set by concatenating multiple copies of the original data set, that is, by forming what we refer to as *the extended data set*

$$(g_1, g_2, \ldots, g_m, g_1, g_2, \ldots, g_m, g_1, g_2, \ldots). \tag{28}$$

The EKF is then applied to the extended data set and takes the form given in the introduction [Eqs. (4)-(8)]. The algorithm has the form

$$H_{km+i} = \lambda H_{km+i-1} + \nabla g_i(\psi_{km+i-1})\nabla g_i(\psi_{km+i-1})', \qquad i = 1, \ldots, m,$$

$$\psi_{km+i} = \psi_{km+i-1} - H_{km+i}^{-1}\nabla g_i(\psi_{km+i-1})g_i(\psi_{km+i-1}), \qquad i = 1, \ldots, m,$$

where $H_0 = 0$ and $\psi_0 = x^0$ is an arbitrary vector. Note that while in the above equations, λ is written as a constant, we will later consider the possibility of changing λ in the course of the algorithm. Also, we assume that the matrix $\nabla g_1(x^0)\nabla g_1(x^0)'$ is invertible, so that H_1^{-1} is well defined. However, it can be shown that the convergence result to be given shortly also holds when H_0 is any positive definite matrix, in which case the invertibility of $\nabla g_1(x^0)\nabla g_1(x^0)'$ is unnecessary.

We will show that when $\lambda = 1$, the EKF version just described typically converges to stationary points of the least squares problem. The basic reason is that the EKF asymptotically resembles a gradient method with diminishing stepsize of order $O(1/k)$. To get a sense of this, assume that the EKF is applied to the extended data set (28) with $\lambda = 1$. Let us denote by x^k the iterate at the end of the kth cycle through the data set, that is,

$$x^k = \psi_{km}, \qquad k = 1, 2, \ldots$$

Then by using Eq. (13) with $i = (k+1)m$ and $\bar{i} = km$, we obtain

$$x^{k+1} = x^k - H_{(k+1)m}^{-1}\left(\sum_{i=1}^{m}\nabla g_i(\psi_{km+i-1})g_i(\psi_{km+i-1})\right). \tag{29}$$

Now $H_{(k+1)m}$ grows roughly in proportion to $k+1$ because, by Eq. (12), we have

$$H_{(k+1)m} = \sum_{j=0}^{k}\sum_{i=1}^{m}\nabla g_i(\psi_{jm+i-1})\nabla g_i(\psi_{jm+i-1})'. \tag{30}$$

It is therefore reasonable to expect that the method tends to make slow progress when k is large, which means that the vectors ψ_{km+i-1} in Eq. (29) are roughly equal to x^k. Thus for large k, the sum in the right-hand side of Eq. (29) is roughly equal to the gradient $\nabla g(x^k)g(x^k)$, while from Eq. (30), $H_{(k+1)m}$ is roughly equal to $(k+1)(\nabla g(x^k)\nabla g(x^k)')$, where $g = (g_1, g_2, \ldots, g_m)$ is the original data set. It follows that for large k, the EKF iteration (29) can be written approximately as

$$x^{k+1} \approx x^k - \frac{1}{k+1}\left(\nabla g(x^k)\nabla g(x^k)'\right)^{-1}\nabla g(x^k)g(x^k), \tag{31}$$

that is, as an approximate Gauss-Newton iteration with diminishing stepsize. Thus, based on generic properties of gradient methods with diminishing stepsize (see e.g. [Pol87]), we can expect convergence to stationary points of the least squares problem, and a sublinear convergence rate.

When $\lambda < 1$, the matrix H_i^{-1} generated by the EKF recursion (22) will typically not diminish to zero, and $\{x^k\}$ may not converge to a stationary point of $\sum_{i=1}^{m}\lambda^{m-i}\|g_i(x)\|^2$. Furthermore, as the following example shows, the sequences $\{\psi_{km+i}\}$ produced by the EKF using Eq. (21), may converge to different limits for different i:

Example 1.

Consider the case where there are two data blocks, $g_1(x) = x - c_1$ and $g_2(x) = x - c_2$, where c_1 and c_2 are given scalars. Each cycle of the EKF consists of two steps. At the second step of the kth cycle, we minimize

$$\sum_{i=1}^{k} \left(\lambda^{2i-1}(x-c_1)^2 + \lambda^{2i-2}(x-c_2)^2\right),$$

which is equal to the following scalar multiple of $\lambda(x-c_1)^2 + (x-c_2)^2$,

$$(1 + \lambda^2 + \cdots + \lambda^{2k-2})\big(\lambda(x-c_1)^2 + (x-c_2)^2\big).$$

Thus at the second step, we obtain the minimizer of $\lambda(x-c_1)^2 + (x-c_2)^2$,

$$\psi_{2k} = \frac{\lambda c_1 + c_2}{\lambda + 1}.$$

At the first step of the kth cycle, we minimize

$$(x-c_1)^2 + \lambda \sum_{i=1}^{k-1} \left(\lambda^{2i-1}(x-c_1)^2 + \lambda^{2i-2}(x-c_2)^2\right),$$

which is equal to the following scalar multiple of $(x-c_1)^2 + \lambda(x-c_2)^2$

$$(1 + \lambda^2 + \cdots + \lambda^{2k-4})\big((x-c_1)^2 + \lambda(x-c_2)^2\big),$$

plus the diminishing term $\lambda^{2k-2}(x-c_1)^2$. Thus at the first step, we obtain approximately (for large k) the minimizer of $(x-c_1)^2 + \lambda(x-c_2)^2$,

$$\psi_{2k-1} \approx \frac{c_1 + \lambda c_2}{1 + \lambda}.$$

We see therefore that within each cycle, there is an oscillation around the minimizer $(c_1 + c_2)/2$ of $(x-c_1)^2 + (x-c_2)^2$. The size of the oscillation diminishes as λ approaches 1.

The preceding example suggests that each sequence $\{\psi_{km+i}\}$, where $i = 1, \ldots, m$, may converge to a stationary point of the function

$$f_i(x) = \sum_{j=1}^{m} \lambda^{m-j} \|g_{j+i}(x)\|^2, \qquad i = 1, \ldots, m,$$

where we use the definition

$$g_j(x) = g_{j \bmod(m)+1}(x) \qquad \text{if } j > m.$$

This is readily shown when the data blocks g_i are linear in view of the definition of ψ_{km+i} as the minimizer of

$$\sum_{j=1}^{km+i} \lambda^{km+i-j} \|g_j(x)\|^2,$$

which can also be written as

$$\sum_{j=1}^{i} \lambda^{km+i-j}\|g_j(x)\|^2 + \big(1+\lambda^2+\cdots+\lambda^{(k-1)m}\big) f_i(x).$$

Since the leftmost summation above vanishes as $k \to \infty$, ψ_{km+i} minimizes $f_i(x)$ asymptotically. In the case of nonlinear data blocks, a related but more complex analysis of the cyclic convergence behavior described above is possible, but this will not be attempted in this paper.

Generally, for a nonlinear least squares problem, the convergence rate tends to be faster when $\lambda < 1$ than when $\lambda = 1$, essentially because the implicit stepsize does not diminish to zero as in the case $\lambda = 1$. For this reason, a hybrid method that uses a different value of λ within each cycle may work best in practice. One may start with a relatively small λ to attain a fast initial rate of convergence, and then progressively increase λ towards 1 in order to attain high solution accuracy. The following proposition shows convergence for the case where λ tends to 1 at a sufficiently fast rate. The idea of the proof is to show that the method involves an implicit stepsize of order $O(1/k)$, and then to apply arguments similar to those used by Tsitsiklis [Tsi84], and Tsitsiklis, Bertsekas, and Athans [TBA86] in their analysis of asynchronous distributed gradient methods, and by Mangasarian and Solodov [MaS93] in their convergence proof of an asynchronous parallel backpropagation method.

Proposition 2. Assume that $\nabla g_i(x)$ has full rank for all x and $i = 1, \ldots, m$, and that for some $L > 0$, we have

$$\|\nabla g_i(x) g_i(x) - \nabla g_i(y) g_i(y)\| \le L\|x-y\|, \qquad \forall\, x, y \in \Re^n,\ i = 1, \ldots, m. \tag{32}$$

Assume also that there is a constant $c > 0$ such that the scalar λ used in the updating formula (22) within the kth cycle, call it $\lambda(k)$, satisfies

$$0 \le 1 - \big(\lambda(k)\big)^m \le \frac{c}{k}, \qquad \forall\, k = 1, 2, \ldots. \tag{33}$$

Then if the EKF applied to the extended data set (28) generates a bounded sequence of vectors ψ_i, the sequence $\{f(x^k)\}$ converges and each of the limit points of $\{x^k\}$ is a stationary point of the least squares problem.

We develop the proof of Prop. 2 through a series of lemmas, all of which implicitly assume the conditions of Prop. 2:

Lemma 2. There exist positive scalars c_1 and c_2 such that for all k, the eigenvalues of the matrices H_{km} lie within the interval $[c_1 k, c_2 k]$.

Proof: We have using the update formula (22), that

$$H_{(k+1)m} = \big(\lambda(k+1)\big)^m H_{km} + \sum_{i=1}^{m} \big(\lambda(k+1)\big)^{m-i} \nabla g_i(\psi_{km+i-1}) \nabla g_i(\psi_{km+i-1})'. \tag{34}$$

Let X be a compact set containing all vectors ψ_i generated by the algorithm, and let B and b be an upper bound and a lower bound, respectively, for the eigenvalues of $\nabla g_i(x)\nabla g_i(x)'$ as x ranges over X. From Eq. (34), it is seen by induction that all eigenvalues of H_{km} are less or equal to $c_2 k$ with $c_2 = mB$. Furthermore, if v_k is the smallest eigenvalue of H_{km}, then from Eqs. (33) and (34) it is seen by induction that

$$v_{k+1} \geq \left(1 - \frac{c}{k+1}\right) v_k + \left(1 - \frac{c}{k+1}\right) mb, \qquad \forall\, k \geq 1. \tag{35}$$

Using this relation, we will prove that $v_k \geq k\beta$ for a sufficiently small but positive value of β. Indeed, let $\overline{k}$ be the minimal positive integer k such that $c/k < 1$, and let β be any positive scalar such that

$$\beta \leq \frac{(\overline{k} + 1 - c)mb}{\overline{k} + 1 + \overline{k}c}.$$

From Eq. (35), it is seen that if $v_{\overline{k}} \geq \beta\overline{k}$, then

$$\begin{aligned} v_{\overline{k}+1} &\geq \left(1 - \frac{c}{\overline{k}+1}\right)\beta\overline{k} + \left(1 - \frac{c}{\overline{k}+1}\right) mb \\ &= \beta(\overline{k}+1) + \frac{(\overline{k}+1-c))mb}{\overline{k}+1} - \frac{(\overline{k}+1+\overline{k}c))\beta}{\overline{k}+1} \\ &\geq \beta(\overline{k}+1). \end{aligned}$$

Similarly, it is shown that $v_k \geq \beta k$ for all $k \geq \overline{k}$. Thus, by taking β equal to the scalar c_1 given below,

$$c_1 = \min\left\{\frac{(\overline{k}+1-c)mb}{\overline{k}+1+\overline{k}c}, \min_{k=1,\ldots,\overline{k}} \frac{v_k}{k}\right\},$$

we see that $v_k \geq c_1 k$ for all $k \geq 1$. **Q.E.D.**

We will use the notation

$$f(x) = \frac{1}{2}\sum_{i=1}^{m} \|g_i(x)\|^2, \tag{36}$$

from which we have

$$\nabla f(x) = \sum_{i=1}^{m} \nabla g_i(x) g_i(x). \tag{37}$$

The next lemma shows that the vector that is multiplied by $H^{-1}_{(k+1)m}$ to obtain the direction used by the EKF differs from the gradient $\nabla f(x^k)$ by a relatively small amount.

Lemma 3. Let

$$e^k = \nabla f(x^k) - \sum_{i=1}^{m} (\lambda(k+1))^{m-i} \nabla g_i(\psi_{km+i-1}) g_i(\psi_{km+i-1}). \tag{38}$$

Then there exists a scalar γ such that for all k

$$\|e^k\| \leq \frac{\gamma\left(1+\|\nabla f(x^k)\|\right)}{k+1}. \tag{39}$$

Proof: We have using Eqs. (37) and (38),

$$e^k = \sum_{i=1}^{m}\left(1-\left(\lambda(k+1)\right)^{m-i}\right)\nabla g_i(x^k)g_i(x^k)$$
$$+\sum_{i=1}^{m}\left(\lambda(k+1)\right)^{m-i}\left(\nabla g_i(x^k)g_i(x^k)-\nabla g_i(\psi_{km+i-1})g_i(\psi_{km+i-1})\right),$$

so from the assumption (32) and the formula (37) for $\nabla f(x^k)$, we obtain

$$\|e^k\| \leq \left(1-\left(\lambda(k+1)\right)^m\right)\|\nabla f(x^k)\| + L\sum_{i=1}^{m}\|x^k-\psi_{km+i-1}\|. \tag{40}$$

We also have using Eq. (21), for all k and $i \geq 2$

$$\|x^k-\psi_{km+i-1}\| \leq \|H_{km}^{-1}\|\sum_{j=1}^{i-1}\|\nabla g_i(\psi_{km+j})g_i(\psi_{km+j})\|.$$

Using the boundedness of ψ_i and Lemma 2, we see that for all i and some $\delta > 0$ we have

$$\|x^k-\psi_{km+i-1}\| \leq \frac{\delta}{k+1}.$$

Combining this relation with Eq. (40) and the assumption $1-\left(\lambda(k+1)\right)^m \leq c/(k+1)$, we obtain the desired relation (39). **Q.E.D.**

The assumption (32) together with Eq. (37), imply that

$$\|\nabla f(x)-\nabla f(y)\| \leq mL\|x-y\|, \qquad \forall\ x,y\in\Re^n. \tag{41}$$

The next lemma is a well-known consequence of this relation. We include the proof for completeness.

Lemma 4. For all x and y, there holds

$$f(x+y) \leq f(x)+y'\nabla f(x)+\frac{mL}{2}\|y\|^2. \tag{42}$$

Proof: Let t be a scalar parameter and let $F(t)=f(x+ty)$. Using Eq. (41), we have

$$f(x+y)-f(x) = F(1)-F(0) = \int_0^1 \frac{dF}{dt}(t)\,dt = \int_0^1 y'\nabla f(x+ty)\,dt$$
$$\leq \int_0^1 y'\nabla f(x)\,dt + \left|\int_0^1 y'\left(\nabla f(x+ty)-\nabla f(x)\right)dt\right|$$
$$\leq \int_0^1 y'\nabla f(x)\,dt + \int_0^1 \|y\|\cdot\|\nabla f(x+ty)-\nabla f(x)\|dt$$
$$\leq y'\nabla f(x)+\|y\|\int_0^1 Lt\|y\|\,dt$$
$$= y'\nabla f(x)+\frac{mL}{2}\|y\|^2.$$

Q.E.D.

We are now ready to prove Prop. 2.

Proof of Prop. 2: We have using the Kalman filter recursion (13) and the definition (38) of e^k,

$$x^{k+1} = x^k - H_{(k+1)m}^{-1}\left(\sum_{i=1}^{m}(\lambda(k+1))^{m-i}\nabla g_i(\psi_{km+i-1})g_i(\psi_{km+i-1})\right) = x^k + d^k,$$

where

$$d^k = -H_{(k+1)m}^{-1}(\nabla f(x^k) - e^k). \tag{43}$$

Using Lemmas 2 and 3, and the fact

$$\|H_{(k+1)m}^{-1}\| \leq \frac{1}{c_2(k+1)},$$

which follows from Lemma 2, it is seen that

$$\begin{aligned} d^{k\prime}\nabla f(x^k) &= -\nabla f(x^k)' H_{(k+1)m}^{-1}\nabla f(x^k) + e^{k\prime}H_{(k+1)m}^{-1}\nabla f(x^k) \\ &\leq -\frac{\|\nabla f(x^k)\|^2}{c_2(k+1)} + \frac{\|e^k\|\,\|\nabla f(x^k)\|}{c_2(k+1)} \\ &\leq -\left(\frac{1}{c_2(k+1)} + O\left(\frac{1}{(k+1)^2}\right)\right)\|\nabla f(x^k)\|^2 + O\left(\frac{1}{(k+1)^2}\right)\|\nabla f(x^k)\|, \end{aligned}$$

and

$$\begin{aligned} \|d^k\|^2 &\leq \|H_{(k+1)m}^{-1}\|^2\left(\|\nabla f(x^k)\| + \|e^k\|\right)^2 \\ &= O\left(\frac{1}{(k+1)^2}\right)\left(\|\nabla f(x^k)\| + O\left(\frac{1+\|\nabla f(x^k)\|}{k+1}\right)\right)^2 \\ &= O\left(\frac{1}{(k+1)^2}\right)\|\nabla f(x^k)\|^2 + O\left(\frac{1}{(k+1)^3}\right)\|\nabla f(x^k)\| + O\left(\frac{1}{(k+1)^4}\right). \end{aligned} \tag{44}$$

Using these relations in Eq. (42), we obtain

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + d^{k\prime}\nabla f(x^k) + \frac{mL}{2}\|d^k\|^2 \\ &\leq f(x^k) - \left(\frac{1}{c_2(k+1)} + O\left(\frac{1}{(k+1)^2}\right)\right)\|\nabla f(x^k)\|^2 \\ &\quad + O\left(\frac{1}{(k+1)^2}\right)\|\nabla f(x^k)\| + O\left(\frac{1}{(k+1)^4}\right). \end{aligned}$$

Thus, since $\|\nabla f(x^k)\|$ is bounded, there exist constants $\beta_1 > 0$ and $\beta_2 > 0$, and a positive integer $\overline{k}$ such that

$$f(x^{k+1}) \leq f(x^k) - \frac{\beta_1}{k}\|\nabla f(x^k)\|^2 + \frac{\beta_2}{k^2}, \qquad \forall\ k \geq \overline{k}. \tag{45}$$

It is well known that if $\{u^k\}$ and $\{d^k\}$ are nonnegative sequences such that $u^{k+1} \leq u^k + \delta^k$ for all k and $\sum_{k=1}^{\infty}\delta^k < \infty$, then $\{u^k\}$ converges; this is a special case of the supermartingale convergence

theorem (see e.g. [Pol87], p. 49, or [BeT89], p. 677). Since $f(x) \geq 0$ for all x, it follows from Eq. (45) that $\{f(x^k)\}$ converges.

From Eq. (45) we have for all $k \geq \overline{k}$

$$f(x^{k+1}) \leq f(x^{\overline{k}}) - \sum_{i=\overline{k}}^{k} \frac{\beta_1}{i} \|\nabla f(x^i)\|^2 + \sum_{i=\overline{k}}^{k} \frac{\beta_2}{i^2}. \tag{46}$$

Since $\lim_{k\to\infty} \sum_{i=\overline{k}}^{k} 1/i = \infty$ and $\lim_{k\to\infty} \sum_{i=\overline{k}}^{k} 1/i^2 < \infty$, we see also that there cannot exist an $\epsilon > 0$ such that $\|\nabla f(x^k)\|^2 > \epsilon$ for all $k \geq \overline{k}$. Therefore, we must have $\liminf_{k\to\infty} \|\nabla f(x^k)\| = 0$.

We will now show that $\|\nabla f(x^k)\| \to 0$. Indeed, assume the contrary, that is, there exists an $\epsilon > 0$ such that $\|\nabla f(x^k)\| > \epsilon$ for all k in an infinite subset of integers $\mathcal{K}$. For each $k \in \mathcal{K}$, let $i(k)$ be the first index i such that $i > k$ and $\|\nabla f(x^i)\| < \epsilon/2$, so that

$$\frac{\epsilon}{2} \leq \|\nabla f(x^k)\| - \|\nabla f(x^{i(k)})\| \leq \|\nabla f(x^k) - \nabla f(x^{i(k)})\| \leq L\|x^k - x^{i(k)}\| \leq L \sum_{i=k}^{i(k)-1} \|d^i\|. \tag{47}$$

Since from Eq. (44) we have $\|d^k\| = O(1/k)$, Eq. (47) implies that for some constant $B_1 > 0$,

$$\frac{\epsilon}{2} \leq B_1 \sum_{i=k}^{i(k)-1} \frac{1}{i}, \qquad \forall\, k \in \mathcal{K}.$$

From Eq. (46) we see that

$$f\left(x^{i(k)}\right) \leq f(x^k) - \beta_1 \left(\frac{\epsilon}{2}\right)^2 \sum_{i=k}^{i(k)-1} \frac{1}{i} + \sum_{i=k}^{i(k)-1} \frac{\beta_2}{i^2}, \qquad \forall\, k \in \mathcal{K}.$$

Since $\{f(x^k)\}$ converges and $\lim_{k\to\infty} \sum_{i=k}^{i(k)-1} \beta_2/i^2 = 0$, it follows that

$$\lim_{k\to 0,\, k\in\mathcal{K}} \sum_{i=k}^{i(k)-1} \frac{1}{i} = 0,$$

contradicting the earlier conclusion $\frac{\epsilon}{2} \leq B_1 \sum_{i=k}^{i(k)-1} 1/i$ for all $k \in \mathcal{K}$. Therefore, $\|\nabla f(x^k)\| \to 0$, and it follows that every limit point of $\{x^k\}$ is a stationary point of f. **Q.E.D.**

Note that the proof of Lemma 2 carries through even if the initial matrix H_0 is any positive definite matrix rather than $H_0 = 0$. As a result Prop. 2 also holds when H_0 is some positive definite matrix, in which case it is unnecessary to assume that $\nabla g_1(x^0)\nabla g_1(x^0)'$ is invertible. More generally, our method of proof shows thats the convergence characteristics of the method are maintained by any scheme that varies λ and/or H in a way that the crucial Lemma 2 holds.

In practice the method seems to converge considerably faster if λ is initially less than 1 and is progressively increased towards 1 in a judicious manner. On the other hand an implicit diminishing stepsize as indicated by Lemma 2 is essential for the convergence of the method, and such

a stepsize induces a generically sublinear convergence rate. This characteristic is shared with the backpropagation method where, to achieve a linear convergence rate, it is essential to use a stepsize that is bounded away from 0, but when such a stepsize is used, the method tends to converge to a nonoptimal point [Luo91]. Thus the development of an incremental least squares method with linear convergence rate remains an important open research question.

We finally note that the boundedness assumption on the sequence of vectors ψ_i is a substantial weakness of Prop. 2. It is not easy to remove this assumption because the algorithm does not have an explicit stepsize mechanism to control the magnitude of the initial iterates. On the other hand one can employ the device of projecting the iterates ψ_i on a compact set that is known to contain an optimal solution, and to use a projection version of the EKF of the type introduced in [Ber82a], and [Ber82b], Section 1.5. Projecting the iterates on a compact set is a well-known approach to enhance the theoretical convergence properties of the EKF (see [Lju79]).

REFERENCES

[AWT69] Athans, M., Wishner, R. P., and Bertolini, A., "Suboptimal State Estimation for Continuous Time Nonlinear Systems from Discrete Noisy Measurements," IEEE Trans. on Aut. Control, Vol. AC-13, 1968, pp. 504-514.

[AnM79] Anderson, B. D. O., and Moore, J. B., "Optimal Filtering," Prentice-Hall, Englewood Cliffs, N. J., 1979.

[BeT89] Bertsekas, D. P., and Tsitsiklis, J. N.,"Parallel and Distributed Computation: Numerical Methods," Prentice-Hall, Englewood Cliffs, N. J., 1989.

[Ber82a] Bertsekas, D. P., "Projected Newton Methods for Optimization Problems with Simple Constraints," SIAM J. Contr. and Optimization, Vol. 20, 1982, pp. 221-246.

[Ber82b] Bertsekas, D. P., "Constrained Optimization and Lagrange Multiplier Methods," Academic Press, NY, 1982.

[Dav76] Davidon, W. C., "New Least Squares Algorithms," J. of Optimization Theory and Applications, Vol. 18, 1976, pp. 187-197.

[Gai93] Gaivoronski, A. A., "Convergence Analysis of Parallel Backpropagation Algorithm for Neural Networks," Symposium on Parallel Optimization 3, Madison, July 7-9, 1993.

[Gri93] Grippo, L., "A Class of Unconstrained Minimization Methods for Neural Network Training," Symposium on Parallel Optimization 3, Madison, July 7-9, 1993.

[Jaz70] Jazwinski, A. H., "Stochastic Processes and Filtering Theory," Academic Press, NY, 1970.

[KuC78] Kushner, H. J., and Clark, D. S., "Stochastic Approximation Methods for Constrained and Unconstrained Systems," Springer-Verlag, NY, 1978.

[Lju77] Ljung, L., "Analysis of Recursive Stochastic Algorithms," IEEE Trans. on Aut. Control, Vol. AC-22, 1977, pp. 551-575.

[Lju79] Ljung, L., "Asymptotic Behavior of the Extended Kalman Filter as a Parameter Estimator for Linear Systems," IEEE Trans. on Aut. Control, Vol. AC-24, 1979, pp. 36-50.

[LuT93] Luo, Z. Q., and Tseng, P., "Analysis of an Approximate Gradient Projection Method with Applications to the Back Propagation Algorithm," Dept. Elec. and Computer Eng., McMaster Univ., Hamilton, Ontario and Dept. of Math., Univ. Washington, Seattle, August 1993.

[Luo91] Luo, Z. Q., "On the Convergence of the LMS Algorithm with Adaptive Learning Rate for Linear Feedforward Networks," Neural Computation, Vol. 3, 1991, pp. 226-245.

[MaS93] Mangasarian, O. L., and Solodov, M. V., "Serial and Parallel Backpropagation Convergence Via Nonmonotone Perturbed Minimization," Computer Science Dept., Computer Sciences Technical Report No. 1149, Univ. of Wisconsin-Madison, April 1993.

[Man93] Mangasarian, O. L, "Mathematical Programming in Neural Networks," ORSA J. on Computing, Vol. 5, 1993, pp. 349-360.

[Meh71] Mehra, R. K., "A Comparison of Several Nonlinear Filters for Reentry Vehicle Tracking," IEEE Trans. Aut. Control, Vol. AC-16, 1971, pp. 307-319.

[Pap82] Pappas, T. N., "Solution of Nonlinear Equations by Davidon's Least Squares Method," M.S. Thesis, Dept. of Electrical Enginering and Computer Science, Mass. Institute of Technology, Cambridge, MA, 1982.

[PoT73] Poljak, B. T, and Tsypkin, Y. Z., "Pseudogradient Adaptation and Training Algorithms," Automation and Remote Control, 1973, pp. 45-68.

[Pol87] Poljak, B. T, "Introduction to Optimization," Optimization Software Inc., N.Y., 1987.

[RRK92] Ruck, D. W., Rogers, S. K., Kabrisky, M., Maybeck, P. S., and Oxley, M. E., "Comparative Analysis of Backpropagation and the Extended Kalman Filter for Training Multilayer Perceptrons,"

IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 14, 1992, pp. 686-691.

[TBA86] Tsitsiklis, J. N., Bertsekas, D. P., and Athans, M., "Distributed Asynchronous Deterministic and Stochastic Gradient Optimization Algorithms," IEEE Trns. on Aut. Control, Vol. AC-31, 1986, pp. 803-812.

[THS77] Tenney, R. R., Hebbert, R. S., and Sandell, N. R., Jr, "Tracking Filter for Maneuvering Sources," IEEE Trns. on Aut. Control, Vol. AC-22, 1977, pp. 246-251.

[Tsi84] Tsitsiklis, J. N., "Problems in Decentralized Decision Making and Computation," Ph.D. Dissertation, Dept. of Electrical Enginering and Computer Science, Mass. Institute of Technology, Cambridge, MA, 1984.

[Urs80] Ursin, B., "Asymptotic Convergence Properties of the Extended Kalman Filter Using Filtered State Estimates," IEEE Trans. on Aut. Control, Vol. AC-25, 1980, pp. 1207-1211.

[Whi89a] White, H., "Some Asymptotic Results for Learning in Single Hidden-Layer Feedforward Network Models," J. Am. Statistical Association, Vol. 84, 1989.

[Whi89b] White, H., "Learning in Artificial Neural Networks: A Statistical Perspective," Neural Computation, Vol. 1, 1989, pp. 425-464.

[WaT90] Watanabe, K., and Tzafestas, S. G., "Learning Algorithms for Neural Networks with the Kalman Filters," J. Intelligent and Robotic Systems, Vol. 3, 1990, pp. 305-319.

[WeM80] Weiss, H., and Moore, J. B., "Improved Extended Kalman Filter Design for Passive Tracking," IEEE Trans. on Aut. Control, Vol. AC-25, 1980, pp. 807-811.

[WiH60] Widrow, B., and Hoff, M. E., "Adaptive Switching Circuits," Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record, part 4, 1960, pp. 96-104.

[WiS85] Widrow, B., and Stearns, S. D., "Adaptive Signal Processing," Prentice-Hall, Englewood Cliffs, N. J., 1985.