J Optim Theory Appl (2012) 154:187–208
DOI 10.1007/s10957-012-0001-1

Cyclic Seesaw Process for Optimization and Identification

James C. Spall

Received: 11 February 2011 / Accepted: 25 January 2012 / Published online: 10 March 2012

Abstract A known approach to optimization is the cyclic (or alternating or block coordinate) method, where the full parameter vector is divided into two or more subvectors and the process proceeds by sequentially optimizing each of the subvectors, while holding the remaining parameters at their most recent values. One advantage of such a scheme is the preservation of potentially large investments in software, while allowing for an extension of capability to include new parameters for estimation. A specific case of interest involves cross-sectional data that is modeled in state–space form, where there is interest in estimating the mean vector and covariance matrix of the initial state vector as well as certain parameters associated with the dynamics of the underlying differential equations (e.g., power spectral density parameters). This paper shows that, under reasonable conditions, the cyclic scheme leads to parameter estimates that converge to the optimal joint value for the full vector of unknown parameters. Convergence conditions here differ from others in the literature. Further, relative to standard search methods on the full vector, numerical results here suggest a more general property of faster convergence for seesaw as a consequence of the more "aggressive" (larger) gain coefficient (step size) possible.

Keywords System identification · Parameter estimation · Alternating optimization · Cyclic optimization · Block coordinate optimization · Recursive estimation · Nondifferentiable

Communicated by Johannes O. Royset.

J.C. Spall (✉)
Applied Physics Laboratory, The Johns Hopkins University, Laurel, MD 20723-6099, USA
e-mail: james.spall@jhuapl.edu

J.C. Spall
Department of Applied Mathematics and Statistics, The Johns Hopkins University, Baltimore, MD 21218, USA

1 Introduction

A known method for optimization is the cyclic (also called alternating or block coordinate) approach, where the full parameter vector is divided into two or more subvectors and the process proceeds by sequentially optimizing the criterion of interest with respect to each of the subvectors while holding the other subvectors fixed. One application of such a method arises in system identification for state–space (dynamical) models, where it is sometimes the case that models are modified to include unknown parameters that may not have been present in an original implementation or that may have been assumed known. For example, in the author's work on a defense system, there has long been interest in estimating the mean vector and covariance matrix for the initial state in a state–space model (e.g., Shumway et al. [1]; Sun [2]). An extensive suite of software has been developed to carry out the estimation based on multiple tests for the system; this sophisticated software produces physically meaningful estimates (i.e., the estimates meet constraints) in an efficient, numerically stable manner. More recently, there has been a need to extend the estimation setting to include unknown parameters associated with the dynamical parameters in the underlying state–space models. There is strong interest in developing methods that preserve the substantial investment in software for estimating only the mean vector and covariance matrix, while allowing for an extension to include dynamical parameters. More generally, this paper provides the theoretical foundation for the cyclic approach to such joint estimation in arbitrary identification and optimization problems.

Because our focus is on the division of the full parameter vector into two subvectors, we sometimes refer to the resulting back-and-forth cyclic process as a "seesaw" process. This contrasts with traditional methods of directly optimizing the full set of all relevant parameters. The seesaw process represents a form of cyclic optimization. The method applies as well to a portioning of the full vector into three or more subvectors. Our discussion is in the context of *minimization* and associated loss functions.

2 Background and Related Literature

Let $\boldsymbol{\theta}$ be a p-dimensional vector representing the unknown parameters to be estimated and $L = L(\boldsymbol{\theta})$ be the loss function to be minimized (e.g., a negative log-likelihood function). According to the seesaw estimation, we represent $\boldsymbol{\theta}$ as composed of two subvectors $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$:

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}^{(1)} \\ \boldsymbol{\theta}^{(2)} \end{pmatrix}.$$

Iteration by iteration, the subvector $\boldsymbol{\theta}^{(1)}$ is estimated based on the most recent value of $\boldsymbol{\theta}^{(2)}$ and, likewise, $\boldsymbol{\theta}^{(2)}$ is estimated based on the most recent value of $\boldsymbol{\theta}^{(1)}$. Thus, there are two subiterations in each full iteration of seesaw, with each subiteration corresponding to the update of $\boldsymbol{\theta}^{(1)}$ or $\boldsymbol{\theta}^{(2)}$. In the application of interest for the author, $\boldsymbol{\theta}^{(1)}$ represents all parameters associated with the mean vector and covariance matrix for the initial state in a state–space model and $\boldsymbol{\theta}^{(2)}$ represents the power spectral

density parameters that enter the process noise covariance matrix (see Sect. 4). We are using the more general $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$ notation because the results are not restricted to this specific allocation of parameters.

A fundamental issue in justifying the seesaw estimation process is the question of convergence. The method cannot be blindly applied without carefully considering conditions for convergence, as demonstrated in the simple counterexample in Achtziger [3]. In particular, consider the linear programming problem of minimizing $\theta^{(1)} + 2\theta^{(2)}$ subject to $\theta^{(1)} + \theta^{(2)} = 1$, where $\theta^{(1)}$ and $\theta^{(2)}$ are two nonnegative scalar parameters. At the feasible point $\boldsymbol{\theta} = (0, 1)^T$ (superscript T is transpose), it is not possible to change either of $\theta^{(1)}$ or $\theta^{(2)}$ (separately) to reduce L. Hence, the method would be stuck at the suboptimal loss $L(\boldsymbol{\theta}) = 2$, not being able to reach $L(\theta^*) = 1$ at $\theta^* = (1, 0)^T$. Another simple counterexample (suggested by a reviewer of this paper) is a bilinear problem, where the loss function to be minimized is a product of two scalar parameters, $\theta^{(1)}$ and $\theta^{(2)}$, on the domain $[-1, 1] \times [-1, 1]$. When $\theta^{(1)} = 0$ and $\theta^{(2)} = 0$, it is not possible to change either parameter alone to reduce the loss function toward its minimum of -1; hence, the seesaw process can get stuck. We present a theorem and supporting corollaries below that preclude such counterexamples and give sufficient conditions for convergence to an optimum.

Note that the seesaw idea is a generalization of a known method within nonlinear programming (sometimes called the Gauss–Seidel method), where a parameter vector is sequentially optimized along each linearly independent coordinate direction (Bazaraa et al. [4, pp. 254–255]; Miller [5, pp. 256–257]). However, it was found that the theory associated with this coordinate-wise method was of little use in showing convergence for the seesaw method here because of our interest in working with *groups* of parameters, not necessarily linearly independent between groups. Others have considered convergence for the cyclic scheme. For example, Bezdek and Hathaway [6] consider a partitioning of $\boldsymbol{\theta}$ into two or more subvectors and show a q-linear convergence rate when the loss function is strictly convex and twice differentiable (q-linear convergence implies that iterate error drops at a rate at least as fast as q times the error in the previous iterate, $0 < q < 1$; see Bazaraa et al. [4, pp. 257–258]). Tseng [7] considers convergence to a stationary, but not necessarily minimum, point for loss functions that include a nondifferentiable and separable contribution (usually added to a nonseparable differentiable contribution). Bertsekas [8, Sect. 2.7] shows convergence to a stationary point for continuously differentiable loss functions when it is possible to fully (and uniquely) minimize the loss in terms of each of the subvectors. Konno [9], Alarie et al. [10], and Audet et al. [11] present convergence results for the cyclic scheme with two subvectors, as emphasized here, when applied to bilinear models having the form $\boldsymbol{\theta}^{(1)T} \boldsymbol{A} \boldsymbol{\theta}^{(2)}$ + linear part, where $\boldsymbol{A}$ is an appropriately dimensioned rectangular matrix and the "linear part" includes linear functions of $\boldsymbol{\theta}^{(1)}$, $\boldsymbol{\theta}^{(2)}$ (Audet et al. [11] also allows for a third parameter vector that is optimized at each subiteration to be included in the seesaw process). An especially appealing aspect of the application of seesaw to bilinear problems is that the two subiterations reduce to linear programming problems. We present some global convergence theory under conditions different than those above.

Let us mention several applications of the cyclic idea. Lee and Park [12] demonstrate numerical convergence and high efficiency, relative to the powerful Levenberg–Marquardt algorithm, for a problem in classification and computer vision. There

have also been applications of the cyclic optimization idea in the context of the expectation-maximization (EM) method for finding maximum likelihood parameter estimates. For example, an approach having some resemblance to the seesaw method is the SAGE method for maximum likelihood estimation (Fessler and Hero [13]). SAGE is based on dividing the overall parameter vector in the "M" step of EM into two parts for carrying out the optimization. One distinction between SAGE and the seesaw method is that seesaw applies to arbitrary loss functions, including those that are not a likelihood function and/or those for which the EM method is not being used for carrying out the optimization. Likewise, Haaland et al. [14], use the cyclic idea (with four subvectors) to carry out the "M" step in the context of parameter estimation for multivariate Gaussian autoregressive hidden Markov models as applied to a problem in temperature control for a large data center and Fessler et al. [15] use the cyclic idea for medical image reconstruction for a class of penalized likelihood problems.

Before proceeding with the main results, let us introduce some notation and basic concepts associated with the identification problem of interest. A formal representation of the parameter estimation problem of interest here is to find the set:

$$\Theta^* := \arg\min_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}) := \left\{\boldsymbol{\theta}^* \in \Theta : L\left(\boldsymbol{\theta}^*\right) \leq L(\boldsymbol{\theta}) \text{ for all } \boldsymbol{\theta} \in \Theta\right\},$$

where $\Theta \subseteq \mathbb{R}^p$ represents the possible values for $\boldsymbol{\theta}$ (i.e., the constraint set for $\boldsymbol{\theta}$). We assume that at least one minimum $\boldsymbol{\theta}^*$ exists; that is, Θ^* is nonempty. The vector elements $\boldsymbol{\theta}^* \in \Theta^* \subseteq \Theta$ are equivalent solutions in the sense that they yield identical values of the loss function. In practice, it is usually sufficient to identify just *one* element of Θ^*. Note that the minimization problem above is well defined when, for example, Θ is compact and L is continuous because it is known that at least one $\boldsymbol{\theta}^*$ exists such that $L(\boldsymbol{\theta}^*) = \min_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta})$ (Polak [16, p. 655]).

3 Algorithm and Convergence Analysis

3.1 Seesaw Algorithm

The estimate at iteration k in the seesaw approach has the form

$$\hat{\boldsymbol{\theta}}_k := \begin{pmatrix} \hat{\boldsymbol{\theta}}_k^{(1)} \\ \hat{\boldsymbol{\theta}}_k^{(2)} \end{pmatrix},$$

with $\hat{\boldsymbol{\theta}}_k^{(1)}$ a function of $\hat{\boldsymbol{\theta}}_{k-1}$, and $\hat{\boldsymbol{\theta}}_k^{(2)}$ a function of $\hat{\boldsymbol{\theta}}_k^{(1)}$ and $\hat{\boldsymbol{\theta}}_{k-1}^{(2)}$. The value $\hat{\boldsymbol{\theta}}_0$ represents an initial guess for $\boldsymbol{\theta}$. It is assumed that the seesaw process satisfies the following relationship:

$$L(\hat{\boldsymbol{\theta}}_{k+1}) \leq L\left(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \hat{\boldsymbol{\theta}}_k^{(2)}\right) \leq L(\hat{\boldsymbol{\theta}}_k) \tag{1}$$

for all k. Further, $\hat{\boldsymbol{\theta}}_{k+1}^{(1)} \neq \hat{\boldsymbol{\theta}}_k^{(1)}$ or $\hat{\boldsymbol{\theta}}_{k+1}^{(2)} \neq \hat{\boldsymbol{\theta}}_k^{(2)}$ only if there is strict reduction in the loss function in stage 1 or 2, respectively, of the seesaw process. Thus, overall, $\hat{\boldsymbol{\theta}}_{k+1} \neq \hat{\boldsymbol{\theta}}_k$ only if

$$L(\hat{\boldsymbol{\theta}}_{k+1}) < L(\hat{\boldsymbol{\theta}}_k). \tag{2}$$

(In practice, checking (1) and (2) requires evaluations of L that are not formally needed in standard implementations of gradient-based methods such as steepest descent, Newton–Raphson, EM, or scoring; see, e.g., Bazaraa et al. [4, Chap. 8]; Spall [17, Chap. 1]; Ng et al. [18]; and Levy [19].) Let $L^* = L(\boldsymbol{\theta}^*)$ for $\boldsymbol{\theta}^* \in \Theta^*$. As in the notation and ordering of operations in (1), we let the per-iteration minima for each of $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$ be denoted by $\boldsymbol{\theta}_k^{*(1)}$ and $\boldsymbol{\theta}_k^{*(2)}$, respectively. That is, $\boldsymbol{\theta}_k^{*(1)} = \arg\min_{\boldsymbol{\theta}^{(1)}} L(\boldsymbol{\theta}^{(1)}, \hat{\boldsymbol{\theta}}_{k-1}^{(2)})$ and $\boldsymbol{\theta}_k^{*(2)} = \arg\min_{\boldsymbol{\theta}^{(2)}} L(\hat{\boldsymbol{\theta}}_k^{(1)}, \boldsymbol{\theta}^{(2)})$. Therefore, $\boldsymbol{\theta}_k^{*(1)}$ is a function of $\hat{\boldsymbol{\theta}}_{k-1}^{(2)}$ while $\boldsymbol{\theta}_k^{*(2)}$ is a function of $\hat{\boldsymbol{\theta}}_k^{(1)}$.

More formally, the seesaw process can be described according to the following steps:

Step 0: Set $k = 0$ and choose initial condition $\hat{\boldsymbol{\theta}}_0$.

Step 1a: From $\hat{\boldsymbol{\theta}}_k$, determine a candidate value of $\boldsymbol{\theta}^{(1)}$ from whatever search algorithm is being used. Set $\hat{\boldsymbol{\theta}}_{k+1}^{(1)} = \boldsymbol{\theta}^{(1)}$ if $L(\boldsymbol{\theta}^{(1)}, \hat{\boldsymbol{\theta}}_k^{(2)}) < L(\hat{\boldsymbol{\theta}}_k)$; else set $\hat{\boldsymbol{\theta}}_{k+1}^{(1)} = \hat{\boldsymbol{\theta}}_k^{(1)}$.

Step 1b: From $\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \hat{\boldsymbol{\theta}}_k^{(2)}$, determine a candidate value of $\boldsymbol{\theta}^{(2)}$ from whatever search algorithm is being used. Set $\hat{\boldsymbol{\theta}}_{k+1}^{(2)} = \boldsymbol{\theta}^{(2)}$ if $L(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \boldsymbol{\theta}^{(2)}) < L(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \hat{\boldsymbol{\theta}}_k^{(2)})$; else set $\hat{\boldsymbol{\theta}}_{k+1}^{(2)} = \hat{\boldsymbol{\theta}}_k^{(2)}$.

Step 2: Replace k with $k + 1$ and go to Step 1a; terminate as appropriate.

3.2 Convergence Analysis

An important issue in this partial decoupling of the estimation process is the question of convergence: Under what conditions does the seesaw estimation process lead to convergence of the loss $L(\hat{\boldsymbol{\theta}}_k)$ and/or estimate $\hat{\boldsymbol{\theta}}_k$ to the corresponding optimal L or $\boldsymbol{\theta}$ as the number of iterations in the estimation process increase? We now present a theorem and supporting corollaries that gives sufficient conditions for convergence.

Theorems 3.1 and 3.2 below consider the convergence of $L(\hat{\boldsymbol{\theta}}_k)$ and $\hat{\boldsymbol{\theta}}_k$ for continuous, but not necessarily differentiable, loss functions. Theorem 3.2 gives a more-checkable special case of a main condition in Theorem 3.1. Corollary 3.1 pertains to continuously differentiable loss functions that are *pseudoconvex* (e.g., Bazaraa et al. [4, pp. 113–115]; Miller [5, p. 558]). Pseudoconvexity is a significant generalization of convexity to include functions that do not have the classical "bowl shape." However, as with convexity, pseudoconvex functions have the property that if the loss function gradient $\boldsymbol{g}(\boldsymbol{\theta}) = \mathbf{0}$ at some point $\boldsymbol{\theta}$, then this $\boldsymbol{\theta}$ corresponds to a global minimum $\boldsymbol{\theta}^*$. The loss function is pseudoconvex iff for each $\bar{\boldsymbol{\theta}}, \bar{\bar{\boldsymbol{\theta}}} \in \Theta$,

$$L(\bar{\bar{\boldsymbol{\theta}}}) < L(\bar{\boldsymbol{\theta}}) \quad \text{implies} \quad g(\bar{\boldsymbol{\theta}})^T(\bar{\bar{\boldsymbol{\theta}}} - \bar{\boldsymbol{\theta}}) < 0, \tag{3}$$

where Θ is a convex set. Note that pseudoconvexity does not guarantee uniqueness of the global minimum. However, under stronger conditions of *strict* pseudoconvexity, $\boldsymbol{\theta}^*$ is unique (L is strictly pseudoconvex iff for each distinct $\bar{\boldsymbol{\theta}}, \bar{\bar{\boldsymbol{\theta}}} \in \Theta$, $L(\bar{\bar{\boldsymbol{\theta}}}) \le L(\bar{\boldsymbol{\theta}})$ implies $g(\bar{\boldsymbol{\theta}})^T(\bar{\bar{\boldsymbol{\theta}}} - \bar{\boldsymbol{\theta}}) < 0$; see, e.g., Bazaraa et al. [4, pp. 112 and 116]). Corollaries 3.3 and 3.4 generalize the two-stage seesaw process to an M-stage process, where $\boldsymbol{\theta}$ is divided into M subvectors, $2 \le M \le p$.

Theorem 3.1 *Suppose that Θ is a compact, convex set and that $L(\boldsymbol{\theta})$ is continuous on Θ. Suppose that at any $\boldsymbol{\theta} \in \Theta$ with $\boldsymbol{\theta} \notin \Theta^*$, it is possible to change at least one of $\boldsymbol{\theta}^{(1)}$ or $\boldsymbol{\theta}^{(2)}$ to yield a reduction in L and that the two-stage algorithm with properties (1) and (2) reduces L with respect to $\boldsymbol{\theta}^{(1)}$ or $\boldsymbol{\theta}^{(2)}$ in the sense that at least one of (4a) or (4b) holds for each $k = 0, 1, 2, \ldots$:*

$$L\big(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \hat{\boldsymbol{\theta}}_{k}^{(2)}\big) \le L\big(\boldsymbol{\theta}_{k+1}^{*(1)}, \hat{\boldsymbol{\theta}}_{k}^{(2)}\big) + \alpha_k^{(1)} \quad \text{if } \boldsymbol{\theta}^{(1)} \text{ is changed or} \tag{4a}$$

$$L\big(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \hat{\boldsymbol{\theta}}_{k+1}^{(2)}\big) \le L\big(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \hat{\boldsymbol{\theta}}_{k+1}^{*(2)}\big) + \alpha_k^{(2)} \quad \text{if } \boldsymbol{\theta}^{(2)} \text{ is changed,} \tag{4b}$$

where the sequences $\alpha_k^{(1)}$ and $\alpha_k^{(2)}$ are nonnegative and converge to 0. Then

$$L(\hat{\boldsymbol{\theta}}_k) \to L^* \quad \text{as } k \to \infty. \tag{5}$$

Further, if $\boldsymbol{\theta}^*$ is unique (i.e., Θ^* is the singleton $\boldsymbol{\theta}^*$), then

$$\hat{\boldsymbol{\theta}}_k \to \boldsymbol{\theta}^* \quad \text{as } k \to \infty. \tag{6}$$

Remark 3.1 While (2) and the finite lower bound on L ensure that some limiting point L' exists (as used in the proof of Theorem 3.1 below), those two conditions alone are not sufficient to guarantee that $L' = L^*$. For example, if $-\infty < L^* < -2$, $L(\hat{\boldsymbol{\theta}}_{k+1}) = L(\hat{\boldsymbol{\theta}}_k) - 1/2^k$, and $L(\hat{\boldsymbol{\theta}}_0) = 0$, then (2) is satisfied, but $L' = -\sum_{k=0}^{\infty} 1/2^k = -2 \neq L^*$.

Remark 3.2 By the reduction property in (1) and (2), we have $0 \le \alpha_k^{(1)} \le L(\hat{\boldsymbol{\theta}}_k) - L(\boldsymbol{\theta}_{k+1}^{*(1)}, \hat{\boldsymbol{\theta}}_k^{(2)})$ and $0 \le \alpha_k^{(2)} \le L(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \hat{\boldsymbol{\theta}}_k^{(2)}) - L(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \hat{\boldsymbol{\theta}}_{k+1}^{*(2)})$.

Proof Because L is continuous and Θ is bounded, we have $L^* > -\infty$. Further, $L(\hat{\boldsymbol{\theta}}_k)$ is monotonically nonincreasing by (1), implying that $\lim_{k\to\infty} L(\hat{\boldsymbol{\theta}}_k)$ exists (e.g., Apostol [20, p. 185]). Let L' be the limiting point (i.e., $\lim_{k\to\infty} L(\hat{\boldsymbol{\theta}}_k) = L'$). Let us prove that $L' = L^*$ (i.e., (5) holds). Because Θ is a bounded set, there exists a convergent subsequence $\{\hat{\boldsymbol{\theta}}_{k_j}\}$ as $j \to \infty$ (Fleming [21, p. 47]). Further, by the convergence of $L(\hat{\boldsymbol{\theta}}_k)$ and continuity of L on the compact set Θ, we know that the limiting point for this subsequence is some $\boldsymbol{\theta}' \in \Theta$ such that $L(\boldsymbol{\theta}') = L'$. We show $L' = L^*$ by demonstrating that a contradiction results when $L' > L^*$.

Consider the point $\boldsymbol{\theta}'$ ($\boldsymbol{\theta}'$ is apportioned according to $\boldsymbol{\theta}'^{(1)}$, $\boldsymbol{\theta}'^{(2)}$ corresponding to the first- and second-stage parameters $\boldsymbol{\theta}^{(1)}$, $\boldsymbol{\theta}^{(2)}$). Because $L(\boldsymbol{\theta}') > L^*$, it is known by assumption that a change in at least one of $\boldsymbol{\theta}'^{(1)}$ or $\boldsymbol{\theta}'^{(2)}$ will reduce L. Suppose that a change applied to $\boldsymbol{\theta}'^{(1)}$ reduces L. Hence, there exists a point $\boldsymbol{\theta}'' \equiv ((\boldsymbol{\theta}''^{(1)})^T, (\boldsymbol{\theta}'^{(2)})^T)^T$ and $\varepsilon > 0$ such that $L(\boldsymbol{\theta}') - L(\boldsymbol{\theta}'') \ge \varepsilon$. Let $\boldsymbol{\Delta}^{(1)} \equiv \boldsymbol{\theta}''^{(1)} - \boldsymbol{\theta}'^{(1)}$. The continuity of L implies that for all $0 < \varepsilon_0 < \varepsilon$ there exists an $N(\varepsilon_0) > 0$ such that

$$\big|L\big(\hat{\boldsymbol{\theta}}_{k_j}^{(1)} + \boldsymbol{\Delta}^{(1)}, \hat{\boldsymbol{\theta}}_{k_j}^{(2)}\big) - L\big(\boldsymbol{\theta}''\big)\big| \le \varepsilon_0 \tag{7}$$

for all $j \ge N(\varepsilon_0)$. Hence, from (7),

$$\begin{aligned} L\big(\boldsymbol{\theta}'\big) - L\big(\hat{\boldsymbol{\theta}}_{k_j}^{(1)} + \boldsymbol{\Delta}^{(1)}, \hat{\boldsymbol{\theta}}_{k_j}^{(2)}\big) &= \big[L\big(\boldsymbol{\theta}'\big) - L\big(\boldsymbol{\theta}''\big)\big] + \big[L\big(\boldsymbol{\theta}''\big) - L\big(\hat{\boldsymbol{\theta}}_{k_j}^{(1)} + \boldsymbol{\Delta}^{(1)}, \hat{\boldsymbol{\theta}}_{k_j}^{(2)}\big)\big] \\ &\ge \varepsilon - \varepsilon_0 \\ &> 0 \end{aligned} \tag{8}$$

for all $j \geq N(\varepsilon_0)$. However, by the guaranteed reduction property of the algorithm in (4a), it is known that at any $j \geq N(\varepsilon_0)$

$$\begin{aligned} L\big(\hat{\boldsymbol{\theta}}_{k_j+1}^{(1)}, \hat{\boldsymbol{\theta}}_{k_j}^{(2)}\big) &\leq L\big(\boldsymbol{\theta}_{k_j+1}^{*(1)}, \hat{\boldsymbol{\theta}}_{k_j}^{(2)}\big) + \alpha_{k_j}^{(1)} \\ &\leq L\big(\hat{\boldsymbol{\theta}}_{k_j}^{(1)} + \boldsymbol{\Delta}^{(1)}, \hat{\boldsymbol{\theta}}_{k_j}^{(2)}\big) + \alpha_{k_j}^{(1)} \\ &\leq L' - \varepsilon + \varepsilon_0 + \alpha_{k_j}^{(1)}, \end{aligned} \tag{9}$$

where we have used (8) to obtain the last inequality. By the fact that $\alpha_k^{(1)} \to 0$, it follows that there exists an integer N_1 such that $\alpha_k^{(1)} < \varepsilon - \varepsilon_0$ for all $k \geq N_1$. Taking $j \geq \max\{N_1, N(\varepsilon_0)\}$, we have

$$L(\hat{\boldsymbol{\theta}}_{k_j+1}) < L'. \tag{10}$$

The monotonicity of (1) leads to a contradiction of (10) with $\lim_{k\to\infty} L(\hat{\boldsymbol{\theta}}_k) = L' > L(\boldsymbol{\theta}^*)$. Based on assumption (4b), analogous reasoning applies for a change applied to $\boldsymbol{\theta}'^{(2)}$ that reduces L. Hence, it has been shown that (5) is true.

Let us now show (6). Suppose by contradiction that (6) is not true. That is, there exists a subsequence $\{\hat{\boldsymbol{\theta}}_{k_j}\}$ such that $\|\hat{\boldsymbol{\theta}}_{k_j} - \boldsymbol{\theta}^*\| \geq \eta$ for some $\eta > 0$ and all $j = 0, 1, 2, \ldots$. By the boundedness of Θ, there exists a subsequence of $\{\hat{\boldsymbol{\theta}}_{k_j}\}$ (i.e., a sub-subsequence of $\{\hat{\boldsymbol{\theta}}_k\}$) converging to some point $\boldsymbol{\theta}''' \in \Theta$ such that $\|\hat{\boldsymbol{\theta}}''' - \boldsymbol{\theta}^*\| \geq \eta$ (Fleming [21, p. 47]). By the continuity of L on Θ, it is known that $L(\boldsymbol{\theta}''')$ exists and, because $\boldsymbol{\theta}^*$ is unique, $L(\boldsymbol{\theta}''') > L(\boldsymbol{\theta}^*)$. Hence, the associated sub-subsequence of $L(\hat{\boldsymbol{\theta}}_k)$ converges to $L(\boldsymbol{\theta}''') > L(\boldsymbol{\theta}^*)$, which violates (5), indicating that (6) must be true. □

Theorem 3.2 below presents the same convergence conclusions under alternative conditions to (4a, 4b). These conditions may be more checkable in some cases. In particular, the alternative conditions below imply that at each iteration the search is reducing L by at least some fixed fraction $0 < \gamma \leq 1$ of the possible improvement in the direction of at least one of the parameter subvectors $\boldsymbol{\theta}^{(1)}$ or $\boldsymbol{\theta}^{(2)}$. For example, if $\gamma = 0.1$, then it is known that the search will always yield an improvement of at least 10 percent of the maximum possible improvement in at least one of the two subvectors. If $\gamma = 1$, then the search is such that L is *minimized* in at least one of $\boldsymbol{\theta}^{(1)}$ or $\boldsymbol{\theta}^{(2)}$ at each iteration, corresponding to one of the conditions in the above-mentioned convergence result in Bertsekas [8, Sect. 2.7].

Theorem 3.2 *Suppose that all conditions of Theorem 3.1 hold except that* (4a) *and* (4b) *are replaced by*

$$\frac{L(\hat{\boldsymbol{\theta}}_k) - L(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \hat{\boldsymbol{\theta}}_k^{(2)})}{L(\hat{\boldsymbol{\theta}}_k) - L(\boldsymbol{\theta}_{k+1}^{*(1)}, \hat{\boldsymbol{\theta}}_k^{(2)})} \geq \gamma \quad \text{if } \boldsymbol{\theta}^{(1)} \text{ is changed or} \tag{11a}$$

$$\frac{L(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \hat{\boldsymbol{\theta}}_k^{(2)}) - L(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \hat{\boldsymbol{\theta}}_{k+1}^{(2)})}{L(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \hat{\boldsymbol{\theta}}_k^{(2)}) - L(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \boldsymbol{\theta}_{k+1}^{*(2)})} \geq \gamma \quad \text{if } \boldsymbol{\theta}^{(2)} \text{ is changed,} \tag{11b}$$

where $0 < \gamma \leq 1$. *Then the conclusions shown in* (5) *and* (6) *hold.*

Proof The proof follows exactly as in Theorem 3.1 with the exception of replacing the statement at (9) by the fact that the guaranteed reduction property of the algorithm in (4a) implies that at any $j \geq N(\varepsilon_0)$

$$\begin{aligned} L\big(\hat{\boldsymbol{\theta}}^{(1)}_{k_j+1}, \hat{\boldsymbol{\theta}}^{(2)}_{k_j}\big) &\leq \gamma\big[L\big(\boldsymbol{\theta}^{*(1)}_{k_j+1}, \hat{\boldsymbol{\theta}}^{(2)}_{k_j}\big) - L(\hat{\boldsymbol{\theta}}_{k_j})\big] + L(\hat{\boldsymbol{\theta}}_{k_j}) \\ &\leq \gamma\big[L\big(\hat{\boldsymbol{\theta}}^{(1)}_{k_j} + \boldsymbol{\Delta}^{(1)}, \hat{\boldsymbol{\theta}}^{(2)}_{k_j}\big) - L(\hat{\boldsymbol{\theta}}_{k_j})\big] + L(\hat{\boldsymbol{\theta}}_{k_j}) \\ &= \gamma L\big(\hat{\boldsymbol{\theta}}^{(1)}_{k_j} + \boldsymbol{\Delta}^{(1)}, \hat{\boldsymbol{\theta}}^{(2)}_{k_j}\big) + (1-\gamma)L(\hat{\boldsymbol{\theta}}_{k_j}). \end{aligned} \tag{12}$$

Hence, by the convergence of $L(\hat{\boldsymbol{\theta}}_{k_j})$, it is known that for any ε_1 satisfying $0 < \varepsilon_1 < \gamma(\varepsilon - \varepsilon_0)/(1-\gamma)$ ($\varepsilon_1 < \infty$ when $\gamma = 1$), there exists an $N_1(\varepsilon_1) > 0$ such that $L(\hat{\boldsymbol{\theta}}_{k_j}) - L' \leq \varepsilon_1$ for all $j \geq N_1(\varepsilon_1)$. Using (8) in (12) implies that for all $j \geq \max\{N_1(\varepsilon_1), N(\varepsilon_0)\}$,

$$\begin{aligned} L\big(\hat{\boldsymbol{\theta}}^{(1)}_{k_j+1}, \hat{\boldsymbol{\theta}}^{(2)}_{k_j}\big) &\leq \gamma\big(L' - \varepsilon + \varepsilon_0\big) + (1-\gamma)\big(L' + \varepsilon_1\big) \\ &= L' + \gamma(-\varepsilon + \varepsilon_0) + (1-\gamma)\varepsilon_1 \\ &< L'. \end{aligned}$$

The remainder of the proof proceeds exactly as below (10). □

Note that a sequence satisfying conditions (11a, 11b) of Theorem 3.2 also satisfies (4a, 4b) Theorem 3.1. That is, any seesaw sequence that satisfies the assumptions of Theorem 3.2 will give rise by the proof of Theorem 3.2 to a sequence $L(\hat{\boldsymbol{\theta}}_k) \to L^*$. It follows easily from this result that the sequences $L(\boldsymbol{\theta}^{*(1)}_{k+1}, \hat{\boldsymbol{\theta}}^{(2)}_k)$ and $L(\hat{\boldsymbol{\theta}}^{(1)}_k, \boldsymbol{\theta}^{*(2)}_{k+1})$ also converge to L^* and, therefore, that the sequences $L(\hat{\boldsymbol{\theta}}_k) - L(\boldsymbol{\theta}^{*(1)}_{k+1}, \hat{\boldsymbol{\theta}}^{(2)}_k)$ and $L(\hat{\boldsymbol{\theta}}_k) - L(\hat{\boldsymbol{\theta}}^{(1)}_k, \boldsymbol{\theta}^{*(2)}_{k+1})$ converge to 0. Hence, taking $\alpha^{(1)}_k = (1+\gamma)[L(\hat{\boldsymbol{\theta}}_k) - L(\boldsymbol{\theta}^{*(1)}_{k+1}, \hat{\boldsymbol{\theta}}^{(2)}_k)]$ and $\alpha^{(2)}_k = (1+\gamma)[L(\hat{\boldsymbol{\theta}}_k) - L(\hat{\boldsymbol{\theta}}^{(1)}_k, \boldsymbol{\theta}^{*(2)}_{k+1})]$ we see that the seesaw sequence does indeed satisfy (4a, 4b) of Theorem 3.1. Note, however, that this choice of $\alpha^{(1)}_k$ and $\alpha^{(2)}_k$ merely establishes the *existence* of at least one sequence satisfying (4a, 4b), which is not a practical a priori choice of $\alpha^{(1)}_k$ and $\alpha^{(2)}_k$ because it requires that the values of L during the search be known in advance.

The corollary below shows that pseudoconvexity is sufficient to satisfy the key condition in Theorems 3.1 and 3.2 requiring that it be possible to change one of $\boldsymbol{\theta}^{(1)}$ or $\boldsymbol{\theta}^{(2)}$ to yield a reduction in L at any $\boldsymbol{\theta} \notin \Theta^*$. Let $\boldsymbol{g}^{(m)}(\cdot) = \partial L/\partial \boldsymbol{\theta}^{(m)}$, $m = 1$ or 2. For some conditions relative to the behavior on the boundary of Θ, we need to refer to subvectors of $\boldsymbol{\theta}^{(1)}$ or $\boldsymbol{\theta}^{(2)}$ (sub-subvectors of $\boldsymbol{\theta}$). In particular, it is assumed that there exists a partitioning of each of $\boldsymbol{\theta}^{(m)}$ into distinct sub-subvectors $\boldsymbol{\theta}^{(m;j)}$, $j = 1, 2, \ldots, n(m)$, such that $\boldsymbol{\theta}^{(m)} = (\boldsymbol{\theta}^{(m;1)T}, \ldots, \boldsymbol{\theta}^{(m;n(m))T})^T$ for $m = 1$ or 2. Two important special cases are when the sub-subvectors are the p coordinates of $\boldsymbol{\theta}$ (each subvector of $\boldsymbol{\theta}^{(m)}$ corresponds to one component of $\boldsymbol{\theta}^{(m)}$) and when the sub-subvectors are the full subvectors themselves (i.e., there are two sub-subvectors, each corresponding to a $\boldsymbol{\theta}^{(m)}$). We let $\boldsymbol{\theta}'^{(m;j)}$ and $\boldsymbol{\theta}^{*(m;j)}$ denote the corresponding sub-subvectors of an arbitrary $\boldsymbol{\theta}' \in \Theta$ and of an arbitrary $\boldsymbol{\theta}^* \in \Theta^*$.

Corollary 3.1 *Suppose that Θ is a compact, convex set and that $L(\boldsymbol{\theta})$ is a pseudoconvex function with continuous gradient $\boldsymbol{g}(\boldsymbol{\theta})$ on Θ. Further, suppose that at any $\boldsymbol{\theta}$ on the boundary of Θ, there exists a partitioning of each of $\boldsymbol{\theta}^{(m)}$, $m = 1$ or 2, into distinct sub-subvectors $\boldsymbol{\theta}^{(m;j)}$ (see above) such that it is possible to make a nonzero change in each sub-subvector along the line segment connecting $\boldsymbol{\theta}'^{(m;j)}$ and $\boldsymbol{\theta}^{*(m;j)}$, with other components of $\boldsymbol{\theta}$ held fixed, such that the new point $\boldsymbol{\theta}$ lies in Θ. Then, at any $\boldsymbol{\theta} \in \Theta$ with $\boldsymbol{\theta} \notin \Theta^*$, there exists a change in at least one of $\boldsymbol{\theta}^{(1)}$ or $\boldsymbol{\theta}^{(2)}$ that yields a reduction in L.*

Proof It is sufficient to show that at an arbitrary $\boldsymbol{\theta}' \in \Theta$ with $\boldsymbol{\theta}' \notin \Theta^*$, a change to at least one of $\boldsymbol{\theta}'^{(1)}$ or $\boldsymbol{\theta}'^{(2)}$ yields a reduction in L. Because $L(\boldsymbol{\theta}') > L^*$, it is known by the fundamental property of pseudoconvexity, (3), that $\boldsymbol{g}(\boldsymbol{\theta}')^T(\boldsymbol{\theta}^* - \boldsymbol{\theta}') < 0$ for any $\boldsymbol{\theta}^* \in \Theta^*$. For an arbitrary $\boldsymbol{\theta}^* \in \Theta^*$, this implies $\boldsymbol{g}^{(m)}(\boldsymbol{\theta}')^T(\boldsymbol{\theta}^{*(m)} - \boldsymbol{\theta}'^{(m)}) < 0$ for at least one of $m = 1$ or 2, where $\boldsymbol{\theta}^{*(m)}$ denotes the mth subvector of $\boldsymbol{\theta}^*$. Let us examine the effect on L of changes in the mth subvector of $\boldsymbol{\theta}$.

If $\operatorname{int}(\Theta)$ (the interior of Θ) is nonempty and if $\boldsymbol{\theta}' \in \operatorname{int}(\Theta)$, then both $\boldsymbol{\theta}' \pm \delta \boldsymbol{e}_r \in \Theta$ for all sufficiently small $\delta > 0$, where $\boldsymbol{e}_r$ is a vector with a one in the rth component and zeroes elsewhere. Because $\boldsymbol{g}(\boldsymbol{\theta}')^T(\boldsymbol{\theta}^* - \boldsymbol{\theta}') < 0$ for any $\boldsymbol{\theta}^* \in \Theta^*$, it is known that $g_r(\boldsymbol{\theta}')(t_r^* - t_r') < 0$ for at least one $r \in \{1, 2, \ldots, p\}$, where $g_r(\cdot)$ is the rth component of $\boldsymbol{g}(\cdot)$ and t_r^* and t_r' are the rth components of $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}'$, respectively. For $\boldsymbol{\theta}' \in \operatorname{int}(\Theta)$, it is known by the continuity of $\boldsymbol{g}(\cdot)$ and convexity of Θ that $g_r(\boldsymbol{\theta}' \pm \lambda\delta\boldsymbol{e}_r)(t_r^* - t_r') < 0$ for all $0 \le \lambda \le 1$ and sufficiently small $\delta > 0$. Hence, because $t_r^* \ne t_r'$ at the given r, the mean-value theorem implies that there exist $0 \le \lambda^{(\pm)} \le 1$ and $\delta^{(\pm)} > 0$ such that

$$\begin{aligned} L(\boldsymbol{\theta}') - L\bigl(\boldsymbol{\theta}' + \delta^{(+)}\boldsymbol{e}_r\bigr) &= -g_r\bigl(\boldsymbol{\theta}' + \lambda^{(+)}\delta^{(+)}\boldsymbol{e}_r\bigr)\delta^{(+)} > 0 \quad \text{if } t_r^* > t_r', \\ L(\boldsymbol{\theta}') - L\bigl(\boldsymbol{\theta}' - \delta^{(-)}\boldsymbol{e}_r\bigr) &= g_r\bigl(\boldsymbol{\theta}' - \lambda^{(-)}\delta^{(-)}\boldsymbol{e}_r\bigr)\delta^{(-)} > 0 \quad \text{if } t_r^* < t_r'. \end{aligned} \tag{13}$$

For $\boldsymbol{\theta}'$ on the boundary of Θ, it is known that there exists a partition of each subvector, $\boldsymbol{\theta}'^{(1)}$ and $\boldsymbol{\theta}'^{(2)}$, such that a change in each sub-subvector in the direction of the corresponding sub-subvector of $\boldsymbol{\theta}^*$, with other components of $\boldsymbol{\theta}'$ remaining fixed, produces a new value of $\boldsymbol{\theta}$ that lies in Θ (in contrast to $\boldsymbol{\theta}' \in \operatorname{int}(\Theta)$, it is possible that no $\boldsymbol{\theta}' \pm \delta\boldsymbol{e}_r$ lie in Θ). Because $\boldsymbol{g}^{(m)}(\boldsymbol{\theta}')^T(\boldsymbol{\theta}^{*(m)} - \boldsymbol{\theta}'^{(m)}) < 0$ for at least one of $m = 1$ or 2, it is known that $\boldsymbol{g}^{(m;j)}(\boldsymbol{\theta}')^T(\boldsymbol{\theta}^{*(m;j)} - \boldsymbol{\theta}'^{(m;j)}) < 0$ for at least one sub-subvector, where $\boldsymbol{g}^{(m;j)} = \partial L/\partial\boldsymbol{\theta}^{(m;j)}$. Suppose a change is made to such a sub-subvector along the line segment connecting $\boldsymbol{\theta}'^{(m;j)}$ and $\boldsymbol{\theta}^{*(m;j)}$ with all other components of $\boldsymbol{\theta}$ held at their values in $\boldsymbol{\theta}'$. That is, a change to $\boldsymbol{\theta}'$ is made that is proportional to $\boldsymbol{\Delta}^{(m;j)} := (0, 0, \ldots, 0, (\boldsymbol{\theta}^{*(m;j)} - \boldsymbol{\theta}'^{(m;j)})^T, 0, \ldots, 0)^T$. The mean-value theorem and continuity of $\boldsymbol{g}^{(m;j)}$ imply that there exist a $0 \le \lambda \le 1$ and sufficiently small $\delta > 0$ such that

$$L(\boldsymbol{\theta}') - L\bigl(\boldsymbol{\theta}' + \delta\boldsymbol{\Delta}^{(m;j)}\bigr) = -\delta g^{(m;j)}\bigl(\boldsymbol{\theta}' + \lambda\delta\boldsymbol{\Delta}^{(m;j)}\bigr)^T\bigl(\boldsymbol{\theta}^{*(m;j)} - \boldsymbol{\theta}'^{(m;j)}\bigr) > 0, \tag{14}$$

where the convexity of Θ ensures that both $\boldsymbol{\theta}' + \delta\boldsymbol{\Delta}^{(m;j)}$ and $\boldsymbol{\theta}' + \lambda\delta\boldsymbol{\Delta}^{(m;j)}$ lie in Θ.

Hence, it is possible to change at least one of $\boldsymbol{\theta}^{(1)}$ or $\boldsymbol{\theta}^{(2)}$ to yield a reduction in L at any point outside of Θ^*. □

The above ideas apply directly when the two-stage seesaw process is generalized to an M-stage process, $M \geq 2$. In particular, suppose that there are vectors $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(M)}$, each processed sequentially in the manner of the two-stage algorithm (so $\boldsymbol{\theta}$ is the stacked vector of $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(M)}$). That is, the vectors are processed sequentially such that

$$L(\hat{\boldsymbol{\theta}}_{k+1}) \leq \cdots \leq L\big(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \hat{\boldsymbol{\theta}}_{k+1}^{(2)}, \ldots, \hat{\boldsymbol{\theta}}_{k}^{(M)}\big) \leq L\big(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \hat{\boldsymbol{\theta}}_{k}^{(2)}, \ldots, \hat{\boldsymbol{\theta}}_{k}^{(M)}\big) \leq L(\hat{\boldsymbol{\theta}}_{k}) \tag{15}$$

subject to $\hat{\boldsymbol{\theta}}_{k+1} \neq \hat{\boldsymbol{\theta}}_k$ only if $L(\hat{\boldsymbol{\theta}}_{k+1}) < L(\hat{\boldsymbol{\theta}}_k)$. Then the obvious modifications to the statements of Theorems 3.1 and 3.2 and Corollary 3.1 apply. In particular, Corollaries 3.2 and 3.3 below are direct extensions of Theorems 3.1 and 3.2, respectively, and Corollary 3.4 is an extension of Corollary 3.1 to M-stage pseudoconvex functions. We let the per-iteration minima for each of $\boldsymbol{\theta}^{(m)}$, $m = 1, 2, \ldots, M$, be denoted by $\boldsymbol{\theta}_k^{*(m)}$. That is,

$$\begin{aligned} &L\big(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \ldots, \hat{\boldsymbol{\theta}}_{k+1}^{(m-1)}, \boldsymbol{\theta}_{k+1}^{*(m)}, \hat{\boldsymbol{\theta}}_{k}^{(m+1)}, \ldots, \hat{\boldsymbol{\theta}}_{k}^{(M)}\big) \\ &\quad \leq L\big(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \ldots, \hat{\boldsymbol{\theta}}_{k+1}^{(m-1)}, \boldsymbol{\theta}^{(m)}, \hat{\boldsymbol{\theta}}_{k}^{(m+1)}, \ldots, \hat{\boldsymbol{\theta}}_{k}^{(M)}\big) \end{aligned}$$

for all $\boldsymbol{\theta}^{(m)}$. So, $\boldsymbol{\theta}_{k+1}^{*(m)}$ is a function of $\hat{\boldsymbol{\theta}}_{k+1}^{(j)}$, $j < m$, and $\hat{\boldsymbol{\theta}}_{k}^{(j)}$, $j > m$.

Corollary 3.2 *Consider an M-stage estimation process. Suppose that Θ is a compact, convex set and that $L(\boldsymbol{\theta})$ is continuous on Θ. Further, suppose that at any $\boldsymbol{\theta} \in \Theta$ with $\boldsymbol{\theta} \notin \Theta^*$, L is such that it is possible to change at least one of $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(M)}$ to yield a reduction in L and that the M-stage algorithm reduces L with respect to at least one of the $\boldsymbol{\theta}^{(m)}$ at each iteration in the sense that (16) holds for at least one $1 \leq m \leq M$ for each $k = 0, 1, 2, \ldots$:*

$$\begin{aligned} &L\big(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \ldots, \hat{\boldsymbol{\theta}}_{k+1}^{(m-1)}, \hat{\boldsymbol{\theta}}_{k+1}^{(m)}, \hat{\boldsymbol{\theta}}_{k}^{(m+1)}, \ldots, \hat{\boldsymbol{\theta}}_{k}^{(M)}\big) \\ &\quad \leq L\big(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \ldots, \hat{\boldsymbol{\theta}}_{k+1}^{(m-1)}, \boldsymbol{\theta}_{k+1}^{*(m)}, \hat{\boldsymbol{\theta}}_{k}^{(m+1)}, \ldots, \hat{\boldsymbol{\theta}}_{k}^{(M)}\big) + \alpha_k^{(m)} \end{aligned} \tag{16}$$

where $\alpha_k^{(m)}$ is nonnegative and converges to 0. Then $L(\hat{\boldsymbol{\theta}}_k) \to L^*$ as $k \to \infty$. Further if $\boldsymbol{\theta}^*$ is unique (i.e., Θ^* is the singleton $\boldsymbol{\theta}^*$), then $\hat{\boldsymbol{\theta}}_k \to \boldsymbol{\theta}^*$ as $k \to \infty$.

Remark 3.3 By the reduction property in (1) and (2), restrictions analogous to those for $M = 2$ in Remark 3.2 apply to the $\alpha_k^{(m)}$.

Proof This result follows very closely along the lines of the proof of Theorem 3.1 when making the obvious modification from $M = 2$, as in Theorem 3.1, to general M. First, we have an L' such that $\lim_{k\to\infty} L(\hat{\boldsymbol{\theta}}_k) = L'$. Then it is known that $L' = L^*$ by demonstrating that a contradiction results when $L' > L^*$.

Consider the point $\boldsymbol{\theta}'$, where $\boldsymbol{\theta}' \in \Theta$ is the limiting point of a subsequence such that $L(\boldsymbol{\theta}') = L'$. We have that $\boldsymbol{\theta}'$ is apportioned according to $\boldsymbol{\theta}'^{(1)}, \boldsymbol{\theta}'^{(2)}, \ldots, \boldsymbol{\theta}'^{(M)}$. Because $L(\boldsymbol{\theta}') > L^*$, it is known by assumption that a change in at least one subvector $\boldsymbol{\theta}'^{(m)}$ will reduce L. Suppose that a change applied to $\boldsymbol{\theta}'^{(1)}$ reduces L. Then the

guaranteed reduction property of the algorithm in (16) and the arguments analogous to those in the proof of Theorem 3.1 show that for any $j \geq N(\varepsilon_0)$ and some N, we get the following expression similar to (9),

$$\begin{aligned} L\big(\hat{\boldsymbol{\theta}}_{k_j+1}^{(1)}, \hat{\boldsymbol{\theta}}_{k_j}^{(2)}, \ldots, \hat{\boldsymbol{\theta}}_{k_j}^{(M)}\big) &\leq L\big(\boldsymbol{\theta}_{k_j+1}^{*(1)}, \hat{\boldsymbol{\theta}}_{k_j}^{(2)}, \ldots, \hat{\boldsymbol{\theta}}_{k_j}^{(M)}\big) + \alpha_{k_j}^{(1)} \\ &\leq L' - \varepsilon + \varepsilon_0 + \alpha_{k_j}^{(1)}, \end{aligned}$$

where $0 < \varepsilon_0 < \varepsilon$. Taking $j \geq N(\varepsilon_0)$ sufficiently large so that $-\varepsilon + \varepsilon_0 + \alpha_{k_j}^{(1)} < 0$, we arrive at the contradiction $L(\hat{\boldsymbol{\theta}}_{k_j+1}) < L'$. The monotonicity of (15), therefore, leads to a further contradiction with the result $\lim_{k\to\infty} L(\hat{\boldsymbol{\theta}}_k) = L' > L(\boldsymbol{\theta}^*)$. Based on assumption (16), analogous reasoning applies for a change applied to $\boldsymbol{\theta}'^{(m)}$, $2 \leq m \leq M$, which reduces L. Hence, it is known that $L(\hat{\boldsymbol{\theta}}_k) \to L^*$ as $k \to \infty$. Further, the arguments in the last paragraph of the proof of Theorem 3.1 carry over directly to show that $\hat{\boldsymbol{\theta}}_k \to \boldsymbol{\theta}^*$ as $k \to \infty$. □

Corollary 3.3 *Consider an M-stage estimation process. Suppose that the conditions of Corollary 3.2 hold except that, at each iteration k, the following replacement to (16) holds for at least one $1 \leq m \leq M$,*

$$\frac{L(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \ldots, \hat{\boldsymbol{\theta}}_{k+1}^{(m-1)}, \hat{\boldsymbol{\theta}}_{k}^{(m)}, \hat{\boldsymbol{\theta}}_{k}^{(m+1)}, \ldots, \hat{\boldsymbol{\theta}}_{k}^{(M)}) - L(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \ldots, \hat{\boldsymbol{\theta}}_{k+1}^{(m-1)}, \hat{\boldsymbol{\theta}}_{k+1}^{(m)}, \hat{\boldsymbol{\theta}}_{k}^{(m+1)}, \ldots, \hat{\boldsymbol{\theta}}_{k}^{(M)})}{L(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \ldots, \hat{\boldsymbol{\theta}}_{k+1}^{(m-1)}, \hat{\boldsymbol{\theta}}_{k}^{(m)}, \hat{\boldsymbol{\theta}}_{k}^{(m+1)}, \ldots, \hat{\boldsymbol{\theta}}_{k}^{(M)}) - L(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \ldots, \hat{\boldsymbol{\theta}}_{k+1}^{(m-1)}, \boldsymbol{\theta}_{k+1}^{*(m)}, \hat{\boldsymbol{\theta}}_{k}^{(m+1)}, \ldots, \hat{\boldsymbol{\theta}}_{k}^{(M)})} \geq \gamma,$$

where $0 < \gamma \leq 1$. Then $L(\hat{\boldsymbol{\theta}}_k) \to L^*$ as $k \to \infty$. Further if $\boldsymbol{\theta}^*$ is unique (i.e., Θ^* is the singleton $\boldsymbol{\theta}^*$), then $\hat{\boldsymbol{\theta}}_k \to \boldsymbol{\theta}^*$ as $k \to \infty$.

Proof The proof is immediate following the proof of Theorem 3.2 when making the obvious modification from $M = 2$ to general $M \geq 2$. □

Corollary 3.4 *Suppose that Θ is a compact, convex set and that $L(\boldsymbol{\theta})$ is a pseudo-convex function with continuous gradient $\boldsymbol{g}(\boldsymbol{\theta})$ on Θ. Further, suppose that at any $\boldsymbol{\theta}$ on the boundary of Θ, there exists a partitioning of each subvector $\boldsymbol{\theta}^{(m)}$, $m = 1, 2, \ldots, M$, into distinct sub-subvectors $\boldsymbol{\theta}^{(m;j)}$, $j = 1, 2, \ldots, n(m)$ (see above), such that it is possible to make a nonzero change in each sub-subvector along the line segment connecting $\boldsymbol{\theta}'^{(m;j)}$ and $\boldsymbol{\theta}^{*(m;j)}$, with other components of $\boldsymbol{\theta}$ held fixed, such that the new point $\boldsymbol{\theta}$ lies in Θ. Then, at any $\boldsymbol{\theta} \in \Theta$ with $\boldsymbol{\theta} \notin \Theta^*$, there exists a change in at least one of $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(M)}$ that yields a reduction in L.*

Proof The proof closely follows the proof of Corollary 3.1. It is sufficient to show that at an arbitrary $\boldsymbol{\theta}' \in \Theta$ with $\boldsymbol{\theta}' \notin \Theta^*$, a change to at least one of $\boldsymbol{\theta}'^{(1)}, \boldsymbol{\theta}'^{(2)}, \ldots,$ or $\boldsymbol{\theta}'^{(M)}$ yields a reduction in L. Because $L(\boldsymbol{\theta}') > L^*$, it is known by (3) that $\boldsymbol{g}(\boldsymbol{\theta}')^T(\boldsymbol{\theta}^* - \boldsymbol{\theta}') < 0$ for any $\boldsymbol{\theta}^* \in \Theta^*$. For $\boldsymbol{\theta}' \in \text{int}(\Theta)$ and an arbitrary $\boldsymbol{\theta}^* \in \Theta^*$, this implies $g_r(\boldsymbol{\theta}')(t_r^* - t_r') < 0$ for at least one $r \in \{1, 2, \ldots, p\}$. The mean-value arguments associated with (13) of Corollary 3.1 now follow directly since the arguments are not limited to the $M = 2$ case. For $\boldsymbol{\theta}'$ on the boundary of Θ, it is known by assumption that there exists a partition of each subvector, $\boldsymbol{\theta}'^{(m)}$, such that a change

in each sub-subvector in the direction of the corresponding sub-subvector of $\boldsymbol{\theta}^*$, with other components of $\boldsymbol{\theta}'$ remaining fixed, produces a new value of $\boldsymbol{\theta}$ that lies in Θ. As in the case above of $\boldsymbol{\theta}'$ not on the boundary, the result to be proved then follows using the mean-value arguments of (14) since the arguments are not limited to the $M = 2$ case. □

3.3 Comments on Theorem Conditions for Strictly Convex Loss Functions, Including Use with Steepest Descent Method

Let us sketch an argument to show that the key lower bound conditions (11a, 11b) in Theorem 3.2 are reasonable for the special case of continuously twice differentiable strictly convex functions (recall that the theorem conditions only involve continuity, not differentiability). The arguments here, which are "local" in the sense of being based on Taylor expansions, are suggestive of the reasonableness of the conditions (11a, 11b), but they do not show formally either the necessity or sufficiency of strict convexity for (11a, 11b). Nevertheless, for small changes in the subvectors, the arguments indicate that (11a, 11b) apply for certain types of algorithms, including the steepest (gradient) descent algorithm, when applied to strictly convex loss functions.

Let us consider (11a) (arguments for (11b) follow in the same manner). Expanding the numerator of the ratio in (11a) around $\hat{\boldsymbol{\theta}}_{k+1}^{(1)}$ and the denominator around $\boldsymbol{\theta}_{k+1}^{*(1)}$, we find that the relevant ratio satisfies

$$\begin{aligned}
&\frac{L(\hat{\boldsymbol{\theta}}_k) - L(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \hat{\boldsymbol{\theta}}_k^{(2)})}{L(\hat{\boldsymbol{\theta}}_k) - L(\boldsymbol{\theta}_{k+1}^{*(1)}, \hat{\boldsymbol{\theta}}_k^{(2)})} \\
&\quad = \frac{\boldsymbol{g}^{(1)}(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \hat{\boldsymbol{\theta}}_k^{(2)})^T (\hat{\boldsymbol{\theta}}_k^{(1)} - \hat{\boldsymbol{\theta}}_{k+1}^{(1)}) + \frac{1}{2}(\hat{\boldsymbol{\theta}}_k^{(1)} - \hat{\boldsymbol{\theta}}_{k+1}^{(1)})^T \bar{\boldsymbol{H}}^{(1)} (\hat{\boldsymbol{\theta}}_k^{(1)} - \hat{\boldsymbol{\theta}}_{k+1}^{(1)})}{\frac{1}{2}(\hat{\boldsymbol{\theta}}_k^{(1)} - \boldsymbol{\theta}_{k+1}^{*(1)})^T \bar{\bar{\boldsymbol{H}}}^{(1)} (\hat{\boldsymbol{\theta}}_k^{(1)} - \boldsymbol{\theta}_{k+1}^{*(1)})} \\
&\quad \approx \frac{(\hat{\boldsymbol{\theta}}_k^{(1)} - \hat{\boldsymbol{\theta}}_{k+1}^{(1)})^T \bar{\boldsymbol{H}}^{(1)} (\hat{\boldsymbol{\theta}}_k^{(1)} - \hat{\boldsymbol{\theta}}_{k+1}^{(1)})}{(\hat{\boldsymbol{\theta}}_k^{(1)} - \boldsymbol{\theta}_{k+1}^{*(1)})^T \bar{\bar{\boldsymbol{H}}}^{(1)} (\hat{\boldsymbol{\theta}}_k^{(1)} - \boldsymbol{\theta}_{k+1}^{*(1)})} \\
&\quad \approx \frac{(\hat{\boldsymbol{\theta}}_k^{(1)} - \hat{\boldsymbol{\theta}}_{k+1}^{(1)})^T \bar{\boldsymbol{H}}^{(1)} (\hat{\boldsymbol{\theta}}_k^{(1)} - \hat{\boldsymbol{\theta}}_{k+1}^{(1)})}{(\hat{\boldsymbol{\theta}}_k^{(1)} - \boldsymbol{\theta}_{k+1}^{*(1)})^T \bar{\boldsymbol{H}}^{(1)} (\hat{\boldsymbol{\theta}}_k^{(1)} - \boldsymbol{\theta}_{k+1}^{*(1)})},
\end{aligned} \tag{17}$$

where $\bar{\boldsymbol{H}}^{(1)}$ and $\bar{\bar{\boldsymbol{H}}}^{(1)}$ represent the upper left blocks, corresponding to $\boldsymbol{\theta}^{(1)}$, of the Hessian matrix of L evaluated at intermediate points between $\hat{\boldsymbol{\theta}}_{k+1}^{(1)}$ and $\hat{\boldsymbol{\theta}}_k^{(1)}$ and between $\boldsymbol{\theta}_{k+1}^{*(1)}$ and $\hat{\boldsymbol{\theta}}_k^{(1)}$, respectively, and at $\hat{\boldsymbol{\theta}}_k^{(2)}$. By the strict convexity of L, it is known that both $\bar{\boldsymbol{H}}^{(1)}$ and $\bar{\bar{\boldsymbol{H}}}^{(1)}$ are positive definite. The "$\approx$" in the second line above follows from $\boldsymbol{g}^{(1)}(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \hat{\boldsymbol{\theta}}_k^{(2)}) \approx \mathbf{0}$ when $\hat{\boldsymbol{\theta}}_{k+1}^{(1)}$ is close to $\boldsymbol{\theta}_{k+1}^{*(1)}$. The "$\approx$" in the third line follows from assuming that the upper left block of the Hessian matrix is approximately constant in the neighborhood of $\hat{\boldsymbol{\theta}}_k^{(1)}$ that includes both $\hat{\boldsymbol{\theta}}_{k+1}^{(1)}$ and $\boldsymbol{\theta}_{k+1}^{*(1)}$ (conditioned on $\hat{\boldsymbol{\theta}}_k^{(2)}$). Using the standard Euclidean vector norm, the compatible matrix spectral norm (i.e., maximum eigenvalue since the matrix is positive definite),

and the symmetric form of matrix square root, we find that the ratio on the right-hand side of (17) satisfies

$$\frac{(\hat{\boldsymbol{\theta}}_k^{(1)}-\hat{\boldsymbol{\theta}}_{k+1}^{(1)})^T\bar{\boldsymbol{H}}^{(1)}(\hat{\boldsymbol{\theta}}_k^{(1)}-\hat{\boldsymbol{\theta}}_{k+1}^{(1)})}{(\hat{\boldsymbol{\theta}}_k^{(1)}-\boldsymbol{\theta}_{k+1}^{*(1)})^T\bar{\boldsymbol{H}}^{(1)}(\hat{\boldsymbol{\theta}}_k^{(1)}-\boldsymbol{\theta}_{k+1}^{*(1)})}\geq\frac{\|(\bar{\boldsymbol{H}}^{(1)})^{1/2}(\hat{\boldsymbol{\theta}}_k^{(1)}-\hat{\boldsymbol{\theta}}_{k+1}^{(1)})\|^2}{\|\bar{\boldsymbol{H}}^{(1)}\|\|\hat{\boldsymbol{\theta}}_k^{(1)}-\boldsymbol{\theta}_{k+1}^{*(1)}\|^2}$$
$$\geq\frac{\|\hat{\boldsymbol{\theta}}_k^{(1)}-\hat{\boldsymbol{\theta}}_{k+1}^{(1)}\|^2}{\|\bar{\boldsymbol{H}}^{(1)}\|\|(\bar{\boldsymbol{H}}^{(1)})^{-1}\|\|\hat{\boldsymbol{\theta}}_k^{(1)}-\boldsymbol{\theta}_{k+1}^{*(1)}\|^2}. \quad (18)$$

Hence, to the extent that the approximation on the right-hand side of (17) is an accurate representation of the ratio in (11a), we know from (18) that the ratio in (11a) is bounded below by a term proportional to the ratio of squared distances $\|\hat{\boldsymbol{\theta}}_k^{(1)}-\hat{\boldsymbol{\theta}}_{k+1}^{(1)}\|^2/\|\hat{\boldsymbol{\theta}}_k^{(1)}-\boldsymbol{\theta}_{k+1}^{*(1)}\|^2$. That is, as long as $\hat{\boldsymbol{\theta}}_{k+1}^{(1)}$ is generated such that the distance between $\hat{\boldsymbol{\theta}}_{k+1}^{(1)}$ and $\hat{\boldsymbol{\theta}}_k^{(1)}$ is at least some consistent fraction of the distance between $\boldsymbol{\theta}_{k+1}^{*(1)}$ and $\hat{\boldsymbol{\theta}}_k^{(1)}$, so that the ratio of squared distances is bounded below by a strictly positive constant, then the lower bound condition $\gamma>0$ in (11a) is satisfied. Analogous reasoning applies to (11b).

As an example of how the ratio of squared distances in the right-hand side of (18) may be bounded below by a strictly positive constant, consider a seesaw version of the steepest descent method, $\hat{\boldsymbol{\theta}}_{k+1}^{(i)}=\hat{\boldsymbol{\theta}}_k^{(i)}-a_k^{(i)}\boldsymbol{g}_k^{(i)}$, $k=0,1,2,\ldots$, $i=1$ or 2, $a_k^{(i)}$ is a nonnegative (scalar) gain number ($a_k^{(i)}$ may be chosen as a constant, as a prespecified decaying sequence, or adaptively via a line search), and $\boldsymbol{g}_k^{(i)}$ is the gradient $\boldsymbol{g}^{(1)}(\hat{\boldsymbol{\theta}}_k)$ or $\boldsymbol{g}^{(2)}(\hat{\boldsymbol{\theta}}_{k+1}^{(1)},\hat{\boldsymbol{\theta}}_k^{(2)})$, as appropriate. If $a_k^{(i)}\geq a>0$ for all k, then, under the assumptions associated with (17),

$$\frac{\|\hat{\boldsymbol{\theta}}_k^{(1)}-\hat{\boldsymbol{\theta}}_{k+1}^{(1)}\|^2}{\|\hat{\boldsymbol{\theta}}_k^{(1)}-\boldsymbol{\theta}_{k+1}^{*(1)}\|^2}\geq\frac{a^2\|\boldsymbol{g}^{(1)}(\hat{\boldsymbol{\theta}}_k)\|^2}{\|\hat{\boldsymbol{\theta}}_k^{(1)}-\boldsymbol{\theta}_{k+1}^{*(1)}\|^2}\approx\frac{a^2\|\bar{\boldsymbol{H}}^{(1)}(\hat{\boldsymbol{\theta}}_k^{(1)}-\boldsymbol{\theta}_{k+1}^{*(1)})\|^2}{\|\hat{\boldsymbol{\theta}}_k^{(1)}-\boldsymbol{\theta}_{k+1}^{*(1)}\|^2}\geq\frac{a^2}{\|(\bar{\boldsymbol{H}}^{(1)})^{-1}\|^2},$$

where the "$\approx$" follows from an expansion of $\boldsymbol{g}^{(1)}$ around the point $(\boldsymbol{\theta}_{k+1}^{*(1)T},\hat{\boldsymbol{\theta}}_k^{(2)T})^T$ and the last inequality follows from:

$$\|\hat{\boldsymbol{\theta}}_k^{(1)}-\boldsymbol{\theta}_{k+1}^{*(1)}\|=\|(\bar{\boldsymbol{H}}^{(1)})^{-1}\bar{\boldsymbol{H}}^{(1)}(\hat{\boldsymbol{\theta}}_k^{(1)}-\boldsymbol{\theta}_{k+1}^{*(1)})\|\leq\|(\bar{\boldsymbol{H}}^{(1)})^{-1}\|\|\bar{\boldsymbol{H}}^{(1)}(\hat{\boldsymbol{\theta}}_k^{(1)}-\boldsymbol{\theta}_{k+1}^{*(1)})\|.$$

Hence, to the extent the local analysis associated with (17) and (18) apply, we know that (11a) holds with γ chosen less than $a^2/(\|\bar{\boldsymbol{H}}^{(1)}\|\|(\bar{\boldsymbol{H}}^{(1)})^{-1}\|^3)$. Likewise, (11b) holds with γ chosen less than $a^2/(\|\bar{\boldsymbol{H}}^{(2)}\|\|(\bar{\boldsymbol{H}}^{(2)})^{-1}\|^3)$, where $\bar{\boldsymbol{H}}^{(2)}$ represents the lower right block, corresponding to $\boldsymbol{\theta}^{(2)}$, of the Hessian matrix of L evaluated at $\hat{\boldsymbol{\theta}}_{k+1}^{(1)}$ and at an intermediate point between $\hat{\boldsymbol{\theta}}_{k+1}^{(2)}$ and $\hat{\boldsymbol{\theta}}_k^{(2)}$. Of course, following the pattern in (17), we assume that the Hessian is approximately constant in the neighborhood of $\hat{\boldsymbol{\theta}}_k^{(2)}$ that includes both $\hat{\boldsymbol{\theta}}_{k+1}^{(2)}$ and $\boldsymbol{\theta}_{k+1}^{*(2)}$ (conditioned on $\hat{\boldsymbol{\theta}}_{k+1}^{(1)}$). Thus, in practice, both $\bar{\boldsymbol{H}}^{(1)}$ and $\bar{\boldsymbol{H}}^{(2)}$ may be evaluated at the known point $\hat{\boldsymbol{\theta}}_k$ (versus unknown intermediate points) in calculating γ. Therefore, a γ can be chosen such that

$$0<\gamma\leq\min\{1,a^2/(\|\bar{\boldsymbol{H}}^{(1)}\|\|(\bar{\boldsymbol{H}}^{(1)})^{-1}\|^3),a^2/(\|\bar{\boldsymbol{H}}^{(2)}\|\|(\bar{\boldsymbol{H}}^{(2)})^{-1}\|^3)\}\leq 1, \quad (19)$$

implying that steepest descent, when applied to continuously twice differentiable strictly convex loss functions, satisfies (11a, 11b) in Theorem 3.2. Because of the local approximations used in the derivations above, the conclusions and bound for γ are most appropriate when operating near $\boldsymbol{\theta}^*$.

4 Summary of Example in State–Space Model Identification

As mentioned in Sect. 1, a motivating application for the seesaw approach is a problem in the identification of parameters in state–space models. This section is a brief discussion of the issues in the motivating application. It is assumed that the process is modeled according to the traditional linear state–space model composed of a state equation and a measurement equation. We observe N independent realizations of the process (i.e., N independent tests). Such cross-sectional identification problems for state–space models have been considered in a number of references, including Goodrich and Caines [22], Shumway et al. [1], and Levy [19]. Each realization is associated with its own state–space model, but $\boldsymbol{\theta}$ is, in general, common across the N models.

For the identification of the defense system of interest to the author, the original focus and software development was aimed at the common mean vector and covariance matrix, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, for the initial states in the state–space model. Later, the interest extended to include power spectral density parameters entering the state-noise covariance matrix. We summarize below the essential aspects of the identification. Greater detail on this state–space implementation is provided in Spall [23].

A standard method for optimization is a gradient-based recursion, which uses the gradient of the log-likelihood function with respect to $\boldsymbol{\theta}$. This form of recursion provides the foundation for many system identification methods, as applied to state–space models (e.g., Goodrich and Caines [22]; Levy [19], and Wills and Ninness [24]). In turn, to compute the gradient of the log-likelihood, it is necessary to compute the gradient of the state–space model terms with respect to $\boldsymbol{\theta}$ (e.g., Goodrich and Caines [22], Segal and Weinstein [25], and Levy [19]). That is, via the Kalman filter and/or smoother, the gradient of the log-likelihood (direct or EM-based) is computable via recursive calculations of the gradient with respect to the relevant terms within the state–space model. Let us consider here the special case of $\boldsymbol{\theta}$ representing the parameters that enter initial state mean and covariance matrix and the state-noise covariance matrices.

Consider first the $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ part of the overall parameter vector $\boldsymbol{\theta}$; these parameters correspond to $\boldsymbol{\theta}^{(1)}$ in the general formulation of Sect. 2. Let us assume that $\boldsymbol{x}_0^{(i)} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{x}_0^{(i)}$ represents the initial state for the ith realization. Let $\boldsymbol{Z}^{(i)} := ((\boldsymbol{z}_0^{(i)})^T, (\boldsymbol{z}_1^{(i)})^T, \ldots, (\boldsymbol{z}_{n_i}^{(i)})^T)^T$ represent the vector of stacked measurements for the ith realization. Then

$$\boldsymbol{Z}^{(i)} = \boldsymbol{C}^{(i)} \boldsymbol{x}_0^{(i)} + \boldsymbol{V}^{(i)}, \tag{20}$$

where $\boldsymbol{C}^{(i)}$ is derived from the measurement and transition matrices of the underlying state–space model and $\boldsymbol{V}^{(i)}$ is derived from the measurement and transition matrices as well as the process and measurement noises. Let $\boldsymbol{\Gamma}^{(i)} = \mathrm{cov}(\boldsymbol{V}^{(i)})$.

If the quantity $\boldsymbol{P}^{(i)} \equiv [(\boldsymbol{C}^{(i)})^T(\boldsymbol{\Gamma}^{(i)})^{-1}\boldsymbol{C}^{(i)}]^{-1}$ exists, each estimate of $\boldsymbol{x}_0^{(i)}$, say $\hat{\boldsymbol{x}}_0^{(i)}$, is defined according to the standard weighted least-squares formula applied to (20):

$$\hat{\boldsymbol{x}}_0^{(i)} = \boldsymbol{P}^{(i)}\big(\boldsymbol{C}^{(i)}\big)^T\big(\boldsymbol{\Gamma}^{(i)}\big)^{-1}\boldsymbol{Z}^{(i)}. \tag{21}$$

As described in Shumway et al. [1], the collection of the above least-squares estimates for $i = 1, 2, \ldots, N$ form a set of sufficient statistics for estimating $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The maximum likelihood estimate for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can then be formed by maximizing the log-likelihood function associated with the independent "data" (sufficient statistics) $\{\hat{\boldsymbol{x}}_0^{(1)}, \hat{\boldsymbol{x}}_0^{(2)}, \ldots, \hat{\boldsymbol{x}}_0^{(N)}\}$; each of these sufficient statistics is normally distributed with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} + \boldsymbol{P}^{(i)}$.

The above approach relies on the existence of the least-squares-based sufficient statistics $\hat{\boldsymbol{x}}_0^{(i)}$ shown in (21). In practical applications, it may be the case that the $\boldsymbol{P}^{(i)}$ do not exist as a consequence of poor observability represented in the $\boldsymbol{C}^{(i)}$ matrices. Hence, the least-squares solution in (21) does not exist. Using results in Rao [26, p. 231], Spall [23] summarizes a maximum likelihood formulation not dependent on the solution in (21), rather using only generalized inverses. Standard numerical methods are used to determine the maximum likelihood estimate for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, although coping with constraints (including the required positive definiteness of $\boldsymbol{\Sigma}$) requires special care; this topic is outside the scope of this paper. The search process for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is the first stage of the seesaw process at each iteration.

We now focus on the second stage of the seesaw process (i.e., that associated with $\boldsymbol{\theta}^{(2)}$). The parameters $\boldsymbol{\theta}^{(2)}$ are associated with the state noise. Consider one realization of the process. In practice, the calculations are carried out for each of the N realizations. For the chosen realization, let $\boldsymbol{Q}_t$ represent the covariance matrix for the state noise at time index t in the *discrete-time* state–space model and $\boldsymbol{\psi}$ represent the vector of unknown parameters (e.g., the power spectral density parameters) associated with the underlying continuous time process noise as represented in a differential equation. The parameters $\boldsymbol{\psi}$ correspond to $\boldsymbol{\theta}^{(2)}$. Note that $\boldsymbol{Q}_t$ is associated with the discrete-time state transition from time t to time $t+1$ (say, from τ_t to $\tau_t + 1$). The parameters in $\boldsymbol{\psi}$ manifest themselves in the discrete-time matrices $\boldsymbol{Q}_t$ based on the connection of the continuous-time dynamics to the discrete-time dynamics.

Then, given a state transition matrix $\boldsymbol{\Phi}(\tau, s)$ found using standard linear systems methods (e.g., Moon and Stirling [27, pp. 20–21]), the derivative of the discrete-time covariance matrix with respect to the mth component of $\boldsymbol{\psi}$, say ψ_m, is

$$\frac{\partial \boldsymbol{Q}_t}{\partial \psi_m} = \int_{\tau_t}^{\tau_{t+1}} \boldsymbol{\Phi}(\tau_{t+1}, \tau)\frac{\partial \boldsymbol{\Omega}(\tau)}{\partial \psi_m}\boldsymbol{\Phi}(\tau_{t+1}, \tau)^T\, d\tau,$$

where $\boldsymbol{\Omega}(\tau)$ is the power spectral density matrix of the continuous time process noise. The above gradient is used as part of the chain rule in forming the gradient of the log-likelihood function as part of the second part of the seesaw process according to the references mentioned above.

A different state–space identification application involving a natural decoupling into two groups of parameters is described in Spall and Garner [28] in the context

of primary parameters and nuisance parameters. The analysis is based on $N = 1$ (i.e., a single realization). The seesaw idea could be used in the nuisance parameter context if the aim was to estimate both primary and nuisance parameters from a given set of data. (The Spall and Garner paper considers only the estimation of the primary parameters, taking the nuisance parameters as "given" based on prior information.)

5 Numerical Analysis

5.1 Overview

In this section, we present a numerical analysis of the seesaw method for three test functions. The test functions are simple functions not directly related to the state–space problem of Sect. 4. These functions are chosen, rather than the state–space-based log-likelihood, because there are many ancillary issues in the real-world implementation of state–space identification that are not pertinent to the generic understanding of the seesaw method (e.g., square-root information forms of Kalman filters and constraints for the covariance matrix part $\boldsymbol{\Sigma}$ of $\boldsymbol{\theta}^{(1)}$).

Although seesaw is not tied to any specific numerical algorithm, we use the steepest (gradient) descent method as the basis of the studies here. The standard form of steepest descent (no seesaw) is

$$\hat{\boldsymbol{\theta}}_{k+1} = \hat{\boldsymbol{\theta}}_k - a_k \boldsymbol{g}(\hat{\boldsymbol{\theta}}_k), \quad k = 0, 1, 2, \ldots, \tag{22}$$

where a_k is a nonnegative (scalar) gain number satisfying certain conditions. (The obvious modification of (22) for seesaw is mentioned in Sect. 3.3.) For example, when L is quadratic with positive definite Hessian matrix and when constant gains $a_k = a$ are used for all k, convergence from any starting point to $\boldsymbol{\theta}^*$ requires $0 < a < 2/\lambda_{\max}$, where $\lambda_{\max}$ is the maximum eigenvalue of the Hessian matrix (Moon and Stirling [27, pp. 639–641]). Other conditions for convergence are available for more general L and/or for nonconstant gains, such as when line searches are used (e.g., Bazaraa et al. [4, Sect. 8.6] and Polak [16, Sect. 1.3]). We use constant gains $a_k = a$ for all k in the studies here. As a check on whether the gain values in the studies are reasonable relative to convergence to $\boldsymbol{\theta}^*$, the implementations of standard steepest descent below use gain values that satisfy the above inequality, where $\lambda_{\max}$ is the maximum eigenvalue of the Hessian matrix at $\boldsymbol{\theta}^*$. Hence, the check is based on a quadratic approximation to the functions being optimized, which, in fact, are not quadratic. Finally, as needed below, the generic representation of the components of $\boldsymbol{\theta}$ is according to $\boldsymbol{\theta} := (t_1, t_2, \ldots, t_p)^T$.

Although the steepest descent method is not likely to be the best algorithm for minimizing any of the functions below, we use it in these studies because it is a foundational method having broad applicability and *reasonable* performance in a range of problems. Further, steepest descent represents a special case of stochastic gradient methods (a.k.a. Robbins–Monro stochastic approximation) (e.g. Spall [17, Chaps. 4 and 5]). Hence, the performance improvement observed here might point to possible improvements in a stochastic environment, as well.

5.2 Simple Quartic Loss Function

The first test case in this study is the simple quartic loss function in Spall [17, Example 1.8], $L(\boldsymbol{\theta}) = t_1^4 + t_1^2 + t_1 t_2 + t_2^2$ with $\Theta = \mathbb{R}^2$. It is easily seen that the global minimum $\boldsymbol{\theta}^* = (0,0)^T$ is the only critical point. We compare the steepest descent method with the seesaw method under a common fixed gain coefficient (step size). The subvectors $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$ here correspond to the two scalar components of $\boldsymbol{\theta}$. We use both a standard steepest descent (22) and a modified steepest descent that exploits a known closed-form solution. In particular, the modified form uses standard steepest descent for the update of t_1 while the closed-form solution $t_2 = -t_1/2$, found by solving the equation $\boldsymbol{g}(\boldsymbol{\theta}) = \boldsymbol{0}$ for $\boldsymbol{\theta}$, is used when updating t_2 (i.e., in the nonseesaw approach, t_2 is updated from the value of t_1 in the previous iteration; in seesaw, t_2 is updated using the value of t_1 in the most recent subiteration).

It is easily seen that L is strictly convex on Θ (positive definite Hessian matrix on Θ). Hence, the lower bound analysis in Sect. 3.3 indicates that the iterate $\hat{\boldsymbol{\theta}}_k$ produced by steepest descent with seesaw converges to $\boldsymbol{\theta}^*$ subject to $\hat{\boldsymbol{\theta}}_k$ staying in the neighborhood of $\boldsymbol{\theta}^*$ for all k, The neighborhood requirement ensures the validity of the local analysis in Sect. 3.3. That is, from (19), we may indentify a value γ satisfying conditions (11a, 11b), thereby ensuring convergence from Theorem 3.2 since the other conditions are automatically satisfied. More broadly, without imposing the local neighborhood restrictions of the analysis in Sect. 3.3, we may consider the application of Corollary 3.1. Although we do not impose explicit constraints on $\boldsymbol{\theta}$, the practical limits of computation operate as if this is a constrained problem. Hence, in practice, with Θ reflecting a hypercube constraint on $\boldsymbol{\theta}$ (i.e., each component of $\boldsymbol{\theta}$ has a fixed lower and upper bound), the conditions of Corollary 3.1 related to pseudoconvexity are satisfied, indicating that it is possible to change at least one of $\boldsymbol{\theta}^{(1)}$ or $\boldsymbol{\theta}^{(2)}$ to yield a reduction in L (i.e., the main condition of Theorems 3.1 and 3.2 is satisfied).

Table 1 compares the performance of steepest descent and seesaw in terms of the error in $\boldsymbol{\theta}$ for two gain values. Each entry in the table is based on $k \leq 50$ iterations using the initial condition $\hat{\boldsymbol{\theta}}_0 = (1,1)^T$. The table gives results for two gain values, $a = 0.15$ and $a = 0.29$. Basic steepest descent (22) is used with the conservative gain $a = 0.15$. The modified steepest descent discussed above is used with the more aggressive (larger) gain, $a = 0.29$. The larger gain provides for faster convergence, but it is close to causing unstable behavior in the algorithm ($a \geq 0.30$ leads to divergence). All of the gain values are well under the above-mentioned upper bound to a, $2/\lambda_{\max} = 2/3$, based on a quadratic approximation to L.

Table 1 indicates that the seesaw method outperforms the standard method with both the conservative and large gain values. Further, these results indicate that the accuracy improves with the larger gain in both the standard and seesaw implementations.

5.3 Rosenbrock Function

The well-known Rosenbrock [29] function has the form, $L(\boldsymbol{\theta}) = 100(t_2 - t_1^2)^2 + (1 - t_1)^2$. It is easily seen that the global minimum over $\Theta = \mathbb{R}^2$ is $\boldsymbol{\theta}^* = (1,1)^T$.

Table 1 Norm values $\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\|$ generated at a sample of iteration counts k while using the standard and seesaw methods. The steepest descent algorithm is implemented for both parameters in trials listed with the gain coefficient $a = 0.15$; the modified steepest descent (including closed-form solution for t_2) is applied for the trials listed below that use the larger gain, $a = 0.29$

k	$a = 0.15$		$a = 0.29$	
	Standard	Seesaw	Standard	Seesaw
0	1.4142	1.4142	1.4142	1.4142
5	0.2215	0.2860	0.2311	0.0048
10	0.0954	0.1152	0.0066	0.00028
25	0.00830	0.0080	9.22×10^{-6}	5.30×10^{-8}
50	0.00014	0.000094	1.59×10^{-10}	3.35×10^{-14}

We follow the pattern of Sect. 5.2 in comparing the steepest descent method with the seesaw method using a standard algorithm form (22) and using a modified steepest descent method that exploits the closed-form solution, $t_2 = t_1^2$. We use the standard initial condition $\hat{\boldsymbol{\theta}}_0 = (-1.2, 1)^T$ (Rosenbrock [29]) in all runs. The topological challenge in optimizing this function is the curved valley that lies between the initial condition and the solution. Note that the function is not quasi-convex (Tseng [7]), and hence not pseudo-convex (Bazaraa et al. [4, p. 569]). Thus, Corollary 3.1 is not applicable; nevertheless, seesaw is numerically convergent, as demonstrated below.

We initially choose a small, constant gain coefficient, $a = 0.0012$, which is the largest value that allows the four implementations—steepest descent and modified steepest descent, each with or without seesaw—to remain stable enough to achieve convergence toward the solution. For this conservative gain value, convergence to $\boldsymbol{\theta}^*$ is relatively slow, although seesaw produces loss values that are less than 10^{-1} and 10^{-8} times the loss values of nonseesaw for steepest descent and modified steepest descent, respectively, at 10,000 iterations. We also conduct a study where the gain a is tuned separately for each of the four implementations, with $a = 0.0012$ or 0.0020 in the nonseesaw implementations of steepest descent and modified steepest descent and $a = 0.0060$ or 0.020 in the corresponding seesaw implementations. The gain value $a = 0.0012$ for standard steepest descent is well under the upper bound to a, $2/\lambda_{\max} = 2/1000.4 = 0.002$, based on a quadratic approximation to L. (The author is unaware of any bound to a available for other implementations of steepest descent.)

Convergence is much faster with the tuned a. Seesaw produces loss values less than 10^{-10} times the loss values of nonseesaw for both the steepest descent and modified steepest descent implementations at 400 iterations. Part of the reason for the relatively greater performance enhancement with seesaw, relative to the common a case, is the fact that it is possible to have a larger ("aggressive") a in seesaw while preserving algorithm stability. The larger a increased the convergence rate.

Table 2 Norm values $\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\|$ generated at a sample of iteration counts k while using the standard steepest descent and seesaw methods for the skewed-quartic loss. The implementation with larger gains (right-hand columns) reflects a tuning process for approximately optimal algorithm performance over $k \leq 1000$

k	$a = 1$		Standard: $a = 2.21$ Seesaw: $a = 5.45$	
	Standard	Seesaw	Standard	Seesaw
0	3.1623	3.1623	3.1623	3.1623
50	0.2835	0.3229	0.7129	0.5362
100	0.1560	0.1756	0.2399	0.0923
500	0.0081	0.0109	0.00023	4.85×10^{-7}
1000	0.00041	0.00056	5.24×10^{-7}	2.54×10^{-13}

5.4 Skewed-Quartic Loss Function

The final test case in this study involves the skewed-quartic loss function from Spall [17, p. 168]:

$$L(\boldsymbol{\theta}) = \boldsymbol{\theta}^T \boldsymbol{B}^T \boldsymbol{B}\boldsymbol{\theta} + 0.1 \sum_{i=1}^{p} (\boldsymbol{B}\boldsymbol{\theta})_i^3 + 0.01 \sum_{i=1}^{p} (\boldsymbol{B}\boldsymbol{\theta})_i^4,$$

where $(\cdot)_i$ represents the ith component of the argument vector $\boldsymbol{B}\boldsymbol{\theta}$, and $\boldsymbol{B}$ is such that $p\boldsymbol{B}$ is an upper triangular matrix of 1's. We consider the unconstrained case with $p = 10$ and $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$ corresponding to the first five and second five components of $\boldsymbol{\theta}$, respectively. The minimum occurs at $\boldsymbol{\theta}^* = \mathbf{0}$ with $L(\boldsymbol{\theta}^*) = 0$; all runs are initialized at $\hat{\boldsymbol{\theta}}_0 = (1, 1, \ldots, 1)^T$ (so $L(\hat{\boldsymbol{\theta}}_0) = 4.178$). We consider only the standard steepest descent method (22), not the modified method used in Sects. 5.2 and 5.3.

We can show that L is strictly convex on Θ by showing that the Hessian matrix $\boldsymbol{H}(\boldsymbol{\theta})$ is positive definite on Θ. To see that $\boldsymbol{H}(\boldsymbol{\theta})$ is positive definite, note that for all $\boldsymbol{\theta}$,

$$\boldsymbol{H}(\boldsymbol{\theta}) = \boldsymbol{B}^T \boldsymbol{B} + 0.6 \sum_{i=1}^{p} (\boldsymbol{B}_i \boldsymbol{\theta}) \boldsymbol{B}_i^T \boldsymbol{B}_i + 0.12 \sum_{i=1}^{p} (\boldsymbol{B}_i \boldsymbol{\theta})^2 \boldsymbol{B}_i^T \boldsymbol{B}_i \geq \frac{1}{4} \boldsymbol{B}^T \boldsymbol{B} > 0,$$

where $\boldsymbol{B}_i$ is the ith row of $\boldsymbol{B}$ and the inequalities are in the usual matrix positive semidefinite or positive definite sense (the matrix lower bound follows from the fact that $\sum_{i=1}^{p} \boldsymbol{B}_i^T \boldsymbol{B}_i = \boldsymbol{B}^T \boldsymbol{B}$). Hence, as discussed in Sect. 5.2 for a different strictly convex function, we know from the analysis in Sect. 3.3 that Theorem 3.2 ensures convergence of $\hat{\boldsymbol{\theta}}_k$ to $\boldsymbol{\theta}^*$ when it is known that $\hat{\boldsymbol{\theta}}_k$ stays in the neighborhood of $\boldsymbol{\theta}^*$ for all k. Without such a local neighborhood requirement, but with Θ reflecting a broader hypercube constraint on $\boldsymbol{\theta}$, we know from Corollary 3.1 (related to pseudoconvexity) that, at a minimum, a change to at least one of $\boldsymbol{\theta}^{(1)}$ or $\boldsymbol{\theta}^{(2)}$ is guaranteed to yield a reduction in L. That is, the main condition of Theorem 3.1 is satisfied.

Table 2 compares the performance of standard steepest descent and seesaw in terms of the error in $\boldsymbol{\theta}$ for a nominal (conservative) gain $a = 1$ and for two gain values, $a = 2.21$ and $a = 5.45$, tuned to provide approximately optimal performance for the standard and seesaw method, respectively. The gain value $a = 2.21$ for standard

steepest descent is slightly under the upper bound to a, $2/\lambda_{\max} = 2/0.8953 = 2.23$, based on a quadratic approximation to L. For the nominal gain, we see that the standard method produces a slightly lower error than seesaw over the $k \leq 1{,}000$ iterations that were considered. On the other hand, for the tuned gains, seesaw produces a significantly lower error than the standard method over the full range of iterations, with an improvement of several orders of magnitude at the higher end of the range of iterations. Similar relative results hold when considering the values of L (as opposed to the error in $\boldsymbol{\theta}$). For the nominal gain case, the seesaw method has a terminal loss that is slightly greater than the loss for the standard method, while in the tuned gain case, the seesaw method has a terminal loss that is over 10 orders of magnitude lower than the loss for the standard method. The numerical results indicate that the "more aggressive" gains that are feasible in the seesaw method provide a much faster rate of convergence, both in terms of the accuracy of $\boldsymbol{\theta}$ and the loss value.

6 Concluding Remarks

This paper has provided a description of a seesaw optimization process—also called cyclic, alternating, or block coordinate process—together with associated convergence theory having conditions that differ from existing convergence results. One advantage of seesaw is the preservation of potentially large investments in software while allowing for an extension to include parameters not covered by the original software. For such a use, the seesaw scheme would require a separate module directed at the new parameters and a master program to control the oscillation between original software and the module devoted to the new parameters. We summarized an application of the seesaw idea to system identification for the parameters in dynamical models represented in state–space form.

Aside from the above advantages relative to software preservation, numerical studies have revealed the desirable property of a *faster* rate of convergence for seesaw optimization, relative to standard joint optimization of all parameters, in the application of steepest descent to three test functions. The faster rate is a consequence of the more "aggressive" (larger) gain coefficient possible in the seesaw algorithm. That is, for the joint optimization, the gain coefficient must be chosen small enough to preserve stability for *all* parameters simultaneously (a gain that is too large will cause the algorithm to go unstable and diverge from the both the initial value and the optimal solution). In contrast, with seesaw, it is possible to pick larger gains for each subvector because not all parameters are being changed simultaneously. The larger gain enhances the rate of convergence via having the parameter iterates move in larger steps toward the optimal value, providing for greater accuracy in the search process (relative to the standard method).

There are several open issues related to seesaw. One is to determine whether the observed faster convergence for seesaw with steepest descent in the numerical studies above is an example of a *general* property of faster convergence for both steepest descent methods and other types of search methods (Newton–Raphson, quasi-Newton, etc.). Associated with the question of generality of faster convergence is the need to develop a formal theory that characterizes the rate of convergence to the optimum relative to nonseesaw methods and whether there are "best" step sizes and partitionings

of $\boldsymbol{\theta}$ (of course, factors other than speed of convergence may govern the choice of partitioning, as in the state–space example discussed above). Another issue is whether the "third parameter vector" idea in Audet et al. [11] can be included in seesaw. That is, in the special case of bilinear loss functions, Audet et al. [11] include a third parameter vector, say $\boldsymbol{\beta}$, that is optimized at every subiteration of seesaw: subiteration 1 optimizes $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\beta}$, subiteration 2 optimizes $\boldsymbol{\theta}^{(2)}$ and $\boldsymbol{\beta}$, and so on. It is not clear at this time whether such an idea could be included in the more general setting (not necessarily bilinear) of this paper.

Finally, while the seesaw idea is described above for deterministic optimization, it would also be of interest to evaluate whether seesaw could lead to improved convergence rates in stochastic approximation algorithms, such as in stochastic gradient methods (a.k.a. Robbins–Monro stochastic approximation [SA]), finite-difference SA, and simultaneous perturbation SA (e.g., Spall [17, Chaps. 5–7]). There is reason to believe the advantage will carry over to the stochastic case because larger gain values are known to also improve convergence rates in stochastic approximation (see Spall [17, pp. 113–114]). The author has yet to formally pursue the extension to SA algorithms. Even without the stochastic extension, however, seesaw provides advantages in implementation and convergence for optimization problems encountered in practice.

Acknowledgements This work was partially supported by US Navy Contract N00024-03-D-6606 and a JHU/APL Sabbatical Professorship. I appreciate comments from Dr. Steve Corley (JHU/APL) on a key aspect of Theorem 3.1 and assistance from former student. John Rumbavage. and current student, Qi Wang, with the numerical study in Sect. 5. Preliminary versions of parts of this paper were presented at the 2006 American Control Conference, the 2011 Conference on Information Sciences and Systems, and the 2011 IEEE Conference on Decision and Control; these conference versions did not include the complete theory of this paper and did not include the numerical study here.

References

1. Shumway, R.H., Olsen, D.E., Levy, L.J.: Estimation and tests of hypotheses for the initial mean and covariance in the Kalman filter model. Commun. Stat., Theory Methods **10**, 1625–1641 (1981)
2. Sun, F.K.: A maximum likelihood algorithm for the mean and covariance of nonidentically distributed observations. IEEE Trans. Autom. Control **27**(1), 245–247 (1982)
3. Achtziger, W.: On simultaneous optimization of truss geometry and topology. Struct. Multidiscip. Optim. **33**, 285–304 (2007)
4. Bazaraa, M.S., Sherali, H.D., Shetty, C.M.: Nonlinear Programming: Theory and Algorithms, 2nd edn. Wiley, New York (1993)
5. Miller, R.E.: Optimization: Foundations and Applications. Wiley, New York (2000)
6. Bezdek, J.C., Hathaway, R.J.: Convergence of alternating optimization. Neural Parallel Sci. Comput. **11**(4), 351–368 (2003)
7. Tseng, P.: Convergence of a block coordinate descent method for nondifferentiable minimization. J. Optim. Theory Appl. **109**(3), 475–494 (2001)
8. Bertsekas, D.: Nonlinear Programming, 2nd edn. Athena Scientific, Belmont (1999)
9. Konno, H.: A cutting plane algorithm for solving bilinear programs. Math. Program. **11**, 14–27 (1976)
10. Alarie, S., Audet, C., Jaumard, B., Savard, G.: Concavity cuts for disjoint bilinear programming. Math. Program., Ser. A **90**(2), 373–398 (2001)
11. Audet, C., Brimberg, J., Hansen, P., Le Digabel, S., Mladenović, N.: Pooling problem: alternate formulations and solution methods. Manag. Sci. **50**(6), 761–776 (2004)
12. Lee, S., Park, F.C.: Cyclic optimization algorithms for simultaneous structure and motion recovery in computer vision. Eng. Optim. **40**(5), 403–419 (2008)

13. Fessler, J.A., Hero, A.O.: Space-alternating generalized expectation–maximization algorithm. IEEE Trans. Signal Process. **42**, 2664–2677 (1994)
14. Haaland, B., Min, W., Qian, P.Z.G., Amemiya, Y.: A statistical approach to thermal management of data centers under steady state and system perturbations. J. Am. Stat. Assoc. **105**(491), 1030–1041 (2010)
15. Fessler, J.A., Ficaro, E.P., Clinthorne, N.H., Lange, K.: Grouped-coordinate ascent algorithms for penalized-likelihood transmission image reconstruction. IEEE Trans. Med. Imaging **16**(2), 166–175 (1997)
16. Polak, E.: Optimization: Algorithms and Consistent Approximations. Springer, New York (1997)
17. Spall, J.C.: Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control. Wiley, Hoboken (2003)
18. Ng, S.K., Krishnan, T., McLachlan, G.J.: The EM algorithm. In: Gentle, J.E., Härdle, W., Mori, Y. (eds.) Handbook of Computational Statistics. Springer, New York (2004). Chap. II.5
19. Levy, L.J.: Generic maximum likelihood identification algorithms for linear state space models. In: Proceedings of the Conference on Information Sciences and Systems (CISS), March 1995, Baltimore, MD, pp. 659–667 (1995)
20. Apostol, T.M.: Mathematical Analysis, 2nd edn. Addison-Wesley, Reading (1974)
21. Fleming, W.: Functions of Several Variables. Springer, New York (1977)
22. Goodrich, R.L., Caines, P.E.: Linear system identification from nonstationary cross-sectional data. IEEE Trans. Autom. Control **24**, 403–411 (1979)
23. Spall, J.C.: Cyclic seesaw optimization with applications to state-space model identification. In: Proceedings of the 45th Annual Conference on Information Sciences and Systems (CISS), 23–25 March 2011, Baltimore, MD (2011)
24. Wills, A., Ninness, B.: On gradient-based search for multivariable system estimates. IEEE Trans. Autom. Control **53**(1), 298–306 (2008)
25. Segal, M., Weinstein, E.: A new method for evaluating the log-likelihood gradient, the Hessian, and the Fisher information matrix. IEEE Trans. Inf. Theory **35**(3), 682–687 (1989)
26. Rao, C.R.: Linear Statistical Inference and its Applications, 2nd edn. Wiley, New York (1973)
27. Moon, T.K., Stirling, W.C.: Mathematical Methods and Algorithms for Signal Processing. Prentice Hall, Upper Saddle River (2000)
28. Spall, J.C., Garner, J.P.: Parameter identification for state-space models with nuisance parameters. IEEE Trans. Aerosp. Electron. Syst. **26**(6), 992–998 (1990)
29. Rosenbrock, H.H.: An automatic method for finding the greatest or least value of a function. Comput. J. **3**(3), 175–184 (1960)