2011 50th IEEE Conference on Decision and Control and
European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

Cyclic Seesaw Optimization and Identification

James C. Spall (james.spall@jhuapl.edu)

The Johns Hopkins University
Applied Physics Laboratory (Laurel, MD) and
Department of Applied Mathematics and Statistics (Baltimore, MD).

***Abstract*—**In the seesaw (or cyclic or alternating) method for optimization and identification, 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 the scheme is the preservation of large investments in software while allowing for an extension of capability to include new parameters for estimation. A specific case involves cross-sectional data represented in state-space form, where there is interest in estimating the mean vector and covariance matrix of the initial state vector as well as parameters associated with the dynamics of the underlying differential equations. 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 as a consequence of the more "aggressive" (larger) gain coefficient (step size) possible in the seesaw algorithm.

***Keywords*—**System identification; parameter estimation; alternating optimization; cyclic optimization; block coordinate optimization; recursive estimation.

I. BACKGROUND

In the seesaw (or cyclic, alternating, or block coordinate) approach to optimization and identification, 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. More generally, this paper provides the theoretical foundation and examples for the cyclic approach to such joint estimation in arbitrary identification and optimization problems.

Spall (2006) discussed a "seesaw" process that partially separates the estimation of the original parameters (e.g., $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as above) from the estimation of the other parameters. This contrasts with traditional methods of directly optimizing for the full set of all relevant parameters. We present here some convergence theory and numerical results that go beyond Spall (2006). We also demonstrate by example that seesaw optimization may provide a *faster* rate of convergence to the solution relative to standard optimization methods on the full vector. The examples suggest a more general property of faster convergence as a consequence of the more "aggressive" (larger) gain coefficient (step size) possible in the seesaw algorithm (this conjecture is not proven here).

Let $\boldsymbol{\theta}$ be a p-dimensional vector representing the unknown parameters to be estimated. According to the seesaw estimation, we represent $\boldsymbol{\theta}$ as composed of two subvectors $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$:

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

Iteration by iteration, the subvector $\boldsymbol{\theta}^{(1)}$ is estimated conditioned on the most recent value of $\boldsymbol{\theta}^{(2)}$ and, likewise, $\boldsymbol{\theta}^{(2)}$ is estimated based on the most recent value of $\boldsymbol{\theta}^{(1)}$. In the typical application of interest for the author, $\boldsymbol{\theta}^{(1)}$ represents all parameters associated with $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ and $\boldsymbol{\theta}^{(2)}$ represents the power spectral density parameters that enter the process noise covariance matrix (see Section III).

Note, however, that the method cannot be blindly applied without considering conditions for convergence, as shown in simple counterexamples (e.g., Achtziger, 2007).

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., 1993, pp. 254–255). Seesaw works with *groups* of parameters. Others have considered convergence for the cyclic scheme. For example, Hathaway and Bezdek (2003) 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 (Bazaraa et al., 1993, pp. 257–258, discusses q-linear convergence). Tseng (2001) considers convergence to a stationary, but not necessarily minimum, point for functions that include a non-differentiable and separable contribution (usually added to a non-separable differentiable contribution). Bertsekas (1999, Sect. 2.7) shows convergence to a stationary point for continuously differentiable functions when it is possible to fully (and uniquely) minimize the loss

Acknowledgments: This work was partially supported by U.S. Navy Contract N00024-03-D-6606 and a JHU/APL Sabbatical Professorship. I appreciate the assistance from former student John Rumbavage with the numerical study in Section IV.

in terms of each of the subvectors. We present global convergence theory different from that above.

Let us mention several applications of the cyclic idea. Lee and Park (2008) 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 cyclic optimization idea in the context of the expectation-maximization (EM) method for finding maximum likelihood parameter estimates. For example, Haaland et al. (2010), use the cyclic idea (with four subvectors) to carry out the "M" step of EM 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.

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:

$$\Theta^* \equiv \arg\min_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}) \equiv \{\boldsymbol{\theta}^* \in \Theta : L(\boldsymbol{\theta}^*) \le L(\boldsymbol{\theta}) \text{ for all } \boldsymbol{\theta} \in \Theta\},$$

where $L = L(\boldsymbol{\theta})$ is the loss function to be minimized (e.g., a negative log-likelihood function), $\Theta \subseteq \mathbb{R}^p$ represents the possible values for $\boldsymbol{\theta}$ (i.e., the constraint set for $\boldsymbol{\theta}$), and Θ^* is assumed to be non-empty. The 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^*.

II. Convergence Analysis

This section presents a theorem and supporting corollary that give sufficient conditions for convergence of $\hat{\boldsymbol{\theta}}_k$ to the optimal $\boldsymbol{\theta}$ as the number of iterations in the estimation process increase.

The estimate at iteration k in the seesaw approach is

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

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 $\hat{\boldsymbol{\theta}}_{k-1}^{(2)}$. It is assumed that the seesaw process satisfies the following relationship:

$$L(\hat{\boldsymbol{\theta}}_{k+1}) \le L(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \hat{\boldsymbol{\theta}}_k^{(2)}) \le L(\hat{\boldsymbol{\theta}}_k) \qquad (2.1)$$

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

$$L(\hat{\boldsymbol{\theta}}_{k+1}) < L(\hat{\boldsymbol{\theta}}_k). \qquad (2.2)$$

Let $L^* = L(\boldsymbol{\theta}^*)$ for $\boldsymbol{\theta}^* \in \Theta^*$. Consistent with the notation and ordering in (2.1), 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, $L(\boldsymbol{\theta}_k^{(1)*}, \hat{\boldsymbol{\theta}}_{k-1}^{(2)}) \le L(\boldsymbol{\theta}^{(1)}, \hat{\boldsymbol{\theta}}_{k-1}^{(2)})$ for all $\boldsymbol{\theta}^{(1)}$ and $L(\hat{\boldsymbol{\theta}}_k^{(1)}, \boldsymbol{\theta}_k^{(2)*}) \le L(\hat{\boldsymbol{\theta}}_k^{(1)}, \boldsymbol{\theta}^{(2)})$ for all $\boldsymbol{\theta}^{(2)}$. So, $\boldsymbol{\theta}_k^{(1)*}$ is a function of $\hat{\boldsymbol{\theta}}_{k-1}^{(2)}$ while $\boldsymbol{\theta}_k^{(2)*}$ is a function of $\hat{\boldsymbol{\theta}}_k^{(1)}$.

Corollary 1 to follow pertains to loss functions that are *pseudoconvex* (e.g., Bazaraa et al., 1993, pp. 113–115). Pseudoconvexity is a significant generalization of convexity to include differentiable functions that do not have the classical "bowl shape." However, as with convexity, pseudoconvex functions have the property that if the 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 if for each $\bar{\boldsymbol{\theta}}$, $\bar{\bar{\boldsymbol{\theta}}} \in \Theta$

$$L(\bar{\bar{\boldsymbol{\theta}}}) < L(\bar{\boldsymbol{\theta}}) \text{ implies } \boldsymbol{g}(\bar{\boldsymbol{\theta}})^T(\bar{\bar{\boldsymbol{\theta}}} - \bar{\boldsymbol{\theta}}) < 0, \qquad (2.3)$$

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 when for each distinct $\bar{\boldsymbol{\theta}}$, $\bar{\bar{\boldsymbol{\theta}}} \in \Theta$, $L(\bar{\bar{\boldsymbol{\theta}}}) \le L(\bar{\boldsymbol{\theta}})$ implies $\boldsymbol{g}(\bar{\boldsymbol{\theta}})^T(\bar{\bar{\boldsymbol{\theta}}} - \bar{\boldsymbol{\theta}}) < 0$; see, e.g., Bazaraa et al., 1993, pp. 112 and 116).

Theorem 1 (Spall, 2006). 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 one of $\boldsymbol{\theta}^{(1)}$ or $\boldsymbol{\theta}^{(2)}$ to yield a reduction in L. Let $0 < \gamma \le 1$. Suppose that the two-stage algorithm with properties (2.1) and (2.2) reduces L with respect to $\boldsymbol{\theta}^{(1)}$ or $\boldsymbol{\theta}^{(2)}$ in the sense that at least one of (2.4a) or (2.4b) holds for each $k = 0, 1, 2, \ldots$:

$$\frac{L(\hat{\boldsymbol{\theta}}_k) - L(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \hat{\boldsymbol{\theta}}_k^{(2)})}{L(\hat{\boldsymbol{\theta}}_k) - L(\boldsymbol{\theta}_{k+1}^{(1)*}, \hat{\boldsymbol{\theta}}_k^{(2)})} \ge \gamma \text{ if } \boldsymbol{\theta}^{(1)} \text{ is changed or} \qquad (2.4a)$$

$$\frac{L(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \hat{\boldsymbol{\theta}}_k^{(2)}) - L(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \hat{\boldsymbol{\theta}}_{k+1}^{(2)})}{L(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \hat{\boldsymbol{\theta}}_k^{(2)}) - L(\boldsymbol{\theta}_{k+1}^{(1)}, \boldsymbol{\theta}_{k+1}^{(2)*})} \ge \gamma \text{ if } \boldsymbol{\theta}^{(2)} \text{ is changed.} \qquad (2.4b)$$

Then,

$$L(\hat{\boldsymbol{\theta}}_k) \to L^* \text{ as } k \to \infty. \qquad (2.5)$$

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

$$\hat{\boldsymbol{\theta}}_k \to \boldsymbol{\theta}^* \text{ as } k \to \infty. \qquad (2.6)$$

Remark on Conditions (2.4a, b). 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 (1999, Sect. 2.7).

The corollary below shows that pseudoconvexity is sufficient to satisfy the key condition in Theorems 1 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 will 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, …, $n(m)$, such that $\boldsymbol{\theta}^{(m)} = [\boldsymbol{\theta}^{(m;1)T},...,\boldsymbol{\theta}^{(m;n(m))T}]^T$ for m = 1 or 2. Two important special cases are when the sub-subvectors are the p coordinates of $\boldsymbol{\theta}$ and when the sub-subvectors are the full subvectors themselves . 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 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 non-zero 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, (2.3), 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 int(Θ) (the interior of Θ) is non-empty and if $\boldsymbol{\theta}' \in$ int(Θ), 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 elements of $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}'$, respectively. For $\boldsymbol{\theta}' \in$ int(Θ), 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 and sufficiently small $\delta > 0$ all $0 \le \lambda \le 1$. Hence, because $t_r^* \ne t_r'$ at this r, the mean-value theorem implies that there exist $\delta^{(\pm)} > 0$ and $0 \le \lambda^{(\pm)} \le 1$ such that

$$L(\boldsymbol{\theta}') - L(\boldsymbol{\theta}' + \delta^{(+)}\boldsymbol{e}_r) = -g_r(\boldsymbol{\theta}' + \lambda^{(+)}\delta^{(+)}\boldsymbol{e}_r)\delta^{(+)} > 0 \text{ if } t_r^* > t_r'$$
$$L(\boldsymbol{\theta}') - L(\boldsymbol{\theta}' - \delta^{(-)}\boldsymbol{e}_r) = g_r(\boldsymbol{\theta}' - \lambda^{(-)}\delta^{(-)}\boldsymbol{e}_r)\delta^{(-)} > 0 \text{ if } t_r^* < t_r'.$$

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 $\boldsymbol{\theta}' \in$ int(Θ), 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 $\boldsymbol{\Delta}^{(m;j)} \equiv [0,0,...,0,\boldsymbol{\theta}^{*(m;j)} - \boldsymbol{\theta}'^{(m;j)},0,...,0]^T$. The mean-value theorem and continuity of $\boldsymbol{g}^{(m;j)}$ imply that there exist a sufficiently small $\delta > 0$ and $0 \le \lambda \le 1$ such that

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

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^*. Q.E.D.

The above ideas apply directly when the two-stage seesaw process is generalized to an M-stage process, $M \ge 3$. In particular, suppose that there are vectors $\boldsymbol{\theta}^{(1)}$, $\boldsymbol{\theta}^{(2)}$,…, $\boldsymbol{\theta}^{(M)}$, each processed sequentially in the manner of the two-stage algorithm. That is, the vectors are processed sequentially such that

$$L(\hat{\boldsymbol{\theta}}_{k+1}) \le \ldots \le L(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \hat{\boldsymbol{\theta}}_{k+1}^{(2)},\ldots,\hat{\boldsymbol{\theta}}_{k}^{(M)}) \le L(\hat{\boldsymbol{\theta}}_{k+1}^{(1)}, \hat{\boldsymbol{\theta}}_{k}^{(2)},\ldots,\hat{\boldsymbol{\theta}}_{k}^{(M)}) \le L(\hat{\boldsymbol{\theta}}_{k})$$

subject to $\hat{\boldsymbol{\theta}}_{k+1} \ne \hat{\boldsymbol{\theta}}_k$ only if $L(\hat{\boldsymbol{\theta}}_{k+1}) < L(\hat{\boldsymbol{\theta}}_k)$. Then, the obvious modifications to the statements of Theorem 1 and Corollary 1 apply.

III. Example in State-Space Model Identification

As mentioned in Section I, a motivating application for the seesaw approach is a problem in the identification of parameters in state-space models. 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 (1979), Shumway et al. (1981), and Levy (1995). 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 (2006), which considers the special case where $\boldsymbol{\theta}$ represents the parameters that enter initial state mean and covariance matrix and the state-noise covariance matrices.

A different state-space identification application involving a natural decoupling into two groups of parameters is described in Spall and Garner (1990) 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.)

IV. Numerical Analysis

4.1 *Overview*

In this section, we present a numerical analysis of the seesaw method for three test functions. Although seesaw is not tied to any specific numerical algorithm, we use the steepest (gradient) descent method in the studies here:

$$\hat{\boldsymbol{\theta}}_{k+1} = \hat{\boldsymbol{\theta}}_k - a_k \boldsymbol{g}(\hat{\boldsymbol{\theta}}_k), \quad k = 0, 1, 2, \ldots, \tag{4.1}$$

where a_k is a non-negative (scalar) gain number satisfying certain conditions. As needed, the generic representation of the elements of $\boldsymbol{\theta}$ is according to $\boldsymbol{\theta} \equiv [t_1, t_2, \ldots, t_p]^T$. We use constant gains $a_k = a$ for all k here.

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 [SA]) (e.g. Spall, 2003, Chaps. 4–5). Hence, the performance improvement observed here might point to possible improvements in a stochastic environment, as well. Second-order-type algorithms (such as quasi-Newton, conjugate gradient, or adaptive SPSA) are also based on (4.1), with a_k being replaced by a matrix typically representing some approximation to the inverse Hessian matrix, built up adaptively as the algorithm proceeds across iterations (Bazaraa et al., 1993, Sect. 8.8; Spall, 2003, Sect. 7.8; or Spall, 2009).

4.2 *Simple Quartic Loss Function*

The first test case in this study is the simple quartic loss function in Spall (2003, 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 (4.1) 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 non-seesaw 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 sub-iteration)

The performance of each method is enhanced by the application of the closed-form solution. Further, we maximize the level of gain a applied to the search algorithm in each method in an attempt to achieve the best possible results. This trial-and-error process involves increasing the constant gain coefficient until it reaches a level where taking it any higher will cause $L(\hat{\boldsymbol{\theta}}_k)$ to diverge. Under this approach, both methods attain their highest accuracy when $a = 0.29$, establishing this value as the maximum gain level in each trial for the simple quartic loss function.

Table 4.1 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 $k \le 50$ 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 (eqn. (4.1)) is used with the gain $a = 0.15$. The modified steepest descent (using closed form) applies 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 ($a \ge 0.30$ leads to divergence).

Table 4.1 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.

Table 4.1. 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 that use the larger gain, $a = 0.29$.

	$a = 0.15$		$a = 0.29$	
k	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
50	0.00014	0.000094	1.59×10^{-10}	3.35×10^{-14}

Using $a = 0.29$, we also implement the closed-form solution for the second component of $\boldsymbol{g}(\boldsymbol{\theta})$ in order to further separate the performances of each method. The loss function values are plotted in Figure 4.1, and accuracy data from the trials using the closed-form solution is included with Table 4.1. These results indicate that maximizing the gain has a significant impact on the seesaw method's ability to improve the performance of the steepest descent algorithm. Therefore,

we examine below test functions of greater complexity in order to further explore the improvement possible.

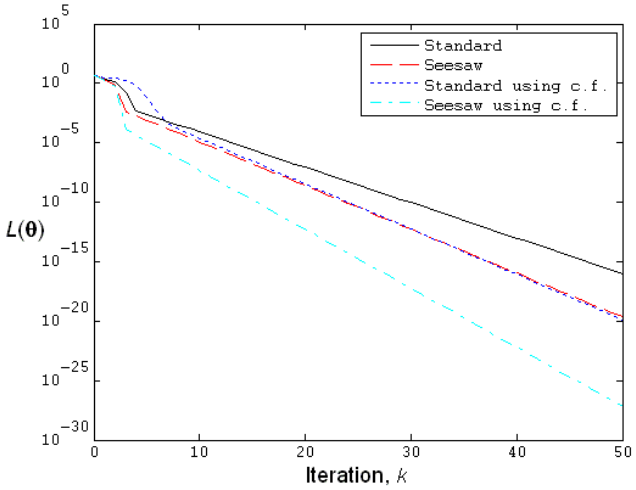

Figure 4.1. Comparison of relative loss function values (on logarithmic scale) generated by the search methods while minimizing the simple quartic loss function. The curves with "using c.f." refer to those runs conducted using the closed-form solution for t_2 in terms of t_1 that exists in this problem. The constant gain coefficient used in each trial is $a = 0.29$.

4.3 *Rosenbrock Function*

The well-known Rosenbrock function has the form, $L(\boldsymbol{\theta}) = 100(t_2 - t_1^2)^2 + (1 - t_1)^2$ (Rosenbrock, 1960). It is easily seen that the global minimum over $\Theta = \mathbb{R}^2$ is $\boldsymbol{\theta}^* = [1, 1]^T$. We follow the pattern of Subsection 4.2 in comparing the steepest descent method with the seesaw method under a common fixed gain coefficient and in sometimes 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, 1960) in all runs. The topological challenge is the curved valley that lies between the initial condition and the solution.

We first chose the 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 towards the solution. For this gain, seesaw produced loss values less than a factor of 10^{-1} and 10^{-8} times that of non-seesaw for steepest descent and modified steepest descent, respectively, at 10,000 iterations. The values of the loss function are displayed in Figure 4.2.

We also conducted a follow-up study where the gain a was tuned separately for each of the four implementations, with $a = 0.0012$ or 0.0020 in the standard (non-seesaw) implementations and $a = 0.0060$ or 0.020 in the seesaw implementations. Seesaw produced loss values less than a factor of 10^{-10} times that of non-seesaw for both the steepest descent and modified steepest descent implementations at 10,000 iterations. Part of the reason for the relatively greater performance enhancement with seesaw, relative to the common a case, was the fact that it was possible to have a larger ("aggressive") a in seesaw while preserving algorithm stability. The larger a increased the convergence rate.

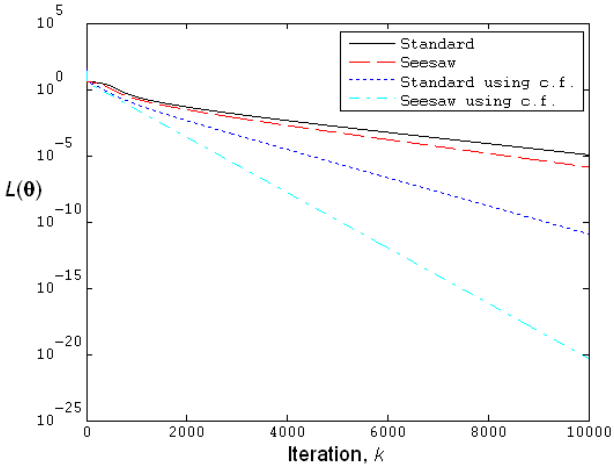

Figure 4.2. Comparison of relative loss function values (on logarithmic scale) generated by the search methods while minimizing the Rosenbrock function. The data representing the trials where we are "using c.f." refer to those which are conducted using the closed-form solution for t_2 in terms of t_1 that exists in this problem. The constant gain coefficient is $a = 0.0012$.

4.4 *Skewed-Quartic Loss Function*

The final test case in this study involves the skewed-quartic loss function from Spall (2003, 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 $\boldsymbol{\theta}^* = \boldsymbol{0}$ 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 (4.1) (not the modified method used in Sections 4.2 and 4.3).

Table 4.2 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. Figure 4.3 shows the relative performance of standard and seesaw for the tuned gains. For the nominal gain, we see that the standard method produces a slightly lower error than seesaw over the $k \le 1000$ 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. The numerical results indicate that the "more

aggressive" gains provide a much faster rate of convergence, both in terms of $\boldsymbol{\theta}$ accuracy and the loss value.

Table 4.2. Norm values $\left\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\right\|$ for 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 $k \le 1000$.

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

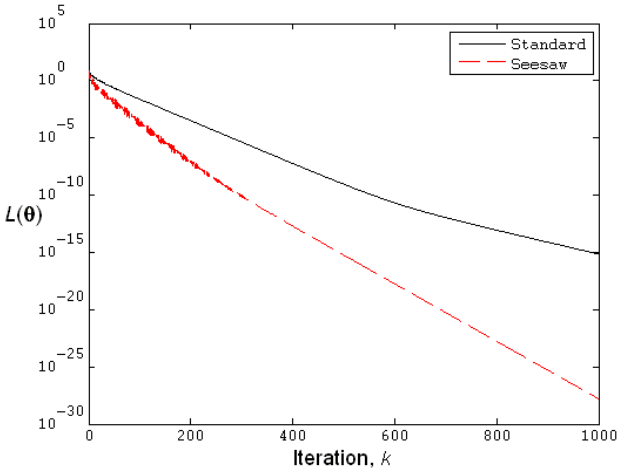

Figure 4.3. Comparison of relative loss function values (on logarithmic scale) generated by each search method while minimizing the skewed-quartic function. Different gain coefficients are used in each trial represented in the figure: a = 2.21 in the trial using the standard method, and a = 5.45 in the trial where seesaw is used.

The results above are consistent with the results for the previous test functions regarding the stabilizing effect that the seesaw method appears to have on search algorithms. These results indicate this technique also increases the efficiency of the overall search process.

V. Concluding Remarks

This paper has provided a description of a seesaw optimization 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 requires a 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.

In addition, numerical studies have revealed the desirable property of a *faster* rate of convergence for seesaw optimization in the three test functions considered as a consequence of the more "aggressive" (larger) gain coefficient possible in the seesaw algorithm. It would also be of interest to evaluate whether the seesaw idea could lead to improved convergence rates in stochastic approximation algorithms such as stochastic gradient methods (a.k.a. Robbins–Monro stochastic approximation [SA]), finite-difference SA, and simultaneous perturbation SA (e.g. Spall, 2003, Chaps. 5-7). Even without the stochastic extension, however, seesaw provides advantages in implementation and convergence for optimization problems encountered in practice.

References

[1] Achtziger, W. (2007), "On Simultaneous Optimization of Truss Geometry and Topology," *Structural and Multidisciplinary Optimization*, vol. 33, pp. 285–304.

[2] Bazaraa, M. S., Sherali, H. D., and Shetty, C. M. (1993), *Nonlinear Programming: Theory and Algorithms* (2nd ed.), Wiley, New York.

[3] Bertsekas, D. (1999), *Nonlinear Programming* (2nd ed.), Athena Scientific, Belmont, MA.

[4] Bezdek, J. C. and Hathaway, R. J. (2003), "Convergence of Alternating Optimization," *Neural, Parallel, and Scientific Computations*, vol. 11(4), pp. 351–368.

[5] Goodrich, R. L. and Caines, P. E. (1979), "Linear System Identification from Nonstationary Cross-Sectional Data," *IEEE Transactions on Automatic Control*, vol. 24, pp. 403–411.

[6] Haaland, B., Min, W., Qian, P. Z. G., and Amemiya, Y. (2010), "A Statistical Approach to Thermal Management of Data Centers Under Steady State and System Perturbations," *Journal of the American Statistical Association*, vol. 105(491), pp. 1030–1041.

[7] Levy, L. J. (1995), "Generic Maximum Likelihood Identification Algorithms for Linear State Space Models," *Proceedings of the Conference on Information Sciences and Systems (CISS)*, March 1995, Baltimore, MD, pp. 659–667.

[8] Rosenbrock, H. H. (1960), "An Automatic Method for Finding the Greatest or Least Value of a Function," *The Computer Journal*, vol. 3, pp. 175–184.

[9] Shumway, R. H., Olsen, D. E., and Levy, L. J. (1981), "Estimation and Tests of Hypotheses for the Initial Mean and Covariance in the Kalman Filter Model," *Communications in Statistics—Theory and Methods*, vol. 10, pp. 1625–1641.

[10] Spall, J. C. (2003), *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*, Wiley, Hoboken, NJ.

[11] Spall, J. C. (2006), "Seesaw Method for Combining Parameter Estimates," *Proceedings of the American Control Conference*, Minneapolis, MN, 14-16 June 2006, pp. 5154–5159 (paper FrB08.1).

[12] Spall, J. C. (2009), "Feedback and Weighting Mechanisms for Improving Jacobian Estimates in the Adaptive Simultaneous Perturbation Algorithm," *IEEE Transactions on Automatic Control*, vol. 54(6), pp. 1216–1229.

[13] Spall, J. C. and Garner, J. P. (1990), "Parameter Identification for State-Space Models with Nuisance Parameters," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 26(6), pp. 992–998.

[14] Tseng, P. (2001), "Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization," *Journal of Optimization Theory and Applications*, vol. 109(3), pp. 475–494.