An Expanded Theoretical Treatment of Iteration-Dependent Majorize-Minimize Algorithms

Matthew W. Jacobson[*] Jeffrey A. Fessler

May 8, 2006

Abstract

The Majorize-Minimize (MM) optimization technique has received considerable attention in signal and image processing applications, as well as in the statistics literature. At each iteration of an MM algorithm, one constructs a *tangent majorant* function that majorizes the given cost function and is equal to it at the current iterate. The next iterate is obtained by minimizing this tangent majorant function, resulting in a sequence of iterates that reduces the cost function monotonically. A well-known special case of MM methods are Expectation-Maximization (EM) algorithms. In this paper, we expand on previous analyses of MM, due to [12, 13], that allowed the tangent majorants to be constructed in iteration-dependent ways. Also, in [13], there was an error in one of the steps of the convergence proof that this paper overcomes.

There are three main aspects in which our analysis builds upon previous work. Firstly, our treatment relaxes many assumptions related to the structure of the cost function, feasible set, and tangent majorants. For example, the cost function can be non-convex and the feasible set for the problem can be any convex set. Secondly, we propose convergence conditions, based on upper curvature bounds, that can be easier to verify than more standard continuity conditions. Furthermore, these conditions in some cases allow for considerable design freedom in the iteration-dependent behavior of the algorithm. Finally, we give an original characterization of the local region of convergence of MM algorithms based on connected (e.g., convex) tangent majorants. For such algorithms, cost function minimizers will locally attract the iterates over larger neighborhoods than is typically guaranteed with other methods. This expanded treatment widens the scope of MM algorithm designs that can be considered for signal and image processing applications, allows us to verify the convergent behavior of previously published algorithms, and gives a fuller understanding overall of how these algorithms behave.

1 Introduction

This paper pertains to the Majorize-Minimize (MM) optimization technique[1] as applied to minimization problems of the form

$$\min.\ \Phi(\boldsymbol{\theta}) \quad \text{s.t.} \quad \boldsymbol{\theta} \in \Theta. \tag{1.1}$$

Here $\Phi(\boldsymbol{\theta}) : \Theta \subset \mathbb{R}^p \to \mathbb{R}$ is a continuously differentiable (but possibly non-convex) cost function, $\mathbb{R}^p$ is the space of length p column vectors,

The MM technique has a long history in a range of literature. In the statistics literature, a prominent example is the Expectation Maximization (EM) methodology (commonly attributed to [9]) which is an application of MM to maximum likelihood estimation. Further examples can be found in in [14, 15, 20, 22]). The interest in maximum likelihood estimation for tomographic image reconstruction subsequently lead to many examples of EM, and more general MM algorithms, in the image processing literature (e.g., [28, 21, 6, 7, 8, 30, 32]). MM has also received considerable attention in the signal processing literature, including [24, 3, 19, 23, 4].

An MM algorithm is one that reduces Φ monotonically by minimizing a succession of approximations to

[*] This work was supported in part by NIH/NCI grant 1P01 CA87634

[1] The technique has gone by various other names as well, such as optimization transfer, SAGE, and iterative majorization. The term MM was coined in [22].

Φ, each of which majorizes Φ in a certain sense. An MM algorithm uses what we call a *majorant generator* $\phi(\cdot;\cdot)$ to associate a given *expansion point* $\boldsymbol{\theta}^i$ with what we call a *tangent majorant* $\phi(\cdot;\boldsymbol{\theta}^i)$. In the simplest case (illustrated for a 1D cost function in Figure 1), a tangent majorant satisfies $\Phi(\boldsymbol{\theta}) \leq \phi(\boldsymbol{\theta};\boldsymbol{\theta}^i)$ for all $\boldsymbol{\theta} \in \Theta$ and $\Phi(\boldsymbol{\theta}^i) = \phi(\boldsymbol{\theta}^i;\boldsymbol{\theta}^i)$. That is, $\phi(\cdot;\boldsymbol{\theta}^i)$ majorizes Φ with equality at $\boldsymbol{\theta}^i$. The constrained minimizer $\boldsymbol{\theta}^{i+1} \in \Theta$ of $\phi(\cdot;\boldsymbol{\theta}^i)$ satisfies $\Phi(\boldsymbol{\theta}^{i+1}) \leq \Phi(\boldsymbol{\theta}^i)$. Repeating these steps iteratively, one obtains a sequence of feasible vectors $\{\boldsymbol{\theta}^i\}$ such that $\{\Phi(\boldsymbol{\theta}^i)\}$ is monotone non-increasing.

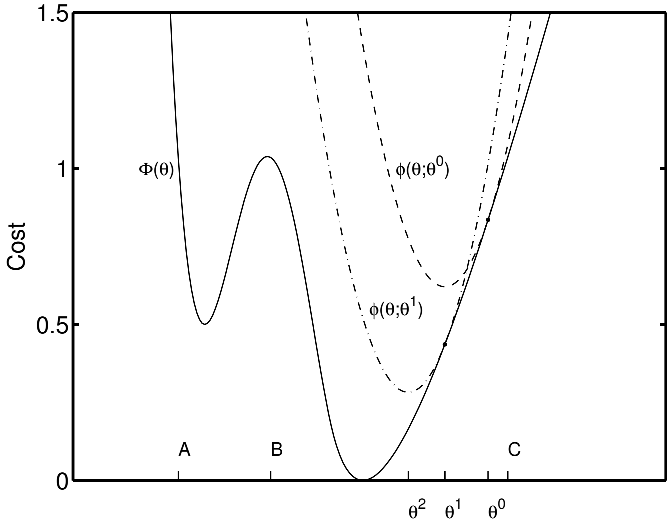

Figure 1: One-dimensional illustration of an MM algorithm.

A more elaborate form of MM was introduced in [12] that allow an iteration-dependent sequence $\{\phi^i(\cdot;\cdot)\}$ of majorant generators to be used, rather than just a single $\phi(\cdot;\cdot)$. This generalization allows considerable freedom in choosing the form of the majorant generator at a given iteration. For example, its form can be adaptively determined based on the observed progress of the algorithm over previous iterations. In addition, one can allow the tangent majorants $\{\phi^i(\cdot;\boldsymbol{\theta}^i)\}$ to depend on an i-dependent subset of the components of $\boldsymbol{\theta}$. The latter results in iterative steps that, similar to coordinate descent, reduce $\Phi(\boldsymbol{\theta})$ as a function of subsets of the optimization variables. This technique, which we call *block alternation*, can simplify algorithm design, because the majorization requirement need be satisfied only with respect to the variables being updated. Furthermore, because the majorization requirement is easier to satisfy, there is empirical evidence that tangent majorants obtained this way may approximate Φ better (leading to faster convergence) than non-block alternating alternatives. An example of where the block alternation technique lead to faster convergence was presented in [12]. Block alternating MM has also seen subsequent use in [24, 11, 3, 19, 23, 4].

The reasons why the MM technique has been attractive to algorithm designers are mixed, and some of the work in this paper may motivate some new reasons. Historically, the main appeal of MM is perhaps that it often leads to algorithms in which the iteration updates are given by simple closed-form formulas (e.g., [28, 6, 7, 10]) and hence, in these cases, tend to be easy to implement. This is in contrast to standard gradient descent methods that employ numerical line searches to ensure global convergence. For large-scale problems, the efficient implementation of line search operations can require complicated customized software implementation, as well as special hardware resources. As an example, one can consider the minimization of the Poisson loglikelihood function encountered in fully 3D Positron Emission Tomography (PET) image reconstruction, e.g., [26]. There, efficiency demands that line searches be implemented in sinogram space. Doing so in turn necessitates considerable RAM, such as would be available on a parallel computing platform. It is likely that, for this reason, investigators in the field of 3D PET have looked to MM alternatives such as [28, 7]. A related reason why MM is attractive is that, when the iteration update computations are simple, one might hope for reduced overall CPU time. This benefit is harder to guarantee, because it demands not only that the tangent majorants be simple to compute/minimize, but also that they provide accurate approximations to Φ, and these two design requirements can conflict. Hence, one sometimes sees examples of MM in the literature that, although easy to implement, are in fact quite slow (e.g., [28]). Conversley, a successful instance of MM acceleration was presented in a logisitic regression example in [22, Example 11]. There, the MM algorithm was found to be competitive with Newton's method. In this paper (see Section 5), we suggest what might be a third benefit of MM. Namely, we discuss how the unusual local convergence properties of MM might be harnessed by certain non-convex minimization strategies.

The overall endeavour of this paper is to revisit and expand the MM convergence analysis of [13]. The scope of [13] is the only one that we know of that includes simultaneously the case where the majorant generator sequence $\{\phi^i(\cdot;\cdot)\}$ can vary non-trivially with i and, furthermore, where minima may lie at constraint boundaries. Our treatment makes three principal contributions to the work begun there. (In the course of our analysis, we also remedy an error in [13], see Remark 4.5).

Our first contribution is to rework the analysis of iteration-dependent MM while relaxing many specific structural assumptions made in [13] on the form of the constraints, the cost function Φ, and the tangent majorants. For example, in [13], only non-negativity constraints were considered.[2] Conversely, in this paper, Θ can be any convex set or, in the case of block alternating MM, any convex set appropriately decomposable into a Cartesian product. Furthermore, in [13], Φ and the tangent majorants were both assumed to be strictly convex. In the present treatment, cases are considered where neither of the two are even convex. Flexibility is also introduced in the domain over which the tangent majorants are defined. In [13], the tangent majorant domains were assumed to be all of Θ, whereas here, the domains can be strict subsets of Θ. Lastly, in [13], the tangent majorants were assumed twice-differentiable, whereas in our analysis, only once-differentiability is assumed. These generalizations widen the range of applications to which [13] is applicable and provide a more flexible framework for algorithm design. Moreover, they allow us to verify the convergence (or at least the asymptotic stationarity) of some previously published block alternating MM algorithms, which the convergence analysis in [13] was not general enough to cover. Among these are the algorithm proposed in [11, Section 6] for the joint estimation of attenuation and activity images in PET. They also include algorithm designs that we proposed (see [17] and [16, Section 6.6]) for a motion-corrected PET image reconstruction application. The convergence analysis in [13] does not apply to these examples because they involve non-convex cost functions, and for various other reasons. Further motivating examples for these generalizations are discussed in [18, Section 6].

Our second contribution is an alternative set of convergence conditions requiring local upper curvature bounds. In the MM literature involving i-independent majorant generators (e.g., [29, 21, 25]), convergence proofs usually invoke an assumption that the $\{\phi^i(\cdot;\cdot)\}$ are continuous (jointly in both arguments). This continuity assumption admits an analysis using Zangwill's convergence theorem [31, p. 91]. In [13], this line of analysis was generalized to iteration-dependent majorant generators under certain additional conditions, and the present paper continues to study these. In addition, however, we show that the continuity condition can be relaxed in favor of a requirement that the tangent majorant curvatures are uniformly locally upper bounded in the region of the expansion points. This latter condition is sometimes more easily verifiable than the standard continuity-based ones. Furthermore, when block alternation is not used, we show that such a curvature bound is sufficient for convergence while admitting considerable freedom in the iteration-dependent behavior of the algorithm (see Remark 4.2).

Our third contribution is an original characterization of the local region of convergence of MM algorithms to local minima. This branch of our analysis is restricted to tangent majorants that are connected (e.g., convex), which is a common practical case. Algorithm designers commonly design tangent majorants that are convex to facilitate minimization. Our results show that the associated MM algorithm will be attracted to a local minima from essentially any point within a basin-like region surrounding that minimum. The same is not generally true of standard derivative-based algorithms. This property has important implications for the tendency of common kinds of MM designs to become trapped at local minima in non-convex minimization problems. As mentioned, however, we also discuss how this property might be harnessed by some established non-convex minimization strategies.

The rest of the paper is organized as follows. In Section 2, we formalize the class of MM algorithms considered in this paper. Next, in Section 3, we give a few additional mathematical preliminaries and describe various conditions imposed in the subsequent analysis. Our analysis begins in Section 4, where we study the global convergence of both block alternating and non-block alternating MM. In this section, the principal step is showing the stationarity of MM limit points under conditions alluded to above. (This asymptotic stationarity property is often used as a definition for "convergence" in the nonlinear optimization literature.) Once asymptotic stationarity is established, convergence of MM in norm can be

[2] Readers who closely scrutinize [13] will see that the line of proof there would also apply to general polyhedral constraints. It would not, however, apply to curved constraints, convex or otherwise.

proved (and we do so in Theorem 4.4) in a standard way by imposing discreteness assumptions[3] on the set of stationary points of (1.1). Section 5 gives our analysis of the local region of convergence for MM, and its relation to capture basins. A concluding summary follows in Section 6.

2 Mathematical Description of MM Algorithms

In this section, we describe the class of MM algorithms considered in this paper. With no loss of generality, we assume that the feasible set Θ is a Cartesian product of $M \leq p$ convex sets, i.e.,

$$\Theta = \Theta_1 \times \Theta_2 \times \ldots \times \Theta_M, \tag{2.1}$$

where $\Theta_m \subset \mathbb{R}^{p_m}$, $m = 1, \ldots, M$ and $\sum_{m=1}^{M} p_m = p$. Since Θ is assumed convex, such a representation is always possible with $M = 1$.

To facilitate discussion, we first introduce some indexing conventions. Given $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \in \Theta$, we can represent $\boldsymbol{\theta}$ as a vertical concatenation[4] of vector partitions $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_M)$ where $\boldsymbol{\theta}_m \in \Theta_m$, $m = 1, \ldots, M$. If $\mathcal{S} = \{m_1, m_2, \ldots, m_q\}$ is a subset of $\{1, \ldots, M\}$, then we write

$$\begin{aligned}
\boldsymbol{\theta}_{\mathcal{S}} &= (\boldsymbol{\theta}_{m_1}, \boldsymbol{\theta}_{m_2}, \ldots, \boldsymbol{\theta}_{m_q}) \\
\Theta_{\mathcal{S}} &= \Theta_{m_1} \times \Theta_{m_2} \times \ldots \times \Theta_{m_q} \\
\mathbb{R}_{\mathcal{S}} &= \mathbb{R}^{p_{m_1} + p_{m_2} + \ldots + p_{m_q}}
\end{aligned}$$

to indicate certain Cartesian sub-products and their elements. Thus, one can write $\boldsymbol{\theta}_{\mathcal{S}} \in \Theta_{\mathcal{S}} \subset \mathbb{R}_{\mathcal{S}}$. The complement of $\mathcal{S}$ shall be denoted $\tilde{\mathcal{S}}$. We may also represent a given $\boldsymbol{\theta} \in \Theta$ in the partitioned form $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\mathcal{S}}, \boldsymbol{\theta}_{\tilde{\mathcal{S}}})$, and $\Phi(\boldsymbol{\theta})$ may be equivalently written $\Phi(\boldsymbol{\theta}_{\mathcal{S}}, \boldsymbol{\theta}_{\tilde{\mathcal{S}}})$.

Given an index set $\mathcal{S} \subset \{1, \ldots, M\}$ and a point-to-set mapping $D(\cdot)$ such that $\bar{\boldsymbol{\theta}}_{\mathcal{S}} \in D(\bar{\boldsymbol{\theta}}) \subset \Theta_{\mathcal{S}}$ for all $\bar{\boldsymbol{\theta}} \in \Theta$, we define a *majorant generator* $\phi(\cdot;\cdot)$ as a function mapping each $\bar{\boldsymbol{\theta}} \in \Theta$ to what we call a *tangent majorant*, a function $\phi(\cdot;\bar{\boldsymbol{\theta}}) : D(\bar{\boldsymbol{\theta}}) \subset \Theta_{\mathcal{S}} \to \mathbb{R}$ satisfying

$$\begin{aligned}
&\Phi(\boldsymbol{\xi}, \bar{\boldsymbol{\theta}}_{\tilde{\mathcal{S}}}) - \Phi(\bar{\boldsymbol{\theta}}) \\
&\qquad \leq \phi(\boldsymbol{\xi};\bar{\boldsymbol{\theta}}) - \phi(\bar{\boldsymbol{\theta}}_{\mathcal{S}};\bar{\boldsymbol{\theta}}), \qquad \forall \boldsymbol{\xi} \in D(\bar{\boldsymbol{\theta}}).
\end{aligned} \tag{2.2}$$

We call $\bar{\boldsymbol{\theta}}$ the *expansion point* of the tangent majorant. Given the point-to-set mapping $D(\cdot)$, we can also write $\phi(\cdot;\cdot) : \mathcal{D} \to \mathbb{R}$, in which

$$\mathcal{D} = \left\{(\boldsymbol{\xi};\bar{\boldsymbol{\theta}}) \,:\, \boldsymbol{\xi} \in D(\bar{\boldsymbol{\theta}}) \subset \Theta_{\mathcal{S}},\, \bar{\boldsymbol{\theta}} \in \Theta\right\}$$

denotes the domain of the majorant generator.

To design an *MM algorithm*, one selects an initial point $\boldsymbol{\theta}^0 \in \Theta$, a sequence of index sets $\left\{\mathcal{S}^i\right\}_{i=0}^{\infty}$, and a sequence of majorant generators $\left\{\phi^i(\cdot;\cdot) : \mathcal{D}^i \to \mathbb{R}\right\}_{i=0}^{\infty}$ with domains

$$\mathcal{D}^i = \left\{(\boldsymbol{\xi};\bar{\boldsymbol{\theta}}) \,:\, \boldsymbol{\xi} \in D^i(\bar{\boldsymbol{\theta}}) \subset \Theta_{\mathcal{S}^i},\, \bar{\boldsymbol{\theta}} \in \Theta\right\}.$$

where the $D^i(\cdot) \subset \Theta_{\mathcal{S}^i}$ are point-to-set mappings, each satisfying $\bar{\boldsymbol{\theta}}_{\mathcal{S}^i} \in D^i(\bar{\boldsymbol{\theta}})$ for all $\bar{\boldsymbol{\theta}} \in \Theta$. The simplest case is when $D^i(\bar{\boldsymbol{\theta}}) = \Theta_{\mathcal{S}^i}$ and $\mathcal{D}^i = \Theta_{\mathcal{S}^i} \times \Theta$ for all i. This was the assumption made in [13]. This assumption does not hold, however, for the MM algorithms in [6, 17]. Once the majorant generators are chosen, the MM algorithm is implemented by generating an iteration sequence $\left\{\boldsymbol{\theta}^i \in \Theta\right\}_{i=0}^{\infty}$ satisfying,

$$\boldsymbol{\theta}^{i+1}_{\mathcal{S}^i} \in \underset{\boldsymbol{\xi} \in D^i(\boldsymbol{\theta}^i)}{\operatorname{argmin}}\, \phi^i(\boldsymbol{\xi};\boldsymbol{\theta}^i) \tag{2.3}$$

$$\boldsymbol{\theta}^{i+1}_{\tilde{\mathcal{S}}^i} = \boldsymbol{\theta}^{i}_{\tilde{\mathcal{S}}^i}. \tag{2.4}$$

Here, we assume that the set of minimizers in (2.3) is non-empty. We shall refer to the total sequence $\left\{\boldsymbol{\theta}^i\right\}_{i=0}^{\infty}$ produced this way as an *MM sequence*. In the simplest case, in which one chooses $\phi^i(\boldsymbol{\theta}_{\mathcal{S}^i};\bar{\boldsymbol{\theta}}) = \Phi(\boldsymbol{\theta}_{\mathcal{S}^i}, \bar{\boldsymbol{\theta}}_{\tilde{\mathcal{S}}^i})$ for all i, (2.3) and (2.4) become a generalization of block coordinate descent (e.g., [1, p. 267]), in which the coordinate blocks are not necessarily disjoint. By virtue of (2.2) and (2.3), $\{\Phi(\boldsymbol{\theta}^i)\}$ is monotonically non-increasing.

A tangent majorant is a mild generalization of what we call a *true tangent majorant*. A function $\phi(\cdot;\bar{\boldsymbol{\theta}})$ satisfying (2.2) is a true tangent majorant if it also satisfies

$$\phi(\boldsymbol{\xi};\bar{\boldsymbol{\theta}}) \geq \Phi(\boldsymbol{\xi}, \bar{\boldsymbol{\theta}}_{\tilde{\mathcal{S}}}) \qquad \forall \boldsymbol{\xi} \in D(\bar{\boldsymbol{\theta}}), \tag{2.5}$$

$$\phi(\bar{\boldsymbol{\theta}}_{\mathcal{S}};\bar{\boldsymbol{\theta}}) = \Phi(\bar{\boldsymbol{\theta}}). \tag{2.6}$$

[3] Non-isolated stationary points are not generally stable (cf. [1, p. 22]) under perturbations of Φ. Therefore, whether or not an algorithm converges in norm to such points seems mainly a question of theoretical interest. It is for such reasons that algorithm users often settle for algorithms with stationary limit points. Nevertheless, we have done some work on convergence to non-isolated stationary points, which the interested reader can find in [18, Section 7].

[4] In this paper, $(a, b, c, \ldots)$ will always denote the vertical concatentation of vectors/scalars a, b, $c, \ldots$.

That is, $\phi(\cdot;\bar{\boldsymbol{\theta}})$ majorizes $\Phi(\cdot,\bar{\boldsymbol{\theta}}_{\tilde{\mathcal{S}}})$ over $D(\bar{\boldsymbol{\theta}})$ and is tangent to it in the sense that equality holds[5] at $\bar{\boldsymbol{\theta}}_{\mathcal{S}}$. These considerations motivate our choice of the term *tangent majorant*.[6] Any tangent majorant can be made into a true tangent majorant by adding to it an appropriate global constant. Doing so does not influence the update formulae (2.3) and (2.4). The distinction between tangent majorants and true tangent majorants is therefore irrelevant in studying MM sequences. The distinction becomes important, however, when deriving tangent majorants by composition of functions (see [18, Note A.1]).

When the sets $\mathcal{S}^i$ vary non-trivially with the iteration number i, we say that the algorithm is *block alternating* (cf. [12, 13]). Conversely, if all $\mathcal{S}^i = \{1,\ldots,M\}$, then $\Theta_{\mathcal{S}^i} = \Theta$ for all i, and we say that the algorithm is not block alternating (or, that the updates are *simultaneous*). In the latter case, (2.2) simplifies to

$$\begin{aligned}\Phi(\boldsymbol{\xi}) &- \Phi(\bar{\boldsymbol{\theta}}) \\ &\leq \phi(\boldsymbol{\xi};\bar{\boldsymbol{\theta}}) - \phi(\bar{\boldsymbol{\theta}};\bar{\boldsymbol{\theta}}), \qquad \forall \boldsymbol{\xi} \in D(\bar{\boldsymbol{\theta}}),\end{aligned} \tag{2.7}$$

while (2.3) and (2.4) reduce to

$$\boldsymbol{\theta}^{i+1} \in \underset{\boldsymbol{\theta}\in D^i(\boldsymbol{\theta}^i)}{\operatorname{argmin}}\, \phi^i(\boldsymbol{\theta};\boldsymbol{\theta}^i), \tag{2.8}$$

The technique of block alternation can be advantageous because it can be simpler to derive and minimize tangent majorants satisfying (2.2), which involve functions of fewer variables, than tangent majorants satisfying (2.7). Block alternation can also provide faster alternatives to certain non-block alternating algorithm designs [12]. To apply block alternation meaningfully, Θ must be decomposable into the Cartesian product form (2.1) with $M > 1$.

3 Mathematical Preliminaries and Assumptions

In this section, we overview mathematical ideas and assumptions that will arise in the analysis to follow.

3.1 General Mathematical Background

A closed d-dimensional ball of radius r and centered at $x \in \mathbb{R}^d$ is denoted

$$B^d(r,x) \overset{\triangle}{=} \left\{x' \in \mathbb{R}^d \,:\, ||x'-x|| \leq r\right\}.$$

where $||\cdot||$ is the standard Euclidean norm. For the minimization problem (1.1), we shall also use the notation

$$\mathcal{B}_{\mathcal{S}}(r,\boldsymbol{\xi}) \overset{\triangle}{=} \Theta_{\mathcal{S}} \cap \left\{\boldsymbol{\xi}' \in \mathbb{R}_{\mathcal{S}} \,:\, ||\boldsymbol{\xi}'-\boldsymbol{\xi}|| \leq r\right\}.$$

to denote certain constrained balls. Given a set $G \subset \mathbb{R}^d$, the notation $\mathrm{cl}(G)$, $\mathrm{ri}(G)$, and $\mathrm{aff}(G)$ shall denote the closure, relative interior, and affine hull of G, respectively. The notation ∂G will denote the relative boundary, $\mathrm{cl}(G) \setminus \mathrm{ri}(G)$.

A function $f : D \subset \mathbb{R}^d \to \mathbb{R}$ is said to be *connected* on a set $D_0 \subset D$ if (see [27, p. 98]), given any $x,\, y \in D_0$, there exists a continuous function $g : [0,\ 1] \to D_0$ such that $g(0) = x$, $g(1) = y$, and

$$f(g(\alpha)) \leq \max\{f(x), f(y)\}$$

for all $\alpha \in (0,\ 1)$. A set $C \subset \mathbb{R}^d$ is said to be *path-connected* if, given any $x, y \in C$ there exists a continuous function $g : [0,\ 1] \to C$ such that $g(0) = x$ and $g(1) = y$. Convex and quasi-convex functions are simple examples of connected functions with $g(\alpha) = \alpha y + (1-\alpha)x$. Also, it has been shown (e.g., Theorem 4.2.4 in [27, p. 99]) that a function is connected if and only if its sublevel sets are path-connected.

Often, we will need to take gradients with respect to a subset of the components of a function's argument. Given a function $f(x;y)$, we shall denote its gradient with respect to its first argument, x, as $\nabla^{10} f(x;y)$. Likewise, $\nabla^{20} f(x;y)$ shall denote the Hessian with respect to x. An expression like $\nabla_m \Phi(\boldsymbol{\theta})$, $m \in \{1,\ldots,M\}$ shall denote the gradient with respect to the sub-vector $\boldsymbol{\theta}_m \in \Theta_m$ of $\boldsymbol{\theta}$. Similarly, $\nabla_{\mathcal{S}} \Phi(\boldsymbol{\theta})$ is the gradient with respect to $\boldsymbol{\theta}_{\mathcal{S}}$.

A key question in the analysis to follow is whether the limit points of an MM algorithm (i.e., the limits of subsequences of $\{\boldsymbol{\theta}^i\}$) are stationary points of (1.1). By a stationary point of (1.1), we mean a feasible point $\boldsymbol{\theta}^*$ that satisfies the first order necessary optimality condition,

$$\langle \nabla\Phi(\boldsymbol{\theta}^*),\, \boldsymbol{\theta}-\boldsymbol{\theta}^*\rangle \geq 0 \qquad \forall \boldsymbol{\theta} \in \Theta. \tag{3.1}$$

[5] It is also tangent to it in the sense that the directional derivatives of $\phi(\cdot;\bar{\boldsymbol{\theta}})$ and $\Phi(\cdot,\bar{\boldsymbol{\theta}}_{\tilde{\mathcal{S}}})$ match at $\bar{\boldsymbol{\theta}}_{\mathcal{S}}$ except in special circumstances (see [18, Note A.2]).

[6] In some literature, the term *surrogate* has been used, however much more general use of this term has been used in other works. We feel that the term *tangent majorant* is much more descriptive of the kind of surrogate functions used in MM specifically.

Here $\langle \cdot,\, \cdot \rangle$ is the usual Euclidean inner product. Henceforth, when an algorithm produces a sequence $\{\boldsymbol{\theta}^i\}$ whose limit points (if any exist) are stationary points of (1.1), we say that the algorithm and the sequence $\{\boldsymbol{\theta}^i\}$ are *asymptotically stationary* .

3.2 Assumptions on MM Algorithms

Throughout the article, we consider cost functions Φ and tangent majorants $\phi(\cdot;\bar{\boldsymbol{\theta}})$ that are continuously differentiable throughout open supersets of Θ and $D(\bar{\boldsymbol{\theta}})$ respectively. For every $\bar{\boldsymbol{\theta}}$, the domain $D(\bar{\boldsymbol{\theta}})$ is assumed convex. In addition, for a given MM algorithm and corresponding sequence $\{\phi^i(\cdot;\boldsymbol{\theta}^i)\}$, we impose conditions that fall into one of two categories. Conditions in the first category, listed next, are what we think of as regularity conditions. In this list, a condition enumerated (Ri.j) denotes a stronger condition than (Ri), i.e., (Ri.j) implies (Ri). Typical MM algorithms will satisfy these conditions to preclude certain degenerate behavior that could otherwise be exhibited.

(R1) *Feasibility of the algorithm.* The sequence $\{\boldsymbol{\theta}^i\}$ lies in a closed subset of Θ. Thus, any limit point of $\{\boldsymbol{\theta}^i\}$ is feasible.

(R1.1) *Feasibility/boundedness of the algorithm.* The sequence $\{\boldsymbol{\theta}^i\}$ is contained in a compact subset of Θ.

(R2) *First order consistency/continuity.* For each i and $\boldsymbol{\xi} \in \Theta_{\mathcal{S}^i}$, the Gâteaux differential,

$$\eta^i(\boldsymbol{\theta};\boldsymbol{\xi}) \overset{\triangle}{=} \left\langle \nabla^{10}\phi^i(\boldsymbol{\theta}_{\mathcal{S}^i};\boldsymbol{\theta}),\, \boldsymbol{\xi}-\boldsymbol{\theta}_{\mathcal{S}^i} \right\rangle \qquad (3.2)$$

is continuous as a function of $\boldsymbol{\theta}$ throughout Θ. Furthermore,

$$\eta^i(\boldsymbol{\theta}^i;\boldsymbol{\xi}) = \left\langle \nabla_{\mathcal{S}^i}\Phi(\boldsymbol{\theta}^i),\, \boldsymbol{\xi}-\boldsymbol{\theta}^i_{\mathcal{S}^i} \right\rangle . \qquad (3.3)$$

Thus, the directional derivatives of the tangent majorants $\{\phi^i(\cdot;\boldsymbol{\theta}^i)\}$ at their expansion points match those of the cost function in feasible directions.

(R2.1) *Matching gradients.* For every i and $\bar{\boldsymbol{\theta}} \in \Theta_{\mathcal{S}^i}$,

$$\nabla^{10}\phi^i(\bar{\boldsymbol{\theta}}_{\mathcal{S}^i};\bar{\boldsymbol{\theta}}) = \nabla_{\mathcal{S}^i}\Phi(\bar{\boldsymbol{\theta}}). \qquad (3.4)$$

Here, the tangent majorant and cost function derivatives match in *all* directions (not just feasible ones) and at *all* expansion points (not just at the $\{\boldsymbol{\theta}^i\}$). Note that, under (R2.1), the continuity of any $\eta^i(\cdot;\boldsymbol{\xi})$ follows from (3.4) and the fact that Φ is continuously differentiable.

(R3) *Minimum size of tangent majorant domains.* There exists an $r > 0$ such that $\mathcal{B}_{\mathcal{S}^i}(r, \boldsymbol{\theta}^i_{\mathcal{S}^i}) \subset D^i(\boldsymbol{\theta}^i)$ for all i. In other words, each tangent majorant is defined on a feasible neighborhood of some minimum size around its expansion point.

Aside from the above regularity conditions, most results will require specific combinations of the following technical conditions. Similar to before, a condition denoted (Ci.j) implies (Ci).

(C1) *Connected tangent majorants.* Each tangent majorant $\phi^i(\cdot;\boldsymbol{\theta}^i)$ is connected on its respective domain $D^i(\boldsymbol{\theta}^i)$.

(C2) *Finite collection of majorant generators.* The elements of the sequence $\{\phi^i(\cdot;\cdot)\}$ are chosen from a finite set of majorant generators.

(C3) *Continuity of majorant generators in both arguments.* For each fixed i, the majorant generator $\phi^i(\cdot;\cdot)$ is continuous throughout its domain $\mathcal{D}^i$. In addition, for any closed subset $\mathcal{Z}$ of Θ, there exists an $r^i_{\mathcal{Z}} > 0$ such that the set $\left\{(\boldsymbol{\xi},\boldsymbol{\theta}) : \boldsymbol{\xi} \in \mathcal{B}_{\mathcal{S}^i}(r^i_{\mathcal{Z}}, \boldsymbol{\theta}^i_{\mathcal{S}^i}), \boldsymbol{\theta} \in \mathcal{Z}\right\}$ lies in a closed subset of $\mathcal{D}^i$.

(C4) *Regular updating of coordinate blocks.* There exists an integer $J > 0$ and, for each $m \in \{1,\ldots,M\}$, an index set $\mathcal{S}^{(m)}$ containing m, a majorant generator $\phi^{(m)}(\cdot;\cdot)$, and a set $\mathcal{I}_m = \left\{i : \mathcal{S}^i = \mathcal{S}^{(m)}, \phi^i = \phi^{(m)}\right\}$ such that

$$\forall n \geq 0, \exists i \in [n, n+J] \text{ s.t. } i \in \mathcal{I}_m.$$

That is, every sub-vector $\boldsymbol{\theta}_m \in \Theta_m$, $m = 1\ldots M$ of $\boldsymbol{\theta}$ is updated regularly by some $\phi^{(m)}$.

(C5) *Diminishing differences.* $\lim_{i\to\infty} ||\boldsymbol{\theta}^{i+1} - \boldsymbol{\theta}^i|| = 0.$

(C5.1) *Uniform strong convexity.* The sequence $\{\boldsymbol{\theta}^i\}$ has at least one feasible limit point. Also, there exists a $\gamma^- > 0$, such that for all i and $\boldsymbol{\xi}, \boldsymbol{\psi} \in D^i(\boldsymbol{\theta}^i)$,

$$\left\langle \nabla^{10}\phi^i(\boldsymbol{\xi};\boldsymbol{\theta}^i) - \nabla^{10}\phi^i(\boldsymbol{\psi};\boldsymbol{\theta}^i),\, \boldsymbol{\xi} - \boldsymbol{\psi} \right\rangle \\ \geq \gamma^- ||\boldsymbol{\xi} - \boldsymbol{\psi}||^2.$$

In other words, the $\{\phi^i(\cdot;\boldsymbol{\theta}^i)\}$ are *strongly convex* with curvatures that are uniformly lower bounded in i. The fact that (C5.1) implies (C5) is proven in Lemma 3.4(c).

(C6) *Uniform upper curvature bound.* In addition to (R3), there exists a $\gamma^+ \geq 0$, such that for all i and $\boldsymbol{\xi} \in \mathcal{B}_{\mathcal{S}^i}(r, \boldsymbol{\theta}^i_{\mathcal{S}^i})$ (here $\mathcal{B}_{\mathcal{S}^i}(r, \boldsymbol{\theta}^i_{\mathcal{S}^i})$ is as in (R3)),

$$\left\langle \nabla^{10}\phi^i(\boldsymbol{\xi};\boldsymbol{\theta}^i) - \nabla^{10}\phi^i(\boldsymbol{\theta}^i_{\mathcal{S}^i};\boldsymbol{\theta}^i),\, \boldsymbol{\xi} - \boldsymbol{\theta}^i_{\mathcal{S}^i} \right\rangle \\ \leq \gamma^+ ||\boldsymbol{\xi} - \boldsymbol{\theta}^i_{\mathcal{S}^i}||^2.$$

In other words, the curvatures of the tangent majorants are uniformly upper bounded along line segments emanating from their expansion points. The line segments must extend to the boundary of a feasible neighborhood of size r around the expansion points.

There are a variety of standard conditions under which Condition (R1) will hold. The simplest case is if Θ is itself closed. Alternatively, (R1) will hold if one can show that the sublevel sets $\mathrm{sublev}_\tau \Phi \overset{\triangle}{=} \{\boldsymbol{\theta} \in \Theta : \Phi(\boldsymbol{\theta}) \leq \tau\}$ of Φ are closed, which is often a straightforward exercise. In the latter case, with $\tau_0 = \Phi(\boldsymbol{\theta}^0)$, the sublevel set $\mathrm{sublev}_{\tau_0} \Phi$ is closed, and because $\{\Phi(\boldsymbol{\theta}^i)\}$ is montonically non-increasing, it follows that the entire sequence $\{\boldsymbol{\theta}^i\}$ is contained in this set. Similarly, if Θ (or just $\mathrm{sublev}_{\tau_0} \Phi$) is compact, then (R1.1) holds. The closure or compactness of sublevel sets often follows if Φ is coercive, i.e., tends to infinity at the boundary of Θ.

The simplest case in which (R3) holds is when $D^i(\boldsymbol{\theta}) = \Theta_{\mathcal{S}^i}$ for all i and $\boldsymbol{\theta} \in \Theta$. A typical situation in which (C4) holds is if the index sets $\{\mathcal{S}^i\}$ and the majorant generators $\{\phi^i(\cdot;\cdot)\}$ are chosen cyclically. Condition (C5) has frequently been encountered in the study of feasible direction methods (e.g., [27, p. 474]). Condition (C5.1) is a sufficient condition for (C5) that is relatively easy to verify. It is essentially a generalization of Condition 5 in [13].

Remark 3.1 In the MM literature, the stronger condition (R2.1) is used customarily to ensure (R2). However, for constrained problems, this can be excessive as discussed in [18, Note A.3].

Remark 3.2 Equation (3.3) is, in fact, implied whenever $\mathrm{aff}(D^i(\bar{\boldsymbol{\theta}})) = \mathrm{aff}(\Theta_{\mathcal{S}^i})$ and $\bar{\boldsymbol{\theta}}_{\mathcal{S}^i} \in \mathrm{ri}(D^i(\bar{\boldsymbol{\theta}}))$. For details, see [18, Note A.2].

3.3 Lemmas

We now give several lemmas that facilitate the analysis in this paper. Most of these lemmas are slight generalizations of existing results. Their proofs are straightforward exercises and are omitted here, but the reader can find full proofs in [18].

Lemma 3.3 (Functions with curvature bounds) *Suppose $f : D \subset \mathbb{R}^d \to \mathbb{R}$ is a continuously differentiable function on a convex set D and fix $y \in D$.*

(a) If $\langle \nabla f(x) - \nabla f(y),\, x - y \rangle \leq \gamma^+ ||x - y||^2$ for some $\gamma^+ > 0$ and $\forall x \in D$, then likewise

$$f(x) - f(y) \leq \langle \nabla f(y),\, x - y \rangle + \frac{1}{2}\gamma^+ ||x - y||^2.$$

(b) If $\langle \nabla f(x) - \nabla f(y),\, x - y \rangle \geq \gamma^- ||x - y||^2$, for some $\gamma^- > 0$ and $\forall x \in D$, then likewise

$$f(x) - f(y) \geq \langle \nabla f(y),\, x - y \rangle + \frac{1}{2}\gamma^- ||x - y||^2.$$

Lemma 3.4 (Implications of limit points) *Suppose that $\{\boldsymbol{\theta}^i\}$ is an MM sequence with a limit point $\boldsymbol{\theta}^* \in \Theta$. Then*

(a) $\{\Phi(\boldsymbol{\theta}^i)\} \searrow \Phi(\boldsymbol{\theta}^*)$.

(b) If $\boldsymbol{\theta}^{**} \in \Theta$ is another limit point of $\{\boldsymbol{\theta}^i\}$, then $\Phi(\boldsymbol{\theta}^{**}) = \Phi(\boldsymbol{\theta}^*)$.

(c) If (C5.1) *also holds then, $\lim_{i\to\infty} ||\boldsymbol{\theta}^i - \boldsymbol{\theta}^{i+1}|| = 0$.*

Lemma 3.5 (Convergence to isolated stationary points) *Suppose $\{\boldsymbol{\theta}^i\}$ is a sequence of points lying in a compact set $\mathcal{K} \subset \Theta$ and whose limit points $S \subset \mathcal{K}$ are stationary points of* (1.1)*. Let $\mathcal{C}$ denote the set of all stationary points of* (1.1) *in $\mathcal{K}$. If either of the following is true,* *(a)* $\mathcal{C}$ *is a singleton, or*

(b) Condition (C5) *holds and* $\mathcal{C}$ *is a discrete set.*

then $\{\boldsymbol{\theta}^i\}$ *in fact converges to a point in* $\mathcal{C}$.

4 Asymptotic Stationarity and Convergence to Isolated Stationary Points

In this section, we establish conditions under which MM algorithms are asymptotically stationary. Convergence in norm is then proved under standard supplementary assumptions that the stationary points are isolated (see Theorem 4.4). Theorem 4.1, our first result, establishes that non-block alternating MM sequences are asymptotically stationary under quite mild assumptions. Two sets of assumptions are considered. One set involves (C3), a continuity condition similar to that used in previous MM literature (e.g., [29, 13, 25]). The continuity condition is motivated by early work due to Zangwill [31, p. 91], which established a broadly applicable theory for monotonic algorithms.

In the second set, the central condition is (C6), which requires a uniform local upper bound on the tangent majorant curvatures. To our knowledge, we are the first to consider such a condition in the context of MM.[7] Condition (C6) can be easier to verify than (C3). For example, the SPS algorithm of [10] is an example of MM based on quadratic tangent majorants. To verify that it satisfies (C3), one must show that the optimal curvature function $c_i(l_i^n)$ (see [10], Equation (28)) is continuous, which is not apparent from the defining expression. Conversely, to verify (C6), it is sufficient to show that $c_i(l_i^n)$ bounded, a fact that follows readily from the fact that the cost function considered in [10] has globally bounded second derivatives.

Theorem 4.1 (Stationarity without block alternation) *Suppose that all* $\mathcal{S}^i = \{1, \dots, M\}$*, that* $\{\boldsymbol{\theta}^i\}$ *is an MM sequence generated by* (2.8)*, and that the regularity conditions* (R1)*,* (R2)*, and* (R3) *hold. Suppose further that either* (C6) *or the pair of conditions* $\{$(C2), (C3)$\}$ *holds. Then any limit point of* $\{\boldsymbol{\theta}^i\}$ *is a stationary point of* (1.1).

Proof. Suppose $\boldsymbol{\theta}^* \in \Theta$ is a limit point of $\{\boldsymbol{\theta}^i\}$ (it must lie in Θ due to (R1)) and, aiming for a contradiction, let us assume that it is not a stationary point. Then there exists a $\boldsymbol{\theta}' \neq \boldsymbol{\theta}^* \in \Theta$ such that

$$\left\langle \nabla\Phi(\boldsymbol{\theta}^*),\, \frac{\boldsymbol{\theta}' - \boldsymbol{\theta}^*}{||\boldsymbol{\theta}' - \boldsymbol{\theta}^*||} \right\rangle < 0. \tag{4.1}$$

Since $\nabla\Phi$ is continuous, then, with (R2) and (R3), it follows that there exists a constant $c < 0$ and a subsequence $\{\boldsymbol{\theta}^{i_k}\}$ satisfying, for all k,

$$||\boldsymbol{\theta}' - \boldsymbol{\theta}^{i_k}|| \geq \min(r, ||\boldsymbol{\theta}' - \boldsymbol{\theta}^*||/2) \triangleq \bar{t}, \tag{4.2}$$

where r is as in (R3), and

$$\left\langle \nabla^{10}\phi^k(\boldsymbol{\theta}^{i_k}; \boldsymbol{\theta}^{i_k}),\, \frac{\boldsymbol{\theta}' - \boldsymbol{\theta}^{i_k}}{||\boldsymbol{\theta}' - \boldsymbol{\theta}^{i_k}||} \right\rangle \leq c. \tag{4.3}$$

Define the unit-length direction vectors

$$\mathbf{s}^k \triangleq \frac{\boldsymbol{\theta}' - \boldsymbol{\theta}^{i_k}}{||\boldsymbol{\theta}' - \boldsymbol{\theta}^{i_k}||}, \quad \mathbf{s}^* \triangleq \frac{\boldsymbol{\theta}' - \boldsymbol{\theta}^*}{||\boldsymbol{\theta}' - \boldsymbol{\theta}^*||}$$

and, for $t \in [0,\ \bar{t}]$, the scalar functions

$$\begin{aligned} h_k(t) &\triangleq \phi^{i_k}(\boldsymbol{\theta}^{i_k} + t\mathbf{s}^k; \boldsymbol{\theta}^{i_k}) \\ &\quad - \left[\phi^{i_k}(\boldsymbol{\theta}^{i_k}; \boldsymbol{\theta}^{i_k}) - \Phi(\boldsymbol{\theta}^{i_k})\right]. \end{aligned} \tag{4.4}$$

Due to (R3) and (4.2), all h_k are well-defined on this common interval. The next several inequalities follow from (2.8), (2.7), and Lemma 3.4(a), respectively,

$$\begin{aligned} h_k(t) &\geq \phi^{i_k}(\boldsymbol{\theta}^{i_k+1}; \boldsymbol{\theta}^{i_k}) - \left[\phi^{i_k}(\boldsymbol{\theta}^{i_k}; \boldsymbol{\theta}^{i_k}) - \Phi(\boldsymbol{\theta}^{i_k})\right] \\ &\geq \Phi(\boldsymbol{\theta}^{i_k+1}) \qquad (4.5) \\ &\geq \Phi(\boldsymbol{\theta}^*). \qquad (4.6) \end{aligned}$$

The remainder of the proof addresses separately the cases where $\{$(C6)$\}$ and $\{$(C2), (C3)$\}$ hold.

First, assume that (C6) holds. This, together with Lemma 3.3(a), implies that for $t \in [0,\ \bar{t}]$,

$$h_k(t) - h_k(0) \leq \dot{h}_k(0)t + \frac{\gamma^+}{2}t^2.$$

However, $h_k(0) = \Phi(\boldsymbol{\theta}^{i_k})$, while $\dot{h}_k(0) \leq c$ due to (4.3). These observations, together with (4.6), leads to

$$\Phi(\boldsymbol{\theta}^*) - \Phi(\boldsymbol{\theta}^{i_k}) \leq ct + \frac{\gamma^+}{2}t^2 \qquad t \in [0,\ \bar{t}\,].$$

[7] Curvature bounds also arise in the convergence theory of trust-region methods, e.g., [5, pp. 121-2].

Passing to the limit in k,

$$ct + \frac{\gamma^+}{2} t^2 \ge 0, \qquad t \in [0,\ \bar{t}\,].$$

Finally, dividing this relation through by t and letting $t \searrow 0$ yields $c \ge 0$, contradicting the assumption that $c < 0$, and completing the proof for this case.

Now, assume $\{$(C2), (C3)$\}$. In light of (C2), we can redefine our subsequence $\{\boldsymbol{\theta}^{i_k}\}$ so that, in addition to (4.2) and (4.3), $\phi^k(\cdot;\cdot)$ equals some fixed function $\hat{\phi}(\cdot;\cdot)$ for all k. That and (4.5) give, for $t \in [0,\ \bar{t}\,]$,

$$\begin{aligned} h_k(t) &= \hat{\phi}(\boldsymbol{\theta}^{i_k} + t\mathbf{s}^k; \boldsymbol{\theta}^{i_k}) - \left[\hat{\phi}(\boldsymbol{\theta}^{i_k}; \boldsymbol{\theta}^{i_k}) - \Phi(\boldsymbol{\theta}^{i_k})\right] \\ &\ge \Phi(\boldsymbol{\theta}^{i_k+1}). \end{aligned} \tag{4.7}$$

From (R1), we know that $\{\boldsymbol{\theta}^{i_k}\}$ lies in a closed subset $\mathcal{Z}$ of Θ. With (C3), there therefore exists a positive $r_{\mathcal{Z}} \le \bar{t}$ such that $h_k(t)$, as given in (4.7), converges as $k \to \infty$ to $h^*(t) \triangleq \hat{\phi}(\boldsymbol{\theta}^* + t\mathbf{s}^*; \boldsymbol{\theta}^*) - \left[\hat{\phi}(\boldsymbol{\theta}^*; \boldsymbol{\theta}^*) - \Phi(\boldsymbol{\theta}^*)\right]$ for all $t \in [0,\ r_{\mathcal{Z}}]$. Letting $k \to \infty$ in (4.7) therefore yields,

$$h^*(t) \ge \Phi(\boldsymbol{\theta}^*) \qquad \forall t \in [0,\ r_{\mathcal{Z}}]. \tag{4.8}$$

The function $h^*(t)$ is differentiable at $t = 0$ due to (R2). Now, $h_k(0) = \Phi(\boldsymbol{\theta}^{i_k})$, so that in the limit, $h^*(0) = \Phi(\boldsymbol{\theta}^*)$. Thus, we have that (4.8) holds with equality at $t = 0$, from which it follows that

$$\dot{h}^*(0) \ge 0. \tag{4.9}$$

However, $\dot{h}_k(0) \le c$ due to (4.3), and the continuity requirement in (R2) implies that $\dot{h}_k(0)$ converges to $\dot{h}^*(0)$ as $k \to \infty$. Thus, we have in the limit that $\dot{h}^*(0) \le c < 0$, contradicting (4.9). □

Remark 4.2 (Curvature and iteration-dependence) Note in Theorems 4.1 that, when the curvature upper bound (C6) holds, there is essentially no restriction on how $\{\phi^i(\cdot;\cdot)\}$ can depend on i.

The next result addresses the block alternating case, but requires additional conditions, namely (C4) and (C5). (Although, Condition (C2) is no longer required.) These conditions, however, are no stronger than those invoked previously in [13]. Condition (C4) is a generalization of [13, Condition 6]. Condition (C5) is an implied condition in [13], as shown in Lemma 3 in that paper.

Theorem 4.3 (Stationarity with block alternation) *Suppose that* $\{\boldsymbol{\theta}^i\}$ *is an MM sequence generated by* (2.3) *and* (2.4) *and that the regularity conditions* (R1), (R2), *and* (R3) *hold. Suppose, further, that* (C4), (C5) *and either* (C6) *or* (C3) *holds. Then any limit point of* $\{\boldsymbol{\theta}^i\}$ *is a stationary point of* (1.1).

Proof. Suppose $\boldsymbol{\theta}^* \in \Theta$ is a limit point of $\{\boldsymbol{\theta}^i\}$ (it must lie in Θ due to (R1)) and, aiming for a contradiction, let us assume that it is not a stationary point. In light of (2.1), there therefore exists a $\boldsymbol{\theta}' \ne \boldsymbol{\theta}^* \in \Theta$ and an $m \in \{1, \dots, M\}$, such that

$$\left\langle \nabla_m \Phi(\boldsymbol{\theta}^*),\, \boldsymbol{\theta}'_m - \boldsymbol{\theta}^*_m \right\rangle < 0 \tag{4.10}$$

and such that $\boldsymbol{\theta}'_{\tilde{m}} = \boldsymbol{\theta}^*_{\tilde{m}}$, $\forall \tilde{m} \ne m$. Then, with $\mathcal{S}^{(m)}$ as in (C4), it follows from (4.10) that,

$$\left\langle \nabla_{\mathcal{S}^{(m)}} \Phi(\boldsymbol{\theta}^*),\, \frac{\boldsymbol{\theta}'_{\mathcal{S}^{(m)}} - \boldsymbol{\theta}^*_{\mathcal{S}^{(m)}}}{||\boldsymbol{\theta}'_{\mathcal{S}^{(m)}} - \boldsymbol{\theta}^*_{\mathcal{S}^{(m)}}||} \right\rangle < 0. \tag{4.11}$$

Now, consider a subsequence $\{\boldsymbol{\theta}^{i_k}\}$ converging to $\boldsymbol{\theta}^*$. We can assume that $\mathcal{S}^{i_k} = \mathcal{S}^{(m)}$ and $\phi^{i_k} = \phi^{(m)}$, for otherwise, in light of (C4), we could construct an alternative subsequence $\{\boldsymbol{\theta}^{i_k + J_k}\}$, $J_k \le J$ which does have this property. Furthermore, this alternative subsequence would converge to $\boldsymbol{\theta}^*$ due to (C5).

In light of (4.11), we can also choose $\{\boldsymbol{\theta}^{i_k}\}$ so that, similar to the proof of Theorem 4.1,

$$||\boldsymbol{\theta}' - \boldsymbol{\theta}^{i_k}|| \ge \min(r, ||\boldsymbol{\theta}' - \boldsymbol{\theta}^*||/2) \triangleq \bar{t}.$$

and

$$\left\langle \nabla^{10} \phi^{(m)}(\boldsymbol{\theta}^{i_k}_{\mathcal{S}^{(m)}}; \boldsymbol{\theta}^{i_k}),\, \frac{\boldsymbol{\theta}'_{\mathcal{S}^{(m)}} - \boldsymbol{\theta}^{i_k}_{\mathcal{S}^{(m)}}}{||\boldsymbol{\theta}'_{\mathcal{S}^{(m)}} - \boldsymbol{\theta}^{i_k}_{\mathcal{S}^{(m)}}||} \right\rangle \le c.$$

for some $c < 0$. Now define

$$\mathbf{s}^k \triangleq \frac{\boldsymbol{\theta}'_{\mathcal{S}^{(m)}} - \boldsymbol{\theta}^{i_k}_{\mathcal{S}^{(m)}}}{||\boldsymbol{\theta}'_{\mathcal{S}^{(m)}} - \boldsymbol{\theta}^{i_k}_{\mathcal{S}^{(m)}}||}$$

and, for $t \in [0,\ \bar{t}\,]$

$$\begin{aligned} h_k(t) \triangleq\, &\phi^{(m)}(\boldsymbol{\theta}^{i_k}_{\mathcal{S}^{(m)}} + t\mathbf{s}^k; \boldsymbol{\theta}^{i_k}) \\ &- \left[\phi^{(m)}(\boldsymbol{\theta}^{i_k}_{\mathcal{S}^{(m)}}; \boldsymbol{\theta}^{i_k}) - \Phi(\boldsymbol{\theta}^{i_k})\right]. \end{aligned}$$

The form and properties of this $h_k(t)$ is a special case of that defined in (4.4). Under (C6), a verbatim argument as in the proof of Theorem 4.1 therefore leads to the contradiction $c \geq 0$, completing the proof for this case. Likewise, the $h_k(t)$ above has the same form and properties as in (4.7). The arguments in the proof of Theorem 4.1 following (4.7) relied only on (C3), and complete the proof of this theorem as well. □

In the following theorem, we deduce convergence in norm by adding discreteness assumptions on the stationary points of (1.1).

Theorem 4.4 (Convergence in norm) *Suppose* $\{\boldsymbol{\theta}^i\}$ *is an MM sequence satisfying* (R1.1)*, as well as the conditions of either Theorem 4.1 or Theorem 4.3. Suppose, in addition, that either of the following is true.*

(a) The problem (1.1) *has a unique solution as its sole stationary point, or*

(b) Condition (C5) *holds and* (1.1) *has a discrete set of stationary points.*

Then $\{\boldsymbol{\theta}^i\}$ *converges to a stationary point. Moreover, in case* (a)*, the limit is the unique solution of* (1.1).

Proof. Under (R1.1), $\{\boldsymbol{\theta}^i\}$ lies in a compact subset of Θ. Moreover, the limit points of $\{\boldsymbol{\theta}^i\}$ are all guaranteed to be stationary by either Theorem 4.1 or Theorem 4.3. The result then follows from Lemma 3.5. □

Remark 4.5 (An error remedied) The convergence analysis in [13] is less general than stated due to an error in the proof of Lemma 6 in that paper. The error occurs where it is argued "if $\nabla_k^{10}\phi^{(k)}(\boldsymbol{\theta}^i_{\mathcal{S}^{(k)}};\boldsymbol{\theta}^i) > 0$ then $\boldsymbol{\theta}_k^{i+1} > \boldsymbol{\theta}_k^i$". This argument would be valid only if, in addition to what was already assumed, $\phi^{(k)}(\cdot;\boldsymbol{\theta}^i)$ were a function of a single variable. Due to the analysis in the present paper, however, we can claim that the *conclusions* of [13] are indeed valid, even if the arguments are not. This follows from Theorem 4.4(a) above, which implies convergence under conditions no stronger than those assumed in [13].

5 Region of Local Convergence for Connected Tangent Majorants

In the study of minimization algorithms, one often wishes to know over what surrounding region of a strict local minimizer an algorithm is guaranteed to converge to that minimizer. In this section, we characterize this region of capture for MM algorithms that use connected (e.g., convex) tangent majorants. It is a prevalent design choice to make the tangent majorants convex, since this facilitates their minimization. We show in Theorem 5.6 that any unimodal, basin-shaped region surrounding a minimizer is a region of capture.

This is to be contrasted with the standard theory concerning derivative-based methods (e.g., gradient, Newton's, Levenberg-Marquardt). If one examines some standard local convergence proofs (e.g., [1, p. 51, Proposition 1.2.5] and [1, p. 90, Proposition 1.4.1(a)]) for these methods, one finds that capture is only guaranteed in a neighborhood where the derivatives are in sufficiently close agreement with the derivatives at the minimizer. Such a neighbourhood can be a significantly small subset of a basin-shaped region around the minimizer. Even just by considering 1D examples (e.g., Figure 1 in the interval $[B,\, C]$), one can see that the first and second derivatives of a cost function can vary greatly throughout a basin. Thus, our findings suggest that connected tangent majorants lead to larger regions of capture than for non-MM derivative-based algorithms. This property has various practical implications that we shall discuss.

To proceed with our analysis, we require a formal mathematical definition of a "basin". The following definition describes what we call a *generalized basin*. It includes the kind of regions that one traditionally thinks of as a basin-shaped region as a special case.

Definition 5.1 We say that a set $G \subset \Theta$ is a *generalized basin* (with respect to the minimization problem (1.1)) if, for some $\boldsymbol{\theta} \in G$, the following is never violated

$$\Phi(\boldsymbol{\theta}) < \Phi(\tilde{\boldsymbol{\theta}}), \qquad \tilde{\boldsymbol{\theta}} \in \mathrm{cl}(G) \cap \mathrm{cl}(\Theta \setminus G). \qquad (5.1)$$

Moreover, we say that such a $\boldsymbol{\theta}$ is *well-contained* in G.

Thus, a point is well-contained in G if it has lower cost than any point $\tilde{\boldsymbol{\theta}}$ in the common boundary $\mathrm{cl}(G) \cap \mathrm{cl}(\Theta \setminus G)$ between G and its complement. The definition is worded so that $\mathrm{cl}(G) \cap \mathrm{cl}(\Theta \setminus G)$ can be empty. Thus, for example, the whole feasible set Θ always constitutes a generalized basin (provided that it contains some $\boldsymbol{\theta}$), because $\mathrm{cl}(\Theta) \cap \mathrm{cl}(\Theta \setminus \Theta)$ is empty, implying that (5.1) can never be violated.

Remark 5.2 The regions described by Definition 5.1 are a bit more general than traditional notions of a capture basin in a few ways. In particular, the definition requires neither that Φ be unimodal over G, nor that G be path-connected. However, it is straightforward to show that any generalized basin G must have the same dimension as Θ, in the sense that $\mathrm{aff}(G) = \mathrm{aff}(\Theta)$ (see [18, Note A.5]). Thus, for example, if $\Theta = \mathbb{R}^2$, no line segment inside Θ can constitute a generalized basin. This is consistent with common intuition.

Remark 5.3 Any sublevel set $G = \{\boldsymbol{\theta} \in \Theta \,:\, \Phi(\boldsymbol{\theta}) \leq \tau\}$ is a generalized basin so long as τ is not the global minimum value of Φ over Θ. Moreover, any global minimizer $\boldsymbol{\theta}^*$ is well-contained in G.

The following proposition lays the foundation for the results of this section. It asserts that, if the expansion point of a connected tangent majorant is well-contained in a generalized basin G, then any point that decreases the cost value of that tangent majorant (relative to the expansion point) is likewise well-contained in G.

Proposition 5.4 *Suppose that $\phi(\cdot;\bar{\boldsymbol{\theta}})$ is a tangent majorant that is connected on its domain $D(\bar{\boldsymbol{\theta}}) \subset \Theta_{\mathcal{S}}$ and whose expansion point $\bar{\boldsymbol{\theta}} \in \Theta$ is well-contained in a generalized basin G. Suppose, further, that $\boldsymbol{\theta} \in \Theta$ satisfies*

$$\begin{aligned} \boldsymbol{\theta}_{\mathcal{S}} \in D(\bar{\boldsymbol{\theta}}), \quad \boldsymbol{\theta}_{\tilde{\mathcal{S}}} &= \bar{\boldsymbol{\theta}}_{\tilde{\mathcal{S}}}, \\ \phi(\boldsymbol{\theta}_{\mathcal{S}};\bar{\boldsymbol{\theta}}) &\leq \phi(\bar{\boldsymbol{\theta}}_{\mathcal{S}};\bar{\boldsymbol{\theta}}), \end{aligned} \tag{5.2}$$

Then $\boldsymbol{\theta}$ is likewise well-contained in G.

Proof. It is sufficient to show that $\boldsymbol{\theta} \in G$. For taking any $\tilde{\boldsymbol{\theta}} \in \mathrm{cl}(G) \cap \mathrm{cl}(\Theta \setminus G)$, and then combining (5.2), (2.2), and the fact that $\bar{\boldsymbol{\theta}}$ is well-contained in G,

$$\Phi(\boldsymbol{\theta}) \leq \Phi(\bar{\boldsymbol{\theta}}) < \Phi(\tilde{\boldsymbol{\theta}}), \tag{5.3}$$

implying that $\boldsymbol{\theta}$ is also well-contained in G. Aiming for a contradiction, suppose that $\boldsymbol{\theta} \in \Theta \setminus G$. Since $\phi(\cdot;\bar{\boldsymbol{\theta}})$ is connected on $D(\bar{\boldsymbol{\theta}})$, there exists a continuous function $\mathbf{g} : [0,\ 1] \rightarrow \Theta$ with $\mathbf{g}(0) = \bar{\boldsymbol{\theta}}$, $\mathbf{g}(1) = \boldsymbol{\theta}$, and such that, for all $\alpha \in (0,\ 1)$, one has

$$\begin{aligned} [\mathbf{g}(\alpha)]_{\mathcal{S}} &\in D(\bar{\boldsymbol{\theta}}), \\ [\mathbf{g}(\alpha)]_{\tilde{\mathcal{S}}} &= \bar{\boldsymbol{\theta}}_{\tilde{\mathcal{S}}}, \\ \phi([\mathbf{g}(\alpha)]_{\mathcal{S}};\bar{\boldsymbol{\theta}}) &\leq \max\{\phi(\bar{\boldsymbol{\theta}}_{\mathcal{S}};\bar{\boldsymbol{\theta}}),\ \phi(\boldsymbol{\theta}_{\mathcal{S}};\bar{\boldsymbol{\theta}})\} \\ &= \phi(\bar{\boldsymbol{\theta}}_{\mathcal{S}};\bar{\boldsymbol{\theta}}), \end{aligned} \tag{5.4}$$

where the equality in (5.4) is due to (5.2). Also, since $\mathbf{g}(0) = \bar{\boldsymbol{\theta}} \in G$,

$$\alpha^* \stackrel{\triangle}{=} \sup\{\alpha \in [0,1] \,:\, \mathbf{g}(\alpha) \in G\}$$

is well-defined. Finally, let $\boldsymbol{\psi} = \mathbf{g}(\alpha^*)$.Combining the definitions of $\mathbf{g}()$ and α^*, the continuity of $\mathbf{g}()$, and the fact that $\boldsymbol{\theta} \in \Theta \setminus G$, one can readily show that $\boldsymbol{\psi} \in \mathrm{cl}(G) \cap \mathrm{cl}(\Theta \setminus G)$.

Therefore, from the rightmost inequality in (5.3), we have, with $\tilde{\boldsymbol{\theta}} = \boldsymbol{\psi}$,

$$\Phi(\bar{\boldsymbol{\theta}}) < \Phi(\boldsymbol{\psi}) = \Phi([\mathbf{g}(\alpha^*)]_{\mathcal{S}}, \bar{\boldsymbol{\theta}}_{\tilde{\mathcal{S}}}). \tag{5.5}$$

With (2.2), this implies that $\phi([\mathbf{g}(\alpha^*)]_{\mathcal{S}};\bar{\boldsymbol{\theta}}) > \phi(\bar{\boldsymbol{\theta}}_{\mathcal{S}};\bar{\boldsymbol{\theta}})$ contradicting (5.4). □

Using Proposition 5.4, we obtain the following result as an immediate consequence. It articulates a capture property for MM sequences.

Theorem 5.5 (Capture property of MM) *Suppose that $\{\boldsymbol{\theta}^i\}$ is an MM sequence generated by* (2.3) *and* (2.4)*. In addition, suppose that some iterate $\boldsymbol{\theta}^n$ is well-contained in a generalized basin G and that the tangent majorant sequence $\{\phi^i(\cdot;\boldsymbol{\theta}^i)\}_{i=n}^{\infty}$ satisfies* (C1)*. Then likewise $\boldsymbol{\theta}^i$ is well-contained in G for all $i > n$.*

Proof. The result follows from Proposition 5.4 and an obvious induction argument. □

Finally, we obtain the principal result of this section.

Theorem 5.6 (Region of Convergence) *In addition to the assumptions of Theorem 5.5, suppose that the conditions of either Theorem 4.1 or Theorem 4.3 are satisfied. Suppose further that G is bounded and* $\mathrm{cl}(G)$ *contains a single stationary point $\boldsymbol{\theta}^*$. Then $\{\boldsymbol{\theta}^i\}$ converges to $\boldsymbol{\theta}^*$.*

Proof. Since G is bounded, it follows from Theorem 5.5 that the sequence $\{\boldsymbol{\theta}^i\}$ lies in the compact set $\mathcal{K} = \mathrm{cl}(G)$. Moreover, all limit points of $\{\boldsymbol{\theta}^i\}$ are stationary, as assured by either Theorem 4.1 or Theorem 4.3. The conclusions of the theorem then follow from Lemma 3.5(a). □

As mentioned, Theorem 5.6 implies that MM algorithms, based on connected tangent majorants, have wider local regions of capture than traditional derivative-based algorithms generally have. There are a mixture of positive and negative practical implications to this property. Since it is common to use convex (and hence connected) tangent majorants, it is essential for algorithm designers to be aware of these implications.

A positive consequence is that global minimizers will, as a special case, attract the iterates over larger distances. Thus, the algorithm may only require a a moderately good initial guess of the solution to perform well. A negative consequence is that sub-optimal local minimizers will also attract the iterates over larger distances. Thus, if not even a moderately good initial guess is available, the chances of failure can be high, depending on the preponderance of sub-optimal local minima in the graph of Φ.

A potential application of Theorem 5.6 is to non-convex optimization strategies that decompose the problem into a sequence of local minimization steps. These include a method due to [2] called Graduated Non-Convexity (GNC), in which a parametric family of approximations to the cost function Φ are locally minimized at successive increments of the parameter. The sequence of local minimizers are meant to trace a parametric curve to the global minimum of Φ. Another example is the strategy of selecting a mesh of initial points and locally minimizing Φ around each point so as to probe for the global minimum. In these strategies, MM with connected tangent majorants seem an appropriate tool for implementing the local minimization steps since, of course, local minimization tasks benefit from a wide region of convergence.

6 Summary

In this paper, we have revised the analysis of [13] in an expanded framework, introduced alternative convergence conditions, and provided original insights into the locally convergent behavior of iteration-dependent MM. In the course of doing so, we also remedied an error in the previous convergence proof (see Remark 4.5). The core results of our global convergence analysis were Theorems 4.1 and 4.3, which proved asymptotic stationarity for non-block alternating and block alternating MM respectively. The core result of our local convergence analysis was Proposition 5.4, which proved the fundamental property of MM algorithms employing connected tangent majorants to become trapped in basin-like regions of the cost function. Our treatment here, we believe, provides enhanced insight into the behavior of MM, as well as a highly broad and flexible framework for MM algorithm design. The results have been useful in verifying the convergence of previously proposed algorithms for different PET imaging applications [11, 17, 16].

An unresolved theoretical question is whether MM will converge in norm when the stationary points of the optimization problem are non-isolated. It is rare to be able to prove this behavior for iterative optimization algorithms in general. However, it has been proven for the EM algorithm of Shepp and Vardi [28], a prominent example of MM in the field of emission tomography. Thus, it is tempting to think that this behavior may be provable in wider generality within the class of MM algorithms. Our preliminary work on this question in [18] may be a starting point for future analysis.

References

[1] D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, Belmont, 2 edition, 1999.

[2] A. Blake and A. Zisserman. *Visual reconstruction*. MIT Press, Cambridge, MA, 1987.

[3] N. Cadalli and O. Arikan. Wideband maximum likelihood direction finding and signal parameter estimation by using tree-structured EM algorithm. *IEEE Trans. Sig. Proc.*, 47(1):201–6, January 1999.

[4] P. J. Chung and J. F. Böhme. Comparative convergence analysis of EM and SAGE algorithms in DOA estimation. *IEEE Trans. Sig. Proc.*, 49(12):2940–9, December 2001.

[5] A. R. Conn, N.I.M. Gould, and P. Toint. *Trust-region Methods*. MPS/SIAM Series on Optimization, Philadelphia, 2000.

[6] A R De Pierro. On the relation between the ISRA and the EM algorithm for positron emission tomography. *IEEE Tr. Med. Im.*, 12(2):328–33, June 1993.

[7] A R De Pierro. A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography. *IEEE Tr. Med. Im.*, 14(1):132–137, March 1995.

[8] A R De Pierro. On the convergence of an EM-type algorithm for penalized likelihood estimation in emission tomography. *IEEE Tr. Med. Im.*, 14(4):762–5, December 1995.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc. Ser. B*, 39(1):1–38, 1977.

[10] H. Erdoğan and J. A. Fessler. Monotonic algorithms for transmission tomography. *IEEE Tr. Med. Im.*, 18(9):801–14, September 1999.

[11] Hakan Erdoğan. *Statistical image reconstruction algorithms using paraboloidal surrogates for PET transmission scans*. PhD thesis, Univ. of Michigan, Ann Arbor, MI, 48109-2122, Ann Arbor, MI., July 1999.

[12] J. A. Fessler and A. O. Hero. Space-alternating generalized expectation-maximization algorithm. *IEEE Tr. Sig. Proc.*, 42(10):2664–77, October 1994.

[13] J. A. Fessler and A. O. Hero. Penalized maximum-likelihood image reconstruction using space-alternating generalized EM algorithms. *IEEE Tr. Im. Proc.*, 4(10):1417–29, October 1995.

[14] W. J. Heiser. Convergent computation by iterative majorization: theory and applications in multidimensional data analysis. In W. J. Krzanowski, editor, *Recent Advances in Descriptive Multivariate Analysis*, Royal Statistical Society Lecture Note Series. Oxford University Press, New York, 1995.

[15] P. J. Huber. *Robust statistics*. Wiley, New York, 1981.

[16] M. Jacobson. *Approaches to motion-corrected PET image reconstruction from respiratory gated projection data*. PhD thesis, Univ. of Michigan, Ann Arbor, MI, 48109-2122, Ann Arbor, MI, 2006.

[17] M. W. Jacobson and J. A. Fessler. Joint estimation of image and deformation parameters in motion-corrected PET. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 5, pages 3290–4, 2003.

[18] M. W. Jacobson and J. A. Fessler. Properties of MM algorithms on convex feasible sets: extended version. Technical Report 353, Comm. and Sign. Proc. Lab., Dept. of EECS, Univ. of Michigan, Ann Arbor, MI, 48109-2122, November 2004.

[19] L. A. Johnston and V. Krishnamurthy. Finite dimensional smoothers for MAP state estimation of bilinear systems. *IEEE Trans. Sig. Proc.*, 47(9):2444–59, September 1999.

[20] K. Lange. A gradient algorithm locally equivalent to the EM Algorithm. *J. Royal Stat. Soc. Ser. B*, 57(2):425–37, 1995.

[21] K. Lange and R. Carson. EM reconstruction algorithms for emission and transmission tomography. *J. Comp. Assisted Tomo.*, 8(2):306–16, April 1984.

[22] K. Lange, D. R. Hunter, and I. Yang. Optimization transfer using surrogate objective functions. *J. Computational and Graphical Stat.*, 9(1):1–20, March 2000.

[23] A. Logothetis and C. Carlemalm. SAGE algorithms for multipath detection and parameter estimation in asynchronous CDMA systems. *IEEE Trans. Sig. Proc.*, 48(11):3162–74, November 2000.

[24] L. B. Nelson and H. V. Poor. Iterative multiuser receivers for CDMA channels: an EM-based approach. *IEEE Trans. Comm.*, 44(12):1700–10, December 1996.

[25] D. Nettleton. Convergence properties of the EM algorithm in constrained parameter spaces. *The Canadian Journal of Statistics*, 27(3):639–48, 1999.

[26] J. M. Ollinger and A. Goggin. Maximum likelihood reconstruction in fully 3D PET via the SAGE algorithm. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 3, pages 1594–8, 1996.

[27] J. M. Ortega and W. C. Rheinboldt. *Iterative solution of nonlinear equations in several variables*. Academic, New York, 1970.

[28] L. A. Shepp and Y. Vardi. Maximum likelihood reconstruction for emission tomography. *IEEE Tr. Med. Im.*, 1(2):113–22, October 1982.

[29] C. F. J. Wu. On the convergence properties of the EM algorithm. *Ann. Stat.*, 11(1):95–103, March 1983.

[30] D. F. Yu, J. A. Fessler, and E. P. Ficaro. Maximum likelihood transmission image reconstruction for overlapping transmission beams. *IEEE Tr. Med. Im.*, 19(11):1094–1105, November 2000.

[31] W. Zangwill. *Nonlinear programming, a unified approach*. Prentice-Hall, NJ, 1969.

[32] J. Zheng, S. Saquib, K. Sauer, and C. Bouman. Parallelizable Bayesian tomography algorithms with rapid, guaranteed convergence. *IEEE Trans. Im. Proc.*, 9(10):1745–59, October 2000.