Published in final edited form as:
IEEE Trans Image Process. 2007 October ; 16(10): 2411–2422.

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

Matthew W. Jacobson* and **Jeffrey A. Fessler [Fellow, IEEE]**

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 [14,15], that allowed the tangent majorants to be constructed in iteration-dependent ways. Also, in [15], 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 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 guaranteed typically 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

$$\min. \Psi(\boldsymbol{x}) \text{ s.t. } \boldsymbol{x} \in X \subset \mathbb{R}^p. \tag{1}$$

Here Ψ is a continuously differentiable, but possibly non-convex cost function, and X is a convex feasible subset of $\mathbb{R}^p$, the space of real, length p column vectors.

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

[1]The term MM was coined in [24]. The technique has gone by various other names as well, such as optimization transfer, SAGE, and iterative majorization.

The MM technique has a long history in a range of fields. In the statistics literature, a prominent example is the Expectation Maximization (EM) methodology [10] which is an application of MM to maximum likelihood estimation. Further examples can be found in in [16,17,22,24]). The interest in maximum likelihood estimation for tomographic image reconstruction subsequently led to many examples of EM, and more general MM algorithms, in image processing (e.g., [34,23,7,8,9,38,36]). MM has also received considerable attention in the signal processing literature, including [28,4,21,26,5].

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

A more elaborate form of MM was introduced in [14] that allows 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 $\{\phi^i_{\boldsymbol{x}^i}(\cdot)\}$ to be functions of i-dependent subsets of the components of $\boldsymbol{x}$. The latter results in iterative steps that, similar to coordinate descent, reduce $\Psi(\boldsymbol{x})$ as a function of subsets of the optimization variables. This technique, called *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 Ψ better (leading to faster convergence) than non-block alternating alternatives. An example of where block alternation led to faster convergence was presented in [14]. Block alternating MM has also seen subsequent use in [28,30,12,4,21,26,5].

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., [34,7,8,11]) 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. For example, consider the minimization of the Poisson loglikelihood function encountered in fully 3D Positron Emission Tomography (PET) image reconstruction, e.g., [30]. 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 [34,8]. A related reason why MM is attractive is that, when the tangent majorants are computationally simple to manipulate, one might hope for reduced overall CPU time. This benefit is harder to guarantee, because it demands not only that the tangent majorants have a simple form, but also that they provide accurate approximations to Ψ, and these two design requirements can conflict. Hence, one sometimes sees examples of MM in the literature that, although easy to implement, converge quite slowly (e.g., [34]). Conversely, a successful instance of MM acceleration was presented in a logisitic regression example in [24, Example 11]. There, the MM algorithm was found to compare

favorably, in terms of convergence rate, with Newton's method. In this paper (see Section 6), 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 [15]. The scope of [15] 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 [15], see Remark 4.5).

Our first contribution is to rework the analysis of iteration-dependent MM while relaxing many specific structural assumptions made in [15] on the form of the constraints, the cost function Ψ, and the tangent majorants. For example, in [15], only non-negativity constraints were considered, whereas in this paper, X can be any convex set or, in the case of block alternating MM, any convex set appropriately decomposable into a Cartesian product. Furthermore, in [15], Ψ and the tangent majorants were both assumed to be strictly convex. Here, we consider cases in which neither of the two are even convex. Flexibility is also introduced in the domain over which the tangent majorants are defined. In [15], the tangent majorant domains were assumed to be all of X, whereas here, the domains can be strict subsets of X. It is not uncommmon for tangent majorants to have smaller domains than Ψ. It is true, for example of the ML-EM algorithm of [34] (see Section 5) and also for the algorithm designs that we proposed (see [19] and [18, Section 6.6]) for a motion-corrected PET image reconstruction application. Lastly, in [15], 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 [15] 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 not encompassed by the convergence analysis in [15]. Among these are the algorithm proposed in [12, Section 6] and those in [19] and [18, Section 6.6]. The convergence analysis in [15] does not apply to these examples at minimum because they involve non-convex Ψ. The line of proof used in [15] depended critically on the strict convexity of the cost function. Also, as discussed above, some of these algorithms use tangent majorants having domains that are not the entire set X. Further motivating examples for these generalizations are discussed in [20, Section 6].

Our second contribution is an alternative set of convergence conditions requiring local upper curvature bounds. In the MM literature involving iteration-independent majorant generators (e.g., [35,23,29]), 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 [37, p. 91]. In [15], 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, this alternative condition admits considerable freedom in the iteration-dependent behavior of the algorithm (see Remark 4.6).

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. Usually, algorithm developers

design tangent majorants that are convex to facilitate their minimization. Our results show that the associated MM algorithm will be attracted to a strict local minima from essentially any point within a basin-like region surrounding that minimum. The same is not generally true of standard gradient 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. 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 nonblock 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 proved (Theorem 4.4) in a standard way by imposing discreteness assumptions[2] on the set of stationary points of (1). In Section 5, we discuss EM algorithms as a special case of MM algorithms and how certain EM algorithms from the tomographic imaging literature relate to our framework. Finally, Section 6 gives our analysis of the local region of convergence for MM, and its relation to capture basins. A concluding summary follows in Section 7.

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 X is a Cartesian product of $M \leq p$ convex sets, i.e.,

$$X = X_1 \times X_2 \times \ldots \times X_M, \tag{2}$$

where $X_m \subset \mathbb{R}^{p_m}$, $m = 1, \ldots, M$ and $\sum_{m=1}^{M} p_m = p$. Since X is assumed convex, such a representation is always at least trivially accomplishable with $M = 1$.

To facilitate discussion, we first introduce some indexing conventions. Given $\boldsymbol{x} = (x_1, \ldots, x_p) \in X$, we can represent $\boldsymbol{x}$ as a vertical concatenation[3] of vector partitions $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M)$ where $\boldsymbol{x}_m \in X_m$, $m = 1, \ldots, M$. We shall refer to any subsequence $S = (m_1, m_2, \ldots, m_q)$ of $(1, \ldots, M)$ as a *block index*, and use the notation

$$\begin{aligned} \boldsymbol{x}_S &= (\boldsymbol{x}_{m_1}, \boldsymbol{x}_{m_2}, \ldots, \boldsymbol{x}_{m_q}) \\ X_S &= X_{m_1} \times X_{m_2} \times \ldots \times X_{m_q} \\ \mathbb{R}_S &= \mathbb{R}^{p_{m_1} + p_{m_2} + \ldots + p_{m_q}} \end{aligned}$$

to indicate certain Cartesian sub-products and their elements. Thus, one has $\boldsymbol{x}_S \in X_S \subset \mathbb{R}_S$. The block index formed from the complement of S in $(1, \ldots, M)$ shall be denoted $\bar{S}$. It will be necessary at times to view $\Psi(\boldsymbol{x})$ as a function of $\boldsymbol{x}_S$ only with the components of $\boldsymbol{x}_{\bar{S}}$ held fixed. We use the notation

[2] Non-discrete stationary points are not generally stable (cf. [1, p. 22]) under perturbations of Ψ, and so are mainly of theoretical interest.

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

$$\Psi_s^{\boldsymbol{y}}(\boldsymbol{z}) \triangleq \Psi(\boldsymbol{x})\Big|_{\substack{\boldsymbol{x}_s=\boldsymbol{z},\\ \boldsymbol{x}_{\bar{s}}=\boldsymbol{y}_{\bar{s}}}}$$

for this purpose, where $\boldsymbol{z} \in X_S$ and $\boldsymbol{y} \in X$.

Given a block index S and a point-to-set mapping $D(\cdot)$ such that $\boldsymbol{y}_S \in D(\boldsymbol{y}) \subset X_S$ for all $\boldsymbol{y} \in X$, we define a *majorant generator* $\boldsymbol{\phi}_{[\cdot]}(\cdot)$ as a function mapping each $\boldsymbol{y} \in X$ to what we call a *tangent majorant*, a function $\boldsymbol{\phi}_{\boldsymbol{y}}(\cdot) : D(\boldsymbol{y}) \subset X_S \to \mathbb{R}$ satisfying

$$\begin{aligned}\Psi_s^{\boldsymbol{y}}(\boldsymbol{z}) &\quad -\Psi(\boldsymbol{y})\\ &\le \phi_{\boldsymbol{y}}(\boldsymbol{z}) - \phi_{\boldsymbol{y}}(\boldsymbol{y}_s), \quad \forall \boldsymbol{z} \in D(\boldsymbol{y}).\end{aligned} \tag{3}$$

We call y the *expansion point* of the tangent majorant. We then also have $\boldsymbol{\phi}_{[\cdot]}(\cdot) : D \to \mathbb{R}$, in which

$$D = \{(\boldsymbol{z}, \boldsymbol{y}) : \boldsymbol{z} \in D(\boldsymbol{y}) \subset X_s, \boldsymbol{y} \in X\}$$

denotes the domain of the majorant generator. The simplest case is when $D = X_S$ and $D = X_S \times X$.

To design an *MM algorithm*, one selects an initial point $\boldsymbol{x}^0 \in X$, a sequence of block indices $\{S^i\}_{i=0}^{\infty}$, and a sequence of majorant generators $\left\{\phi_{[\cdot]}^i(\cdot) : D^i \to \mathbb{R}\right\}_{i=0}^{\infty}$ with domains

$$D^i = \left\{(\boldsymbol{z}, \boldsymbol{y}) : \boldsymbol{z} \in D^i(\boldsymbol{y}) \subset X_{S^i}, \boldsymbol{y} \in X\right\}.$$

where the $D^i(\cdot) \subset X_{S^i}$ are point-to-set mappings of the type described above. Once the majorant generators are chosen, the MM algorithm is implemented by generating an iteration sequence $\{\boldsymbol{x}^i \in X\}_{i=0}^{\infty}$ satisfying,

$$\boldsymbol{x}_{S^i}^{i+1} \in \underset{\boldsymbol{z} \in D^i(\boldsymbol{x}^i)}{\operatorname{argmin}} \phi_{\boldsymbol{x}^i}^i(\boldsymbol{z}) \tag{4}$$

$$\boldsymbol{x}_{\tilde{S}^i}^{i+1} = \boldsymbol{x}_{\tilde{S}^i}^i. \tag{5}$$

Here, we assume that the set of minimizers in (4) is non-empty. We shall refer to the sequence $\{\boldsymbol{x}^i\}_{i=0}^{\infty}$ produced this way as an *MM sequence*. In the simplest case, in which one chooses $\phi_{\boldsymbol{y}}^i(\boldsymbol{x}_{S^i}) = \Psi_{S^i}^{\boldsymbol{y}}(\boldsymbol{x}_{S^i})$ for all i, (4) and (5) become a generalization of block coordinate descent (e.g., [1, p. 267]), in which the coordinate blocks are not necessarily disjoint. By virtue of (3) and (4), $\{\Psi(\boldsymbol{x}^i)\}$ is monotonically non-increasing.

When the block indices S^i are not all equal to (1, …, M), we say that the algorithm is *block alternating* (cf. [14,15]). When the algorithm is not block alternating, i.e., if all $S^i = (1, \ldots, M)$, then (3) simplifies to

$$\Psi(\boldsymbol{z}) - \Psi(\boldsymbol{y}) \leq \phi_{\boldsymbol{y}}(\boldsymbol{z}) - \phi_{\boldsymbol{y}}(\boldsymbol{y}), \ \forall \boldsymbol{z} \in D(\boldsymbol{y}), \tag{6}$$

while (4) and (5) reduce to

$$\boldsymbol{x}^{i+1} \in \underset{\boldsymbol{x} \in D^i(\boldsymbol{x}^i)}{\operatorname{argmin}} \phi^i_{\boldsymbol{x}^i}(\boldsymbol{x}). \tag{7}$$

Observe in (6) that when $\Psi(\boldsymbol{y}) = \phi_{\boldsymbol{y}}(\boldsymbol{y})$ (something we could always ensure by adding a $\boldsymbol{y}$-dependent constant to $\phi_{\boldsymbol{y}}(\cdot)$), one has $\phi_{\boldsymbol{y}}(\boldsymbol{z}) \geq \Psi(\boldsymbol{z})$ with equality at $\boldsymbol{z} = \boldsymbol{y}$. That is, $\phi_{\boldsymbol{y}}(\cdot)$ majorizes Ψ and is tangent to it at $\boldsymbol{y}$. This is the reason for our choice of the terminology *tangent majorant*.

The technique of block alternation can be advantageous because it can be simpler to derive and minimize tangent majorants satisfying (3), which involve functions of fewer variables, than tangent majorants satisfying (6). Block alternation can also provide faster alternatives to certain non-block alternating algorithm designs [14]. To apply block alternation, X must be decomposable into the Cartesian product form (2) 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 $\boldsymbol{u} \in \mathbb{R}^d$ is denoted

$$B^d(r, \boldsymbol{u}) \triangleq \left\{ \boldsymbol{u}' \in \mathbb{R}^d : \|\boldsymbol{u}' - \boldsymbol{u}\| \leq r \right\}.$$

where $\|\cdot\|$ is the standard Euclidean norm. For the minimization problem (1), we shall also use the notation

$$B_S(r, \boldsymbol{z}) \triangleq X_S \cap \left\{ \boldsymbol{z}' \in \mathbb{R}_S : \|\boldsymbol{z}' - \boldsymbol{z}\| \leq r \right\}.$$

to denote certain constrained balls. Given a set $G \subset \mathbb{R}^d$, the notation cl(G), ri(G), and aff(G) shall denote the closure, relative interior, and affine hull of G, respectively. For a more leisurely discussion of these concepts, see [1,33]. The notation ∂G will denote the relative boundary, cl(G) \ 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 [31, p. 98]), given any $\boldsymbol{u}, \boldsymbol{v} \in D_0$, there exists a continuous function $g : [0, 1] \to D_0$ such that $g(0) = \boldsymbol{u}$, $g(1) = \boldsymbol{v}$, and

$$f(g(\alpha)) \leq \max\{f(\boldsymbol{u}), f(\boldsymbol{v})\}$$

for all $\alpha \in (0, 1)$. A set $C \subset \mathbb{R}^d$ is said to be *path-connected* if, given any $\boldsymbol{u}, \boldsymbol{v} \in C$ there exists a continuous function $g : [0, 1] \to C$ such that $g(0) = \boldsymbol{u}$ and $g(1) = \boldsymbol{v}$. Convex and quasi-convex functions are simple examples of connected functions with $g(\alpha) = \alpha\boldsymbol{v} + (1 - \alpha)\boldsymbol{u}$. Also, it has been shown (e.g., Theorem 4.2.4 in [31, p. 99]) that a function is connected if and only if its sublevel sets are path-connected.

A key question in the analysis to follow is whether the limit points of an MM algorithm (i.e., the limits of subsequences of $\{\boldsymbol{x}^i\}$) are stationary points of (1). By a stationary point of (1), we mean a feasible point $\boldsymbol{x}^*$ that satisfies the first order necessary optimality condition,

$$\langle \nabla\Psi(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq 0 \quad \forall \boldsymbol{x} \in X. \tag{8}$$

Here $\langle \cdot, \cdot \rangle$ is the usual Euclidean inner product. If an algorithm produces a sequence $\{\boldsymbol{x}^i\}$ whose limit points (if any exist) are stationary points of (1), we say that the algorithm and the sequence $\{\boldsymbol{x}^i\}$ are *asymptotically stationary*.

3.2 Assumptions on MM Algorithms

Throughout the article, we consider cost functions Ψ and tangent majorants $\phi_{\boldsymbol{y}}(\cdot)$ that are continuously differentiable throughout open supersets of X and $D(\boldsymbol{y})$ respectively. For every $\boldsymbol{y}$, the domain $D(\boldsymbol{y})$ is assumed convex. In addition, for a given MM algorithm and corresponding sequence $\{\phi^i_{\boldsymbol{x}^i}(\cdot)\}$, 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*. We assume that the sequence $\{\boldsymbol{x}^i\}$ lies in a closed subset of X. Thus, any limit point of $\{\boldsymbol{x}^i\}$ is feasible. There are a variety of standard conditions under which (R1) will hold. The simplest case is if X is itself closed. Alternatively, (R1) will hold if one can show that the sublevel sets $\text{sublev}_\tau \Psi \triangleq \{\boldsymbol{x} \in X : \Psi(\boldsymbol{x}) \leq \tau\}$ of Ψ are closed, which is often a straightforward exercise. For example, the closure of sublevel sets often follows if Ψ is coercive, i.e., tends to infinity at the boundary of X and at infinity. In such cases then, with $\tau_0 = \Psi(\boldsymbol{x}^0)$, the sublevel set $\text{sublev}_{\tau_0} \Psi$ is closed, and because $\{\Psi(\boldsymbol{x}^i)\}$ is montonically non-increasing, it follows that the entire sequence $\{\boldsymbol{x}^i\}$ is contained in this set.

- **(R1.1)** *Feasibility/boundedness of the algorithm*. The sequence $\{\boldsymbol{x}^i\}$ is contained in a compact subset of X. Similar to (R1), if X (or just $\text{sublev}_{\tau_0} \Psi$) is compact, then (R1.1) holds. This again is often the case when Ψ is coercive.

(R2) *Agreement and continuity of first derivatives*. The gradient of every tangent majorant agrees with that of Ψ at its expansion point. Formally, for every i and $\boldsymbol{y} \in X$,

$$\nabla \phi_{\boldsymbol{y}}^{i}(\boldsymbol{y}_{S^i}) = \nabla \Psi_{S^i}^{\boldsymbol{y}}(\boldsymbol{y}_{S^i}). \tag{9}$$

Because Ψ is continuously differentiable, it follows from (9) that $\nabla \phi_{\boldsymbol{y}}^{i}(\boldsymbol{y}_{S^i})$ is continuous with respect to $\boldsymbol{y}$.

(R3) *Minimum size of tangent majorant domains*. Each tangent majorant is defined on a feasible neighborhood of some minimum size around its expansion point. Formally, there exists an $r > 0$ such that $B_{S^i}(r, \boldsymbol{x}_{S^i}^{i}) \subset D^i(\boldsymbol{x}^i)$ for all i. The simplest scenario is when $D^i(\boldsymbol{x}) = X_{S^i}$ for all i and $\boldsymbol{x} \in X$, in which case (R3) holds with any $r > 0$.

Remark 3.1 Equation (9) is, in fact, implied by (3) whenever $\mathrm{aff}(D^i(\boldsymbol{y})) = \mathrm{aff}(X_{S^i})$ and $\boldsymbol{y}_{S^i} \in \mathrm{ri}\,(D^i(\boldsymbol{y}))$. For details, see [20, Note A.2].

Remark 3.2 As discussed in [20, Note A.3], Condition (R2) can be weakened when X_{S^i} is of measure zero in $\mathbb{R}_{S^i}$.

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 $\phi_{\boldsymbol{x}^i}^{i}(\cdot)$ is connected on its respective domain $D^i(\boldsymbol{x}^i)$. The typical case is when the tangent majorants are convex.

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

(C3) *Continuity of majorant generators in both arguments*. For each i, the majorant generator $\phi_{[\cdot]}^{i}(\cdot)$ is continuous on its domain D^i. In addition, for any closed subset $\mathcal{Z}$ of X, there exists an $r_{\mathcal{Z}}^{i} > 0$ such that the set $\left\{(\boldsymbol{z}, \boldsymbol{x}) : \boldsymbol{z} \in B_{S^i}(r_{\mathcal{Z}}^{i}, \boldsymbol{x}_{S^i}^{i}), \boldsymbol{x} \in \mathcal{Z}\right\}$ lies in a closed subset of D^i. This is a generalization of the joint continuity condition for EM proposed in [35].

(C4) *Regular updating of coordinate blocks*. Every sub-vector $\boldsymbol{x}_m \in X_m$, $m = 1 \ldots M$ of $\boldsymbol{x}$ is updated by the algorithm at least once in every sequence of J iterations. Formally, for each m, there exists a block index $S^{(m)}$ containing m, so that the set $I_m = \{i : S^i = S^{(m)}\}$ satisfies,

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

A simple case is when the block indices simply cycle over $(1, \ldots, M)$ according to $S^i = i \bmod M + 1$.

(C5) *Diminishing differences*. The sequence $\boldsymbol{x}^i$ satisfies $\lim_{i \to \infty} \|\boldsymbol{x}^{i+1} - \boldsymbol{x}^i\| = 0$. This condition has frequently been encountered in the study of feasible direction methods (e.g., [31, p. 474]). In the MM context, this condition is implied by the following, often readily verifiable condition,

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

$$\left\langle \nabla\phi^i_{\boldsymbol{x}^i}(\boldsymbol{z}) - \nabla\phi^i_{\boldsymbol{x}^i}(w), \boldsymbol{z} - w \right\rangle \geq \gamma^- \|\boldsymbol{z} - w\|^2.$$

In other words, the $\{\phi^i_{\boldsymbol{x}^i}(\cdot)\}$ are *strongly convex* with curvatures that are uniformly lower bounded in i. The fact that (C5.1) implies (C5) is proved in [20, Section 6.3.3]. Essentially, the proof is done by using Lemma 3.3 (b) in this paper with $\boldsymbol{u} = \boldsymbol{x}^{i+1}$, $\boldsymbol{v} = \boldsymbol{x}^i$, and $f = \boldsymbol{\phi}_{\boldsymbol{x}^i}(\cdot)$. Condition (C5.1) generalizes [15, Condition 5].

(C6) *Uniform upper curvature bound*. In addition to (R3), there exists a $\gamma^+ \geq 0$, such that for all i and $\boldsymbol{z} \in B_{s^i}(r, \boldsymbol{x}^i_{s^i})$ (here r is as in (R3)),

$$\left\langle \nabla\phi^i_{\boldsymbol{x}^i}(\boldsymbol{z}) - \nabla\phi^i_{\boldsymbol{x}^i}(\boldsymbol{x}^i_{s^i}), \boldsymbol{z} - \boldsymbol{x}^i_{s^i} \right\rangle \leq \gamma^+ \|\boldsymbol{z} - \boldsymbol{x}^i_{s^i}\|^2.$$

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. When the tangent majorants are twice differentiable, this is equivalent to saying that the second derivatives are locally bounded by γ^+.

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 and are given in [20, Section 6.3.3].

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 $\boldsymbol{v} \in D$.*

a. *If $\langle \nabla f(\boldsymbol{u}) - \nabla f(\boldsymbol{v}), \boldsymbol{u} - \boldsymbol{v} \rangle \leq \gamma^+ \|\boldsymbol{u} - \boldsymbol{v}\|^2$ for some $\gamma^+ > 0$ and $\forall \boldsymbol{u} \in D$, then likewise*

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

b. *If $\langle \nabla f(\boldsymbol{u}) - \nabla f(\boldsymbol{v}), \boldsymbol{u} - \boldsymbol{v} \rangle \geq \gamma^- \|\boldsymbol{u} - \boldsymbol{v}\|^2$ for some $\gamma^- > 0$ and $\forall \boldsymbol{u} \in D$, then likewise*

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

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

a. $\{\Psi(\boldsymbol{x}^i)\} \searrow \Psi(\boldsymbol{x}^*)$.

b. *If $\boldsymbol{x}^{**} \in X$ is another limit point of $\{\boldsymbol{x}^i\}$,* then $\Psi(\boldsymbol{x}^{**}) = \Psi(\boldsymbol{x}^*)$.

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

Lemma 3.5 (Convergence to isolated stationary points) *Suppose* $\{\boldsymbol{x}^i\}$ *is a sequence in a compact set* $K \subset X$ *and whose limit points* $S \subset K$ *are stationary points of* (1). *Let C denote the set of all stationary points of* (1) *in K. If either of the following is true,*

- **a.** C is a singleton, or
- **b.** Condition (C5) holds and C is a discrete set,

then $\{\boldsymbol{x}^i\}$ *in fact converges to a point in 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., [35,15,29]). The continuity condition is motivated by early work due to Zangwill [37, 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 methods.[4] Condition (C6) can be easier to verify than (C3). For example, the algorithm of [11] is an example of MM based on separable quadratic tangent majorants. The optimal choice of curvatures for these quadratics is derived in [11], and is given by a complicated formula. It is much easier to show that these curvatures are uniformly bounded than to show that they are continuous.

Theorem 4.1 (Stationarity without block alternation) *Suppose that all* $S^i = (1,\ldots, M)$, *that* $\{\boldsymbol{x}^i\}$ *is an MM sequence generated by* (7), *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* $\{\boldsymbol{x}^i\}$ *is a stationary point of* (1).

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

$$\left\langle \nabla\Psi(\boldsymbol{x}*), \frac{\boldsymbol{x}' - \boldsymbol{x}*}{\|\boldsymbol{x}' - \boldsymbol{x}*\|} \right\rangle < 0. \tag{10}$$

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

$$\|\boldsymbol{x}' - \boldsymbol{x}^{i_k}\| \geq \min(r, \|\boldsymbol{x}' - \boldsymbol{x}*\|/2) \triangleq \bar{t}, \tag{11}$$

where r is as in (R3), and

[4] Curvature bounds also arise in the convergence theory of trust-region methods, e.g., [6, pp. 121-2].

$$\left\langle \nabla \phi_{\boldsymbol{x}^{i_k}}^{i_k}(\boldsymbol{x}^{i_k}), \frac{\boldsymbol{x}' - \boldsymbol{x}^{i_k}}{\|\boldsymbol{x}' - \boldsymbol{x}^{i_k}\|} \right\rangle \leq c. \tag{12}$$

Define the unit-length direction vectors

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

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

$$\begin{aligned} h_k(t) \triangleq & \phi_{\boldsymbol{x}^{i_k}}^{i_k}(\boldsymbol{x}^{i_k} + t s^k) \\ & - \left[\phi_{\boldsymbol{x}^{i_k}}^{i_k}(\boldsymbol{x}^{i_k}) - \Psi(\boldsymbol{x}^{i_k}) \right]. \end{aligned} \tag{13}$$

Due to (R3) and (11), all h_k are well-defined on this common interval. The next several inequalities follow from (7), (6), and Lemma 3.4(a), respectively,

$$\begin{aligned} h_k(t) &\geq \phi_{\boldsymbol{x}^{i_k}}^{i_k}(\boldsymbol{x}^{i_k+1}) - [\phi_{\boldsymbol{x}^{i_k}}^{i_k}(\boldsymbol{x}^{i_k}) - \Psi(\boldsymbol{x}^{i_k})] \\ &\geq \Psi(\boldsymbol{x}^{i_k+1}) \end{aligned} \tag{14}$$

$$\geq \Psi(\boldsymbol{x}*). \tag{15}$$

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, $h_k(0) = \Psi(\boldsymbol{x}^{i_k})$, while $\dot{h}_k(0) \leq c$ due to (12). These observations, together with (15), leads to

$$\Psi(\boldsymbol{x}*) - \Psi(\boldsymbol{x}^{i_k}) \leq ct + \frac{\gamma^+}{2} t^2 \quad t \in \left[0, \bar{t}\right].$$

Passing to the limit in k,

$$ct + \frac{\gamma^+}{2} t^2 \geq 0, \quad t \in \left[0, \bar{t}\right].$$

Finally, dividing this relation through by t and letting $t \searrow 0$ yields $c \geq 0$, 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 $\{\boldsymbol{x}^{i_k}\}$ so that, in addition to (11) and (12), $\phi_\cdot^{i_k}([\cdot])$ equals some fixed function $\hat{\boldsymbol{\phi}}_\cdot([\cdot])$ for all k. That and (14) give, for $t \in [0, \bar{t}]$,

$$h_k(t) = \widehat{\phi}_{\boldsymbol{x}^{i_k}}(\boldsymbol{x}^{i_k} + t\mathrm{s}^k) - \left[\widehat{\phi}_{\boldsymbol{x}^{i_k}}(\boldsymbol{x}^{i_k}) - \Psi(\boldsymbol{x}^{i_k})\right] \tag{16}$$

$$\geq \Psi(\boldsymbol{x}^{i_k+1}). \tag{17}$$

From (R1), we know that $\{\boldsymbol{x}^{i_k}\}$ lies in a closed subset $\mathcal{Z}$ of X. With (C3), there therefore exists a positive $r_{\mathcal{Z}} \leq \bar{t}$ such that $h_k(t)$ as given in (16), converges as $k \to \infty$ to $h*(t) \triangleq \widehat{\phi}_{\boldsymbol{x}*}(\boldsymbol{x}* + t\mathrm{s}*) - \left[\widehat{\phi}_{\boldsymbol{x}*}(\boldsymbol{x}*) - \Psi(\boldsymbol{x}*)\right]$ for all $t \in [0, r_{\mathcal{Z}}]$. Letting $k \to \infty$ in (17) therefore yields,

$$h*(t) \geq \Psi(\boldsymbol{x}*) \qquad \forall t \in [0, r_{\mathcal{Z}}]. \tag{18}$$

The function $h*(t)$ is differentiable at $t = 0$ due to (R2). Now, $h_k(0) = \Psi(\boldsymbol{x}^{i_k})$, so that in the limit, $h*(0) = \Psi(\boldsymbol{x}*)$. Thus, we have that (18) holds with equality at $t = 0$, from which it follows that

$$\dot{h}^*(0) \geq 0. \tag{19}$$

However, $\dot{h}_k(0) \leq c$ due to (12). Furthermore, since $\nabla\phi_{\boldsymbol{y}}^i(\boldsymbol{y}_S{}^i)$ is continuous in $\boldsymbol{y}$ due to (R2), we have that $\dot{h}_k(0)$ converges to $\dot{h}*(0)$ as k $\to \infty$. With (19), these observations lead to c $\geq$ lim $\inf_k \dot{h}_k(0) \geq 0$, contradicting again the assumption that $c < 0$.

The following example provides a simple illustration of how an MM algorithm can be non-asymptotically stationary when the assumptions of Theorem 4.1 are not met.

Example 4.2 Consider the 1D problem $X = [0, 1.5]$, and $\Psi(x) = 2 - x$. Take $x^0 = 0$ and let $\{x^i\}$ be the sequence generated via (7) with majorant generator

$$\phi_y^i(x) = \phi_y(x) \triangleq c(y)(x - y)^2 + \Psi(x)$$

$$c(y) \triangleq \begin{cases} 1 & y = 1 \\ \frac{1}{|y-1|} & y \neq 1 \end{cases}$$

The resulting sequence of iterates $\{x^i\}$ and tangent majorants $\phi_{x^i}(\cdot)$ are depicted for several iterations in Figure 2. By induction, one can show that $x^i = 1 - 2^{-i}$. Hence, $\{x^i\}$ converges to 1 which is not a stationary point. This presents no conflict with Theorem 4.1, however. The

tangent majorants do not satisfy condition (C6), since the tangent majorant curvatures $\{c(x^i) = 2^i\}$ tend to infinity. Also, $\phi_y(x)$ is discontinuous at $y = 1$, so (C3) is not satisfied. Consequently, the hypothesis of Theorem 4.1 does not hold.

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 [15]. Condition (C4) is a generalization of [15, Condition 6]. Condition (C5) is an implied condition in [15], as shown in Lemma 3 in that paper.

Theorem 4.3 (Stationarity with block alternation) *Suppose that* $\{\boldsymbol{x}^i\}$ *is an MM sequence generated by* (4) *and* (5). *As in Theorem 4.1, assume that* (R1), (R2), (R3), *and either* (C6) *or the pair of conditions* {(C2), (C3)} *hold. In addition, suppose that* (C4) *and* (C5) *hold. Then any limit point of* $\{\boldsymbol{x}^i\}$ *is a stationary point of* (1).

Proof. Suppose $\boldsymbol{x}^* \in X$ is a limit point of $\{x^i\}$ (it must lie in X due to (R1)) and, aiming for a contradiction, let us assume that it is not a stationary point. In light of (2), there therefore exists $\boldsymbol{x}' \neq \boldsymbol{x}^* \in X$ and $m \in \{1, \ldots, M\}$, such that

$$\left\langle \nabla \Psi_m^{\boldsymbol{x}^*}(\boldsymbol{x}_m^*), \boldsymbol{x}'_m - \boldsymbol{x}_m^* \right\rangle < 0 \tag{20}$$

and such that $\boldsymbol{x}'_{\tilde{m}} = \boldsymbol{x}^*_{\tilde{m}}$, $\forall \tilde{m} \neq m$. Then, with $S^{(m)}$ as in (C4), it follows from (20) that,

$$\left\langle \nabla \Psi_{S^{(m)}}^{\boldsymbol{x}^*}(\boldsymbol{x}*), \frac{\boldsymbol{x}'_{S^{(m)}} - \boldsymbol{x}^*_{S^{(m)}}}{\|\boldsymbol{x}'_{S^{(m)}} - \boldsymbol{x}^*_{S^{(m)}}\|} \right\rangle < 0. \tag{21}$$

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

Similar to the proof of Theorem 4.1, we can also choose $\{\boldsymbol{x}^{i_k}\}$ so that,

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

and, in light of (21) and (R2), so that

$$\left\langle \nabla \phi_{\boldsymbol{x}^{i_k}}^{i_k}(\boldsymbol{x}_{S^{(m)}}^{i_k}), \frac{\boldsymbol{x}'_{S^{(m)}} - \boldsymbol{x}_{S^{(m)}}^{i_k}}{\|\boldsymbol{x}'_{S^{(m)}} - \boldsymbol{x}_{S^{(m)}}^{i_k}\|} \right\rangle \leq c.$$

for some $c < 0$. Now define

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

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

$$h_k(t) \triangleq \phi^{i_k}_{\boldsymbol{x}^{i_k}}(\boldsymbol{x}^{i_k}_{S(m)} + ts^k) - [\phi^{i_k}_{\boldsymbol{x}^{i_k}}(\boldsymbol{x}^{i_k}_{S(m)}) - \Psi(\boldsymbol{x}^{i_k})].$$

Manipulations of this $h_k(t)$ verbatim to those in the proof of Theorem 4.1 lead to the contradiction $c \geq 0$, and complete the proof of this theorem, as well.

The following theorem establishes convergence in norm by adding discreteness assumptions on the stationary points of (1).

Theorem 4.4 (Convergence in norm) *Suppose* $\{\boldsymbol{x}^i\}$ *is an MM sequence satisfying* (R1.1) *and 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) *has a unique solution as its sole stationary point, or*
- **b.** *Condition (C5) holds and* (1) *has a discrete set of stationary points.*

Then $\{\boldsymbol{x}^i\}$ *converges to a stationary point. Moreover, in case* (a), *the limit is the unique solution of* (1).

Proof. Under (R1.1), $\{\boldsymbol{x}^i\}$ lies in a compact subset of X. Moreover, the limit points of $\{\boldsymbol{x}^i\}$ 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 analysis in [15] of MM convergence is less general than stated there due to an error in the proof of Lemma 6 in that paper. The error occurs where it is argued "if $\nabla_k \phi^{(k)}(\boldsymbol{x}^i_{S(k)}; \boldsymbol{x}^i) > 0$ then $\boldsymbol{x}^{i+1}_k > \boldsymbol{x}^i_k$". This argument would be valid only if, in addition to what was already assumed, $\boldsymbol{\phi}^{(k)}(\cdot; \boldsymbol{x}^i)$ were a function of a single variable. Due to the analysis in the present paper, however, we can claim that the *conclusions* of [15] 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 [15].

Remark 4.6 (Curvature and iteration-dependence) In Theorems 4.1 and 4.3, when the curvature upper bound (C6) holds, there is very little restriction on how $\{\phi^i_{[\cdot]}(\cdot)\}$ can depend on i, as compared to when {(C2), (C3)} are invoked.

This is useful, for example, if one wishes to use majorant generators that change adaptively based on several previous iterations of the algorithm sequence $\{\boldsymbol{x}^i\}$. For example, one strategy that can be helpful for certain cost functions is to use a block alternating MM algorithm that monitors the gradient sequence $\{\nabla\Psi(\boldsymbol{x}^i)\}$. When certain gradient components are persistently larger than others over several iterations, one switches to a majorant generator that updates only the variables corresponding to those components, thereby conserving computation. Such majorant generator sequences $\{\phi^i_{[\cdot]}(\cdot)\}$ will not generally satisfy Condition (C2), and so one

could not invoke Theorem 4.3 with {(C2), (C3)}. However, $\{\phi^i_{[\cdot]}(\cdot)\}$ could well be made to satisfy (C6).

5 EM as a Special Case of MM

As discussed in Section 1, the family of Expectation Maximization (EM) algorithms is a prominent special case of MM algorithms for minimizing the negative loglikelihood $\Psi(\boldsymbol{x}) = -\log p_{\boldsymbol{x}}(u)$ of a random measurement vector U. In the classical set-up, one develops an EM algorithm by devising a so-called *complete data* random vector V whose joint distribution with U is of the form[5]

$$p_{\boldsymbol{x}}(u, v) = p(u|v)p_{\boldsymbol{x}}(v), \tag{22}$$

i.e., the conditional distribution of U given V is independent of the unknown parameter vector $\boldsymbol{x}$. An EM algorithm is then just an MM algorithm based on the majorant generator,

$$\phi_{\boldsymbol{y}}(\boldsymbol{x}) = \Psi(\boldsymbol{x}) + \mathrm{KL}[\, p_{\boldsymbol{y}}(v|u) \| p_{\boldsymbol{x}}(v|u)], \tag{23}$$

where KL $[f \| g]$ is the Kullback-Leibler (KL) divergence between two probability distributions f and g. For discrete U and V, this is

$$\mathrm{KL}[\, f \| g] \stackrel{\Delta}{=} \sum_z f(\boldsymbol{z}) \left(-\log \frac{g(z)}{f(z)} \right). \tag{24}$$

For continuous random variables, the sum is to be replaced by an integral. A straightforward consequence of Jensen's inequality is that KL $[f \| g] \geq 0$ with equality iff $f = g$. It follows that (23) satisfies the majorization property (6), which in turn ensures the monotonicity of the algorithm.

The term "Expectation Maximization" comes from the fact that

$$\begin{aligned} \phi_{\boldsymbol{y}}(\boldsymbol{x}) &= -\sum_z \log p_{\boldsymbol{x}}(v) p_{\boldsymbol{y}}(v|u) + \text{const.} \\ &= -\mathrm{E}_{\boldsymbol{y}}(\log p_{\boldsymbol{x}}(V) \big| U = u) + \text{const.}, \end{aligned}$$

which one can readily show by combining (23) with (22). Thus, processing the tangent majorant in each iteration is equivalent to taking a conditional expectation and maximizing the result.

A well established convergence condition for EM is the joint continuity assumption proposed in [35], a forerunner to our Condition (C3). A hazard in the design of MM algorithms can arise due to the singularity in the integrand in (24) at $f = g = 0$. Thus, unless $p_{\boldsymbol{x}}(u|v)$ is bounded away from zero as a function of $\boldsymbol{x}$, this singularity may translate into singularities in the tangent majorant(23), so that Condition (C3) is violated. Algorithms such as these therefore do not satisfy standard regularity conditions for EM (or MM) convergence. Such algorithms would

[5]This form was the one considered in [10]. Generalizations have been proposed in [14,25,27].

also violate Condition (C6), since curvatures in the neighbourhood of a singularity are unbounded.

Examples of these pathological cases are to be seen in certain EM algorithms that have been proposed for emission tomographic imaging. A widely considered statistical model for emission tomography projection measurements is

$$U_i \sim \text{Poiss}\{\sum_j a_{ij} x_j + r_i\}.$$

where $x_j \geq 0$ denote unknown image voxel values, $a_{ij} \geq 0$ are system matrix elements, and $r_i \geq 0$ are mean background radiation measurements. An EM algorithm investigated in [32] was based on the complete data choice,

$$V_{ij}^{(1)} = \text{Poiss}\{a_{ij} x_j\}, \quad V_i^{(2)} = \text{Poiss}\{r_i\},$$
$$U_i = \sum_j V_{ij}^{(1)} + V_i^{(2)}$$

for which (23) becomes,

$$\boldsymbol{e}_j(\boldsymbol{y}) \stackrel{\Delta}{=} \left(\sum_i \frac{u_i}{\sum_j a_{ij} y_j + r_i} \right) \tag{25}$$

$$\phi_{\boldsymbol{y}}(\boldsymbol{x}) = \sum_i a_{ij} x_j - \boldsymbol{e}_j(\boldsymbol{y}) y_j \log x_j + \text{const.} \tag{26}$$

In the particular case $r_i = 0$, this algorithm reduces to the ML-EM algorithm of algorithm of [34]. Clearly this $\boldsymbol{\phi}_{[\cdot]}(\cdot)$ can approach singularities as $(x_j, y_j) \to (0, 0)$, and hence (C3) and (C6) are violated. Although the algorithm has been proven to converge for $r_i = 0$, existing analyses (e.g., [3]) are difficult and very specific to the structures of Ψ and $\boldsymbol{\phi}$. The algorithm has not, to our knowledge, been shown to converge for the more general case $r_i \geq 0$, although some relevant analysis was done in [13].

Another consequence of the singularities in KL is that, when all $y_j > 0$, the domain $D(\boldsymbol{y})$ of the tangent majorant(26) is the interior of the non-negative orthant, which is normally a strict subset of the feasible set X. Normally, there are feasible images $\boldsymbol{x}$ in which some of the voxel values are zero (e.g., when for all i, the mean background radiation terms $r_i > 0$) and such images are excluded from the domain of (26). From similar considerations, one can also see that the tangent majorant fails to satisfy Condition (R3) at $\boldsymbol{y}$ near the boundary of the non-negative orthant.

A modification called ML-EM-3 was proposed in [15] that remedies the singularity issue in most practical cases. ML-EM-3 is based on complete data,

$$V_{ij}^{(1)}=\text{Poiss}\{a_{ij}(x_j+m_j)\},$$
$$V_i^{(2)}=\text{Poiss}\{r_i-\sum_j a_{ij}m_j\},$$
$$U_i=\sum_j V_{ij}^{(1)}+V_i^{(2)}$$

where $m_j \geq 0$ are parameters chosen to satisfy $\Sigma_j m_j \leq r_i$. With this complete data, (23) becomes,

$$\phi_{\boldsymbol{y}}(\boldsymbol{x})=\sum_i a_{ij}x_j - \boldsymbol{e}_j(\boldsymbol{y})(y_j+m_j)\log(x_j+m_j)+\text{const.}$$

In practice, one generally has $r_i > 0$ for all i, and hence can choose $m_j > 0$ for all j. In this case, $\phi_{[\cdot]}(\cdot)$ satisfies both (C3) and (C6). Moreover, the domain of these tangent majorants is the entire non-negative orthant. So, by Theorem 4.1, we can conclude that ML-EM-3 is asymptotically stationary.

In summary, although EM algorithms are special cases of MM, the applicability of the MM theory of this paper and its forerunners (e.g., [35]) depends on the choice of complete data.

6 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 regions of capture and convergence 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 6.4 that such algorithms are captured in any basin-shaped region in the graph of Ψ. If the basin contains a minimizer, then with suitable additional conditions (see Theorem 6.5), the entire basin is a region of convergence to that minimizer.

This is to be contrasted with the standard theory of gradient methods (e.g., steepest descent, Newton's, Levenberg-Marquardt). General gradient methods are driven by the minimization of quadratic approximations to Ψ. These approximations may not majorize Ψ as tangent majorants do. Standard analyses of regions of capture for gradient methods (e.g., [1, p. 51, Proposition 1.2.5] and [1, p. 90, Proposition 1.4.1(a)]) guarantee capture only in a neighbourhood where the derivatives of Ψ are sufficiently similar to the derivatives at the minimum. In Example 6.6, we illustrate how this region of capture can be a strict subset of a basin-shaped region around the minimizer. Thus, our findings suggest that connected tangent majorants lead to larger regions of capture/convergence for MM than for (non-MM) gradient methods. 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 generalizes usual notions of a basin-shaped region.

Definition 6.1 We say that a set $G \subset X$ is a *generalized basin* (with respect to the minimization problem (1)) if, for some $\boldsymbol{x} \in G$, the following is never violated

$$\Psi(\boldsymbol{x})<\Psi(\tilde{\boldsymbol{x}}), \quad \tilde{\boldsymbol{x}} \in \text{cl}(G) \cap \text{cl}(X \backslash G). \tag{27}$$

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

Thus, a point is well-contained in G if it has lower cost than any point $\tilde{\boldsymbol{x}}$ in the common boundary $\mathrm{cl}(G) \cap \mathrm{cl}(X \setminus G)$ between G and its complement.

Remark 6.2 (Special Cases of Basins) Definition 6.1 is worded so that $\mathrm{cl}(G) \cap \mathrm{cl}(X \setminus G)$ can be empty. Thus, for example, the whole feasible set X always constitutes a generalized basin, provided that it contains some $\boldsymbol{x}$. This is because $\mathrm{cl}(X) \cap \mathrm{cl}(X \setminus X)$ is empty, implying that (27) can never be violated by any $\boldsymbol{x}$.

Any sublevel set $G = \{\boldsymbol{x} \in X : \Psi(\boldsymbol{x}) \leq \tau\}$ is a generalized basin so long as τ is not the global minimum value of Ψ over X. Moreover, any global minimizer $\boldsymbol{x}^*$ is well-contained in G. For further discussion of the basic properties of generalized basins, see [18, pp. 104-5,136-7].

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 6.3 *Suppose that $\phi_{\boldsymbol{y}}(\cdot)$ is a tangent majorant that is connected on its domain $D(\boldsymbol{y}) \subset X_S$ and whose expansion point $\boldsymbol{y} \in X$ is well-contained in a generalized basin G. Suppose, further, that $\boldsymbol{x} \in X$ satisfies*

$$\begin{aligned} \boldsymbol{x}_S \in D(\boldsymbol{y}), \quad \boldsymbol{x}_{\tilde{S}} = \boldsymbol{y}_{\tilde{S}}, \\ \phi_{\boldsymbol{y}}(\boldsymbol{x}_S) \leq \phi_{\boldsymbol{y}}(\boldsymbol{y}_S), \end{aligned} \tag{28}$$

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

Proof. It is sufficient to show that $\boldsymbol{x} \in G$. For taking any $\tilde{x} \in \mathrm{cl}(G) \cap \mathrm{cl}(X \setminus G)$, and then combining (28), (3), and the fact that $\boldsymbol{y}$ is well-contained in G,

$$\Psi(\boldsymbol{x}) \leq \Psi(\boldsymbol{y}) < \Psi(\tilde{\boldsymbol{x}}), \tag{29}$$

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

$$\begin{aligned} [\mathbf{g}(\alpha)]_S &\in D(\boldsymbol{y}), \\ [\mathbf{g}(\alpha)]_{\tilde{S}} &= \boldsymbol{y}_{\tilde{S}}, \\ \phi_{\boldsymbol{y}}([\mathbf{g}(\alpha)]_S) &\leq \max\left\{\phi_{\boldsymbol{y}}(\boldsymbol{y}_S), \phi_{\boldsymbol{y}}(x_S)\right\} \\ &= \phi_{\boldsymbol{y}}(\boldsymbol{y}_S), \end{aligned} \tag{30}$$

where the equality in (30) is due to (28). Also, since $\mathbf{g}(0) = \boldsymbol{y} \in G$,

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

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

Therefore, from the rightmost inequality in (29), we have, with $\tilde{\boldsymbol{x}} = \boldsymbol{w}$,¡

$$\Psi(\boldsymbol{y}) < \Psi(\boldsymbol{w}) = \Psi_S^{\boldsymbol{y}}([\mathbf{g}(\alpha^*)]_S). \tag{31}$$

With (3), this implies that $\phi_{\boldsymbol{y}}([\mathbf{g}(\alpha^*)]_S) > \phi_{\boldsymbol{y}}(\boldsymbol{y}_S)$, contradicting (30).

The following consequence of Proposition 6.3 articulates a capture property for MM sequences.

Theorem 6.4 (Capture property of MM) *Suppose that* $\{\boldsymbol{x}^i\}$ *is an MM sequence generated by* (4) *and* (5). *In addition, suppose that some iterate* $\boldsymbol{x}^n$ *is well-contained in a generalized basin G and that the tangent majorant sequence* $\{\phi^i_{\boldsymbol{x}^i}(\cdot)\}_{i=n}^{\infty}$ *satisfies* (C1). *Then likewise* $\boldsymbol{x}^i$ *is well-contained in G for all* $i > n$.

Proof. The result follows from Proposition 6.3 and an obvious induction argument.

Finally, we obtain the principal result of this section.

Theorem 6.5 (Region of Convergence) *In addition to the assumptions of Theorem 6.4, 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* x^*. *Then* $\{\boldsymbol{x}^i\}$ *converges to* $\boldsymbol{x}^*$.

Proof. Since G is bounded, it follows from Theorem 6.4 that the sequence $\left\{\boldsymbol{x}^i\right\}_{i=1}^{\infty}$ lies in the compact set $K = \mathrm{cl}(G)$. Moreover, all limit points of $\{\boldsymbol{x}^i\}$ 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).

Example 6.6 (Gradient Methods with Ad Hoc Steps) Here we illustrate how gradient algorithms with *ad hoc* step size choices may have a radius of capture that does not cover the entire basin surrounding a minimum. Consider the following 1D cost function, also depicted in Figure 3,

$$\Psi(x) = \begin{cases} |x|^3, & |x| \le 1 \\ 3(|x| - 1)e^{-2(|x|-1)} + 1, & |x| > 1. \end{cases}$$

From Figure 3, one can see that the open interval (−1.5, 1.5) is a generalized basin in which all points are well-contained. It will be useful to note also that

$$\left|\frac{d\Psi(x)}{dx}\right| \le 2|x|, \; x\frac{d\Psi(x)}{dx} \ge 0, \quad \forall |x| \le 2/3. \tag{32}$$

Consider now the family of quadratic expansions,

$$\begin{aligned} Q_{s^k}(x; x^k) \;=\; & \Psi(x^k) \\ & + \tfrac{d\Psi(x^k)}{dx}(x - x^k) + \tfrac{1}{2s^k}(x - x^k)^2. \end{aligned} \tag{33}$$

The minimizer x^{k+1} of $Q_s{}^k(\cdot; x^k)$ is given by the gradient algorithm step,

$$x^{k+1} = x^k - s^k \frac{d\Psi(x^k)}{dx}. \tag{34}$$

Suppose we choose an arbitrary and constant step size parameter $s^k = 1$. From (32), it then follows that if x^k lies in the interval $|x| \leq 2/3$, the gradient step (34) cannot cross the origin to a point more distant from the origin than x^k. Consequently, all subsequent iterations of (34) remain trapped in the region $|x| \leq 2/3$ and hence also in the larger basin-like interval (−1.5, 1.5). (One can also show that the algorithm converges monotonically to the minimum when initialized in $|x| < 2/3$, but this is tangential to the point of this example.)

This capture property does not hold, however, for all starting points in (−1.5, 1.5). In particular, if $|x^k| = 1$, then $\left|\frac{d\Psi(x)}{dx}\right| = 3$. With $s^k = 1$, this means that the gradient step (34) will be large enough to put x^{k+1} in the region $|x| > 1.5$. Not only does this step escape from the basin (−1.5, 1.5), but the gradient algorithm will also never return there. For, once $|x^{k+1}| > 1.5$ the direction of the gradient will carry all subsequent iterations of (34) off toward infinity.

This example is to be contrasted with the choice $s^k = 6$. The iterations are then driven by the minimizations of the functions $Q_6(\cdot; x^k)$ which are not only convex, but are also tangent majorants. The latter can be verified, for instance, using Lemma 3.3(a) with $\gamma^+ = 6$ and $f = \Psi$. Theorem 6.4 therefore applies and shows that if x^k lies in (−1.5, 1.5), then all subsequent iterations of the algorithm will as well.

The distinction between these two cases is also illustrated graphically in Figure 3 with $x^k = -1$. There, we see that, since $Q_6(\cdot; -1)$ majorizes Ψ, its minimum is constrained by the graph of Ψ to lie in (−1.5, 1.5). Conversely, since $Q_1(\cdot; -1)$ does not majorize Ψ, its minimum is not constrained in this way.

The above example illustrates how non-MM gradient methods with constant step sizes can escape from a basin. However, it is also easy to see how this would be true even when conventional line search algorithms are used. Generally, line search methods can find any 1D stationary point along the search line and different 1D stationary points can lie in different generalized basins. For example, in Figure 1, the stationary points in the intervals [*A*, *B*] and [*B*, *C*] clearly lie in different generalized basins. The MM steps in Figure 1 respect the basin boundaries, consistent with Theorem 6.4, whereas a line search algorithm would not be expected to.

There are both positive and negative practical implications to Theorem 6.5. Since it is common to use convex (and hence connected) tangent majorants, it is essential for MM 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, if a moderately good initial guess of the solution is available, the chances of getting pulled toward the global solution may be higher. 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 becoming trapped at a sub-optimal local minimum can be high, depending on the preponderance of different minima in the graph of Ψ.

A potential application of Theorem 6.5 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 Ψ are locally minimized at successive increments of the parameter. Another example is the strategy of selecting a mesh of initial points and locally minimizing Ψ 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.

7 Summary

In this paper, we have revised the analysis of [15] 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 6.3, which proved a fundamental property of MM algorithms employing connected tangent majorants, namely that they remain 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 [12,19,18].

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 [34], 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 [20] may be a starting point for future analysis.

Acknowledgments

The authors would like to thank the reviewers for the time that they have invested in this article, as well as for their many helpful suggestions for its improvement.

References

1. Bertsekas, DP. Nonlinear programming. 2. Athena Scientific; Belmont: 1999.
2. Blake, A.; Zisserman, A. Visual reconstruction. MIT Press; Cambridge, MA: 1987.
3. Byrne C. Likelihood maximization for list-mode emission tomographic image reconstruction. IEEE Tr Med Im October;2001 20(10):1084–92.
4. Cadalli N, Arikan O. Wideband maximum likelihood direction finding and signal parameter estimation by using tree-structured EM algorithm. IEEE Trans Sig Proc January;1999 47(1):201–6.
5. Chung PJ, Böhme JF. Comparative convergence analysis of EM and SAGE algorithms in DOA estimation. IEEE Trans Sig Proc December;2001 49(12):2940–9.
6. Conn, AR.; Gould, NIM.; Toint, P. Trust-region Methods. MPS/SIAM Series on Optimization; Philadelphia: 2000.
7. De Pierro AR. On the relation between the ISRA and the EM algorithm for positron emission tomography. IEEE Tr Med Im June;1993 12(2):328–33.
8. De Pierro AR. A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography. IEEE Tr Med Im March;1995 14(1):132–137.
9. De Pierro AR. On the convergence of an EM-type algorithm for penalized likelihood estimation in emission tomography. IEEE Tr Med Im December;1995 14(4):762–5.

10. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J Royal Stat Soc Ser B 1977;39(1):1–38.
11. Erdoğan H, Fessler JA. Monotonic algorithms for transmission tomography. IEEE Tr Med Im September;1999 18(9):801–14.
12. Erdoğan, Hakan. PhD thesis. Univ. of Michigan; Ann Arbor, MI, 48109-2122: Jul. 1999 Statistical image reconstruction algorithms using paraboloidal surrogates for PET transmission scans.
13. Fessler JA, Clinthorne NH, Rogers WL. On complete data spaces for PET reconstruction algorithms. IEEE Trans Nuc Sci August;1993 40(4):1055–61.
14. Fessler JA, Hero AO. Space-alternating generalized expectation-maximization algorithm. IEEE Tr Sig Proc October;1994 42(10):2664–77.
15. Fessler JA, Hero AO. Penalized maximum-likelihood image reconstruction using space-alternating generalized EM algorithms. IEEE Tr Im Proc October;1995 4(10):1417–29.
16. Heiser, WJ. Convergent computation by iterative majorization: theory and applications in multidimensional data analysis. In: Krzanowski, WJ., editor. Recent Advances in Descriptive Multivariate Analysis. Royal Statistical Society Lecture Note Series. Oxford University Press; New York: 1995.
17. Huber, PJ. Robust statistics. Wiley; New York: 1981.
18. Jacobson, M. PhD thesis. Univ. of Michigan; Ann Arbor, MI, 48109-2122: 2006. Approaches to motion-corrected PET image reconstruction from respiratory gated projection data.
19. Jacobson MW, Fessler JA. Joint estimation of image and deformation parameters in motion-corrected PET. Proc IEEE Nuc Sci Symp Med Im Conf 2003;5:3290–4.
20. Jacobson, MW.; Fessler, JA. Properties of MM algorithms on convex feasible sets: extended version. Comm. and Sign. Proc. Lab., Dept. of EECS, Univ. of Michigan; Ann Arbor, MI, 48109-2122: Nov. 2004 Technical Report 353
21. Johnston LA, Krishnamurthy V. Finite dimensional smoothers for MAP state estimation of bilinear systems. IEEE Trans Sig Proc September;1999 47(9):2444–59.
22. Lange K. A gradient algorithm locally equivalent to the EM Algorithm. J Royal Stat Soc Ser B 1995;57 (2):425–37.
23. Lange K, Carson R. EM reconstruction algorithms for emission and transmission tomography. J Comp Assisted Tomo April;1984 8(2):306–16.
24. Lange K, Hunter DR, Yang I. Optimization transfer using surrogate objective functions. J Computational and Graphical Stat March;2000 9(1):1–20.
25. Liu CH, Wu YN. Parameter expansion scheme for data augmentation. J Am Stat Assoc December; 1999 94(448):1264–74.
26. Logothetis A, Carlemalm C. SAGE algorithms for multipath detection and parameter estimation in asynchronous CDMA systems. IEEE Trans Sig Proc November;2000 48(11):3162–74.
27. Meng XL, van Dyk D. The EM algorithm - An old folk song sung to a fast new tune. J Royal Stat Soc Ser B 1997;59(3):511–67.
28. Nelson LB, Poor HV. Iterative multiuser receivers for CDMA channels: an EM-based approach. IEEE Trans Comm December;1996 44(12):1700–10.
29. Nettleton D. Convergence properties of the EM algorithm in constrained parameter spaces. The Canadian Journal of Statistics 1999;27(3):639–48.
30. Ollinger JM, Goggin A. Maximum likelihood reconstruction in fully 3D PET via the SAGE algorithm. Proc IEEE Nuc Sci Symp Med Im Conf 1996;3:1594–8.
31. Ortega, JM.; Rheinboldt, WC. Iterative solution of nonlinear equations in several variables. Academic; New York: 1970.
32. Politte DG, Snyder DL. Corrections for accidental coincidences and attenuation in maximum-likelihood image reconstruction for positron-emission tomography. IEEE Trans Med Imag March; 1991 10(1):82–9.
33. Rockafellar, RT. Convex analysis. Princeton University Press; Princeton: 1970.
34. Shepp LA, Vardi Y. Maximum likelihood reconstruction for emission tomography. IEEE Tr Med Im October;1982 1(2):113–22.
35. Wu CFJ. On the convergence properties of the EM algorithm. Ann Stat March;1983 11(1):95–103.

36. Yu DF, Fessler JA, Ficaro EP. Maximum likelihood transmission image reconstruction for overlapping transmission beams. IEEE Tr Med Im November;2000 19(11):1094–1105.

37. Zangwill, W. Nonlinear programming, a unified approach. Prentice-Hall; NJ: 1969.

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

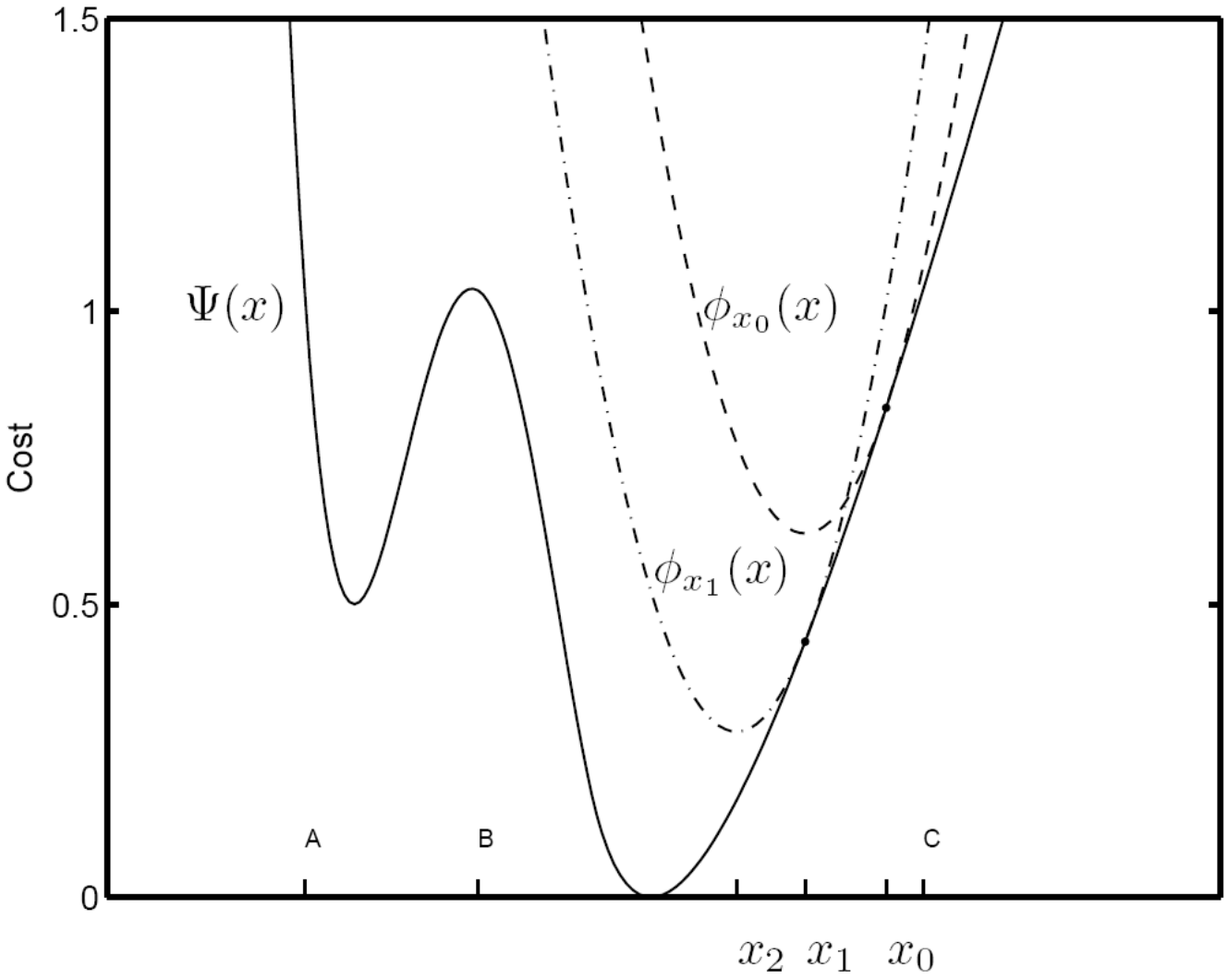

Figure 1.
One-dimensional illustration of an MM algorithm.

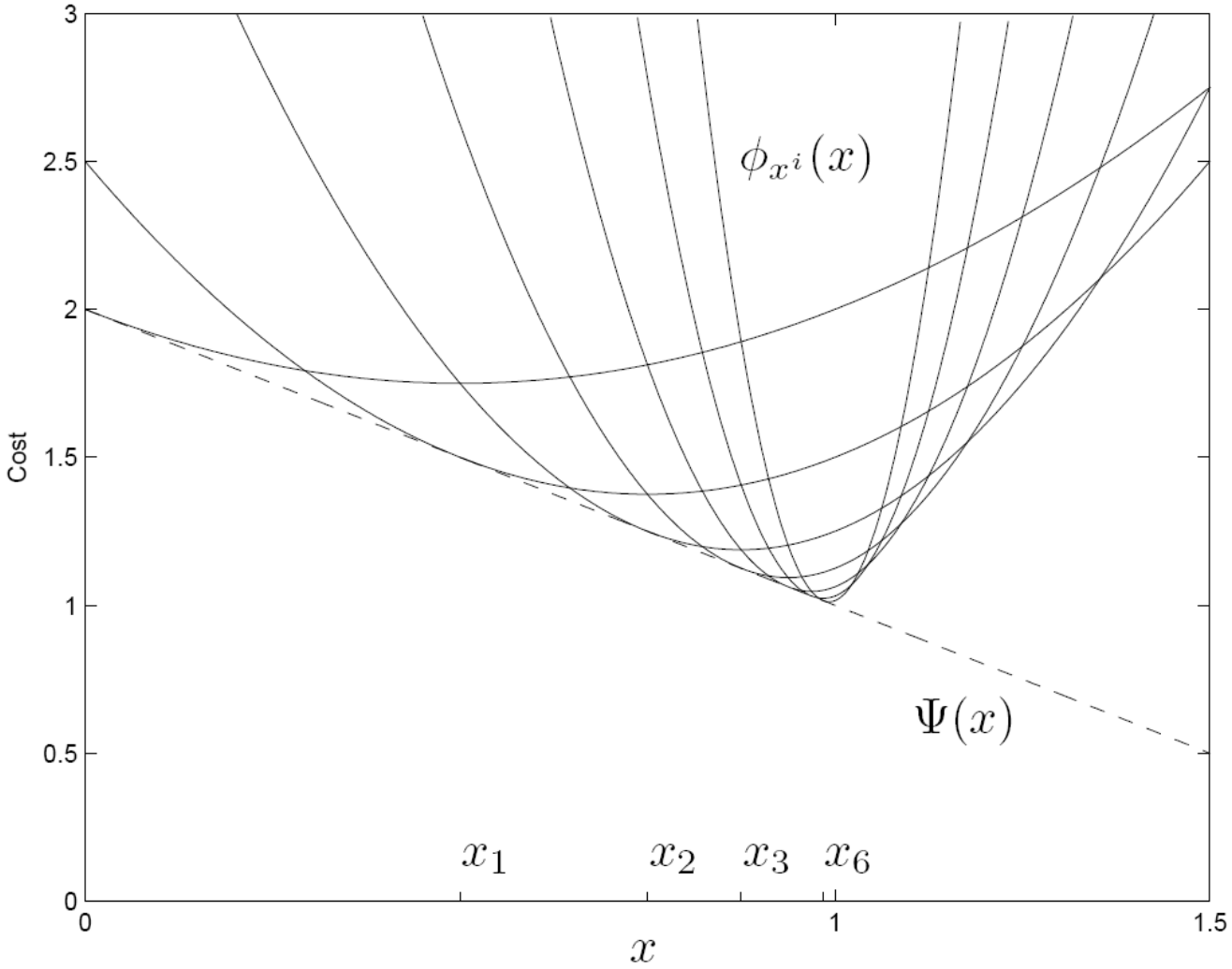

Figure 2.
Illustration of Example 4.2. The MM sequence $\{x^i\}$ converges to a non-stationary point. This is possible since the conditions of Theorem 4.1 are not satisfied.

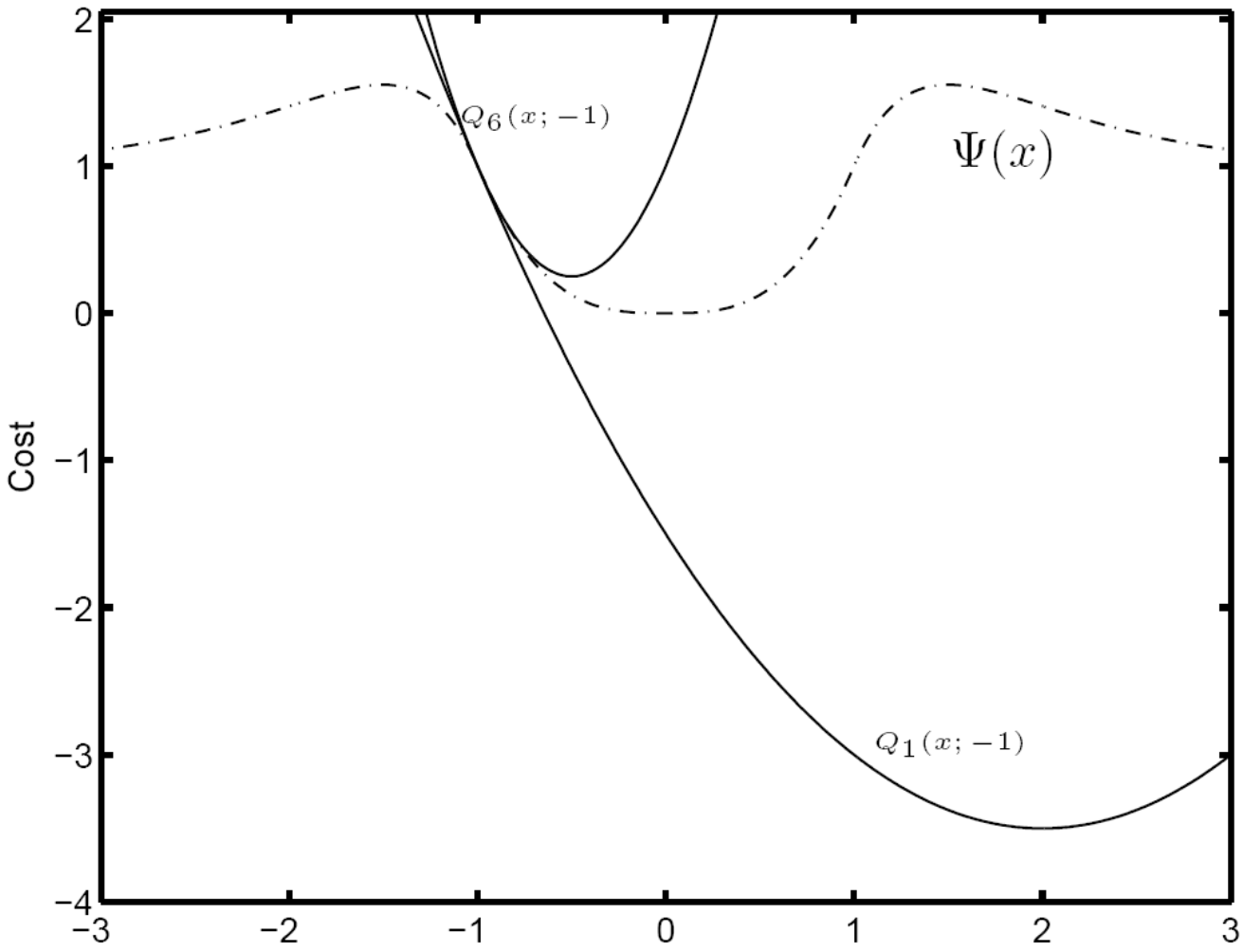

Figure 3.
Comparing the capture properties of MM with gradient methods.