Properties of MM Algorithms on Convex Feasible Sets: Extended Version

Mattthew W. Jacobson* Jeffrey A. Fessler

November 30, 2004

Abstract

We examine some properties of the Majorize-Minimize (MM) optimization technique, generalizing previous analyses. At each iteration of an MM algorithm, one constructs a *tangent majorant* function that majorizes the given cost function (possibly after adding a global constant) and is equal to it at the current iterate. The next iterate is taken from the set of minimizers of this tangent majorant function, resulting in a sequence of iterates that reduces the cost function monotonically. The article studies the behavior of these algorithms for problems with convex feasible sets but possibly non-convex cost functions. We analyze convergence properties in a standard way, showing first that the iteration sequence has stationary limit points under fairly mild conditions. We then obtain convergence results by adding discreteness assumptions on the stationary points of the minimization problem. The case where the stationary points form continua is also examined.

Local convergence results are also developed for algorithms that use *connected* (e.g., convex) tangent majorants. Such algorithms have the property that the iterates cannot leave any basin-like region containing the initial vector. This makes MM useful in various non-convex minimization strategies that involve basin-probing steps. This property also implies that cost function minimizers will locally attract the iterates over larger neighborhoods than can typically be guaranteed with other methods.

Our analysis generalizes previous work in several respects. Firstly, arbitrary convex feasible sets are permitted. The tangent majorant domains are also assumed convex, however they can be strict subsets of the feasible set. Secondly, the cost function and the tangent majorant functions are not required to be more than once continuously differentiable and the tangent majorants are often allowed to be non-convex as well. Thirdly, the technique of coordinate block alternation is considered for feasible sets of a more general Cartesian product form than in previous work.

1 Introduction

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

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

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 row vectors,[2] and Θ is a convex feasible set. An MM algorithm is one that reduces Φ monotonically by minimizing a succession of approximations to 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*

*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 iterative majorization or optimization transfer.

[2]When $\boldsymbol{\theta}$ is a scalar variable, we shall use the notation θ instead.

majorant $\phi(\cdot;\boldsymbol{\theta}^i)$. In the simplest case, 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. The concept is illustrated for a 1D cost function in Figure 1.

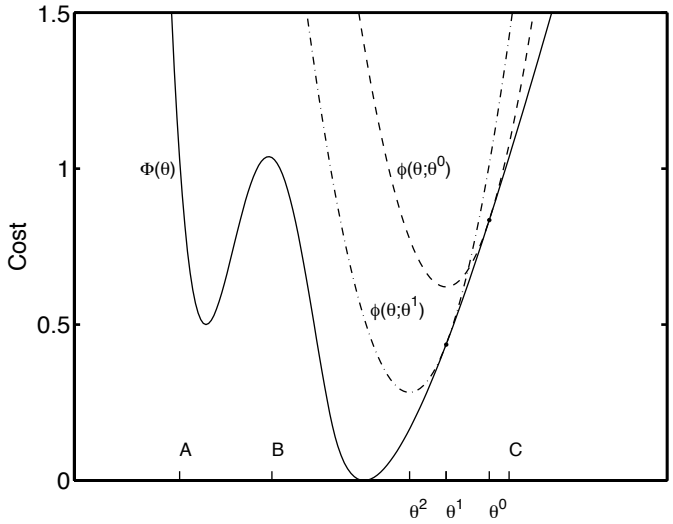

Figure 1: One-dimensional illustration of an MM algorithm.

More elaborate forms of MM have been considered [14, 15] 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 one to choose the form of the majorant generator at a given iteration based on the observed progress of the algorithm. 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 $\boldsymbol{\theta}$. This 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.

The subject of MM has appeared in a range of statistics literature over the years (e.g., [17, 18, 23, 25]). A prominent example is the Expectation Maximization (EM) methodology (commonly attributed to [12]) which is an application of MM to maximum likelihood estimation. In the field of statistical tomographic imaging, maximum likelihood estimators, and EM algorithms for implementing them, achieved early popularity in [33] and [24]. Since then, MM has become an algorithm design technique of increasing prevalence in this field, as is apparent in such works as [9, 10, 11, 35, 36]. The popularity of MM has also reached the commercial medical imaging world. In particular, several manufacturers now package their emission tomography scanners with software for an incremental gradient version (due to [19]) of an MM algorithm.

An MM algorithm is derived by choosing majorant generators of a particular form, and the form selected is based on the algorithm designer's insights into the shape of the cost function in question. One design goal is to choose majorant generators $\{\phi^i(\cdot;\cdot)\}$ so that the corresponding tangent majorants $\{\phi^i(\cdot;\boldsymbol{\theta}^i)\}$ approximate Φ as accurately as possible over as large a region of Θ as possible, all the while respecting the majorization requirement. One expects greater approximation accuracy to result in more rapid descent of $\{\Phi(\boldsymbol{\theta}^i)\}$. At the same time, one tries to choose the majorant generators so that the resulting tangent majorants $\{\phi^i(\cdot;\boldsymbol{\theta}^i)\}$ can be constructed and minimized in a computationally efficient manner. Typically, therefore, one generates tangent majorants that are convex (although we will not make this a global assumption in our analysis) as in Figure 1. Naturally, these design aims can conflict, and this has led to abundant literature that examines alternative designs for particular applications (e.g., [1, 2, 14, 15]).

Since the majorant generators, and hence the descent mechanism of the algorithm, are custom designed to suit the particular cost function at hand, MM algorithms have the potential to outperform more generic algorithms that

do not take the specifics of Φ into account. The latter would include textbook-variety derivative based methods (e.g., gradient projection and conditional gradient). These considerations are particularly important if Φ is of a general type that must be minimized routinely. If a cost function of a specific form is minimized on a repeat basis then, logically, algorithms like MM that are custom-designed to suit this form can be a worthwhile investment. This notion seems to account for the popularity of MM in statistical tomographic imaging. There, one routinely minimizes instances of loglikelihood or other cost functions of a common form, but corresponding to different sets of measured data.

As new minimization problems are encountered, the creativity of algorithm designers leads to new kinds of tangent majorants and corresponding MM algorithms. Because of this trend, it is desirable to have as general a theory as possible regarding the kinds of tangent majorants that one can use and the kinds of cost functions that one can apply them to. To our knowledge, the most significant recent advancements in the theory of MM are [14, 15, 30]. Prior to these works, MM algorithms studied in the literature used only a single majorant generator $\phi(\cdot;\cdot)$. Furthermore, convergence analyses mostly treated convergence to interior points of the feasible set only. Those analyses that did consider convergence to the boundary were valid only for specific problems and tangent majorants (e.g., [7, 33]). Beginning with [14, 15], the majorant generators were allowed to be iteration-dependent, resulting in various benefits like block alternation. In [14], only convergence to interior points of the feasible set was considered. The work of [15] extended [14] and tried to address convergence to the boundary when Θ was the non-negative orthant and when Φ was strictly convex. However due to an error (see Remark 4.6), the extension is valid only under much more restrictive assumptions than intended. In [30], a very unified treatment of constraints is given, one that covers even non-convex constraints. However, the analysis there does not allow iteration-dependent majorant generators, as in [14, 15].

Although more general than their forerunners, the collective fruits of [14, 15, 30] have lately proven insufficient to cover various contemporary problems and MM algorithms devised for them. One limitation is that [14, 15] do not treat constrained, non-convex problems comprehensively. Constraints are considered in [15], but the analysis contains an error as mentioned above. Furthermore, only the case in which Φ is strictly convex and in which Θ is the non-negative orthant is addressed. These restrictions apply to certain penalized likelihood minimization problems commonly encountered in emission tomography, and it is these problems that the authors of [15] had in mind when they made that analysis. However, more recent work in medical imaging applications has given rise to more complicated cost functions, that are not convex, but for which MM algorithms can be derived (e.g., [1, 35, 20, 21]). Moreover, the kinds of constraints now encountered in medical imaging go beyond mere non-negativity constraints. For example, in non-rigid image registration, feasible sets of a more complicated polyhedral form may be desired (see [22, p. 60]) to ensure physically realistic solutions. These more complicated constraints would be covered by [30] so long as iteration-independent majorant generators were used. However, iteration-dependent $\phi^i(\cdot;\cdot)$ are often desirable, at minimum because they allow block alternation, which can simplify algorithm design.

Further limitations of [15] are that the tangent majorants are required to be twice differentiable, strongly convex, and defined throughout the feasible set Θ. In [21], we derived several kinds of convex tangent majorants for a non-convex problem. However, the tangent majorant domains were strict subsets of Θ. Furthermore, many imaging problems involve cost functions with convex, but only once-differentiable penalty terms. This motivates certain MM algorithm designs with only once-differentiable convex tangent majorants. The limitations of [15] with respect to tomographic imaging problems are discussed in greater detail in Section 6. However, one could surely point to applications in other fields where these limitations are problematic as well.

In this article, we generalize the analysis in [15] resulting in a much more versatile algorithm design framework. Our analysis is more general than [15] in several respects. Firstly, arbitrary convex feasible sets are permitted in our framework. In this way, we marry some of the generality of [30] with that of [15]. Secondly, the tangent majorant domains can be strict subsets of the feasible set. Thirdly, the technique of block alternation is considered for feasible sets of a more general Cartesian product form. Fourth, Φ is not required to be convex and the tangent majorants $\{\phi^i(\cdot;\boldsymbol{\theta}^i)\}$ are often allowed to be non-convex as well. Finally, our analysis does not require the cost function and

the tangent majorants to be more than once continuously differentiable. Since we treat only convex feasible sets, the scope of possible constraints is more restrictive than in [30]. However, unlike [30], the use of iteration-dependent tangent majorants is covered in the presence of constraints (and hence also, the error in [15] is remedied).

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. In Section 4, we analyze the asymptotic behavior of MM. Results are developed showing stationarity of MM limit points in both the block alternating and non-block alternating case. In each case, two sets of conditions are applied. One set involves traditional kinds of continuity assumptions on the majorant generators. None of these conditions are more restrictive than [15]. The second set involves local curvature bounds on the tangent majorants. We then deduce convergence of MM in norm (see Theorem 4.5) in a standard way by imposing discreteness assumptions on the set of stationary points of (1.1). Non-isolated stationary points are not generally stable (cf. [5, 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 Section 7.

When Φ is non-convex, local behavior of MM becomes an important issue and is the subject of Section 5. Here we restrict our attention to the case where the tangent majorants have path-connected level sets (e.g., as in the case when the tangent majorants are convex). For this family of tangent majorants, it is shown that the iterates $\{\boldsymbol{\theta}^i\}$ are captured by local basin-like regions in the graph of Φ. This property allows us to derive a local analogue, namely Theorem 5.6, to the convergence described in Theorem 4.5. An implication of Theorem 5.6 is that local convergence will occur over a larger neighborhood of a global minimizer than can typically be guaranteed with more standard algorithms. In addition, various non-convex minimization strategies involve basin-probing steps. The basin capture property of connected tangent majorants makes MM algorithms particularly suitable for implementing these steps.

Sections 6 and 7 deal with topics of special interest. In Section 6, we discuss the relevance of our results to two recent problems in tomographic imaging. One problem is the joint estimation of radiotracer concentration and attenuation variables given both emission and transmission tomography data. The second problem is the estimation of radiotracer concentration given emission tomography measurements corrupted by anatomical motion. Section 7 considers the question of whether convergence in norm will occur to stationary points that are *not* isolated. This question might be of theoretical interest only, because of stability issues alluded to above. However, various MM algorithms ([7, 33, 28]) have been observed to converge in norm, even to non-isolated minima, so specialists are apt to wonder if this behavior can be proved more generally. For single variable problems, we show, in Theorem 7.2, that, if the iterate sequence $\{\theta^i\}$ is bounded and the $\{\phi^i(\cdot;\theta^i)\}$ have a uniform lower curvature bound, then convergence is assured, regardless of whether or not continua of stationary points exist. Moreover, we argue (see Example 7.1) that these conditions are about as weak as one can consider. For multi-variable problems, we find that these conditions are insufficient for convergence. This is demonstrated via an example in $\mathbb{R}^2$ (see Example 7.5). At this time, we are unable to extend Theorem 7.2 to the multi-variable case. However, the aforementioned 2D example provides some intuition as to what conditions may be sufficient, and so is a useful starting point for future work.

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 concatenation 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 $\Phi(\boldsymbol{\theta}_{\mathcal{S}}, \boldsymbol{\theta}_{\tilde{\mathcal{S}}})$.

The following example illustrates these indexing conventions and gives the flavor of the problems that this framework is meant to accommodate. For more extensive examples, see Section 6.

Example 2.1 Consider the following minimization problem with $\Theta_1 \subset \mathbb{R}^3$, $\Theta_2 \subset \mathbb{R}$, and $\Theta_3 \subset \mathbb{R}^2$ specified by the constraints as indicated.

$$\text{min.} \quad \Phi(\theta_1, \ldots, \theta_6) = \left\{ \sum_{j=1}^{6} \theta_j \right\} - \log \left\{ \sum_{j=1}^{6} \theta_j \right\}$$

subject to

$$\begin{aligned}\Theta_1 &\left\{ \begin{array}{l} \theta_1,\, \theta_2,\, \theta_3 \geq 0 \\ \theta_1 + \theta_2 + \theta_3 = 10 \\ \theta_1 + 2\theta_2 = 5 \end{array} \right. \\ \Theta_2 &\left\{ \; 1 \leq \theta_4 \leq 6 \right. \\ \Theta_3 &\left\{ \; \theta_5^2 + \theta_6^2 \leq 9. \right.\end{aligned}$$

Thus, we obtain the decomposition $\Theta = \Theta_1 \times \Theta_2 \times \Theta_3$, a particular case of (2.1) with $p = 6$, $M = 3$, $p_1 = 3$, $p_2 = 1$, and $p_3 = 2$. Given $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_6)$ then, according to our indexing conventions, $\boldsymbol{\theta}_1 = (\theta_1, \theta_2, \theta_3)$, $\boldsymbol{\theta}_2 = (\theta_4)$, and $\boldsymbol{\theta}_3 = (\theta_5, \theta_6)$. If, for example, we let $\mathcal{S} = \{1, 3\}$, then $\boldsymbol{\theta}_{\mathcal{S}} = (\theta_1, \theta_2, \theta_3, \theta_5, \theta_6)$, $\Theta_{\mathcal{S}} = \Theta_1 \times \Theta_3$, and $\mathbb{R}_{\mathcal{S}} = \mathbb{R}^5$. Also, $\tilde{\mathcal{S}} = \{2\}$, $\boldsymbol{\theta}_{\tilde{\mathcal{S}}} = \theta_4$, $\Theta_{\tilde{\mathcal{S}}} = \Theta_2$, and $\mathbb{R}_{\tilde{\mathcal{S}}} = \mathbb{R}$. Observe, as in the case of Θ_1 above, that any Θ_m can have an empty interior in its corresponding space $\mathbb{R}^{p_m}$. That is, we are not assuming that the Θ_m are *solid* subsets of $\mathbb{R}^{p_m}$.

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

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

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. An equivalent way of expressing (2.2) is

$$\min_{\boldsymbol{\xi} \in D(\bar{\boldsymbol{\theta}})} \{ \phi(\boldsymbol{\xi}; \bar{\boldsymbol{\theta}}) - \Phi(\boldsymbol{\xi}, \bar{\boldsymbol{\theta}}_{\tilde{\mathcal{S}}}) \} = \phi(\bar{\boldsymbol{\theta}}_{\mathcal{S}}; \bar{\boldsymbol{\theta}}) - \Phi(\bar{\boldsymbol{\theta}}). \tag{2.3}$$

Certain properties of tangent majorants (see, for example, Note A.2) are more obvious from this definition.

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 [15]. This assumption does not hold, however, for the MM algorithms in [9, 21]. 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,

$$\boldsymbol{\theta}^{i+1}_{\mathcal{S}^i} \in \operatorname*{argmin}_{\boldsymbol{\xi} \in D^i(\boldsymbol{\theta}^i)} \phi^i(\boldsymbol{\xi};\boldsymbol{\theta}^i) \tag{2.4}$$

$$\boldsymbol{\theta}^{i+1}_{\bar{\mathcal{S}}^i} = \boldsymbol{\theta}^i_{\bar{\mathcal{S}}^i}. \tag{2.5}$$

Here, we assume that the set of minimizers in (2.4) 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 $\phi^i(\boldsymbol{\theta}_{\mathcal{S}^i};\bar{\boldsymbol{\theta}}) = \Phi(\boldsymbol{\theta}_{\mathcal{S}^i}, \bar{\boldsymbol{\theta}}_{\bar{\mathcal{S}}^i})$ for all i, (2.4) and (2.5) become a generalization of block coordinate descent (e.g., [5, p. 267]), in which the coordinate blocks are not necessarily disjoint. By virtue of (2.2) and (2.4), $\{\Phi(\boldsymbol{\theta}^i)\}$ 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

$$\phi(\boldsymbol{\xi};\bar{\boldsymbol{\theta}}) \geq \Phi(\boldsymbol{\xi}, \bar{\boldsymbol{\theta}}_{\bar{\mathcal{S}}}) \qquad \forall \boldsymbol{\xi} \in D(\bar{\boldsymbol{\theta}}), \tag{2.6}$$

$$\phi(\bar{\boldsymbol{\theta}}_{\mathcal{S}};\bar{\boldsymbol{\theta}}) = \Phi(\bar{\boldsymbol{\theta}}). \tag{2.7}$$

That is, $\phi(\cdot;\bar{\boldsymbol{\theta}})$ majorizes $\Phi(\cdot, \bar{\boldsymbol{\theta}}_{\bar{\mathcal{S}}})$ over $D(\bar{\boldsymbol{\theta}})$ and is tangent to it in the sense that equality holds[3] at $\bar{\boldsymbol{\theta}}_{\mathcal{S}}$. These considerations motivate our choice of the term *tangent majorant*.[4] We abbreviate (2.6) and (2.7) via the notation,

$$\phi(\cdot;\bar{\boldsymbol{\theta}}) \underset{D(\bar{\boldsymbol{\theta}})}{\overset{\bar{\boldsymbol{\theta}}}{\succ}} \Phi(\cdot, \bar{\boldsymbol{\theta}}). \tag{2.8}$$

The relational operator $\underset{D(\bar{\boldsymbol{\theta}})}{\overset{\bar{\boldsymbol{\theta}}}{\succ}}$ describes a partial ordering between functions on $D(\bar{\boldsymbol{\theta}})$. 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 (2.4) and (2.5). 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 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. [14, 15]). 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

$$\Phi(\boldsymbol{\xi}) - \Phi(\bar{\boldsymbol{\theta}}) \leq \phi(\boldsymbol{\xi};\bar{\boldsymbol{\theta}}) - \phi(\bar{\boldsymbol{\theta}};\bar{\boldsymbol{\theta}}) \qquad \forall \boldsymbol{\xi} \in D(\bar{\boldsymbol{\theta}}), \tag{2.9}$$

while (2.4) and (2.5) reduce to

$$\boldsymbol{\theta}^{i+1} \in \operatorname*{argmin}_{\boldsymbol{\theta} \in D^i(\boldsymbol{\theta}^i)} \phi^i(\boldsymbol{\theta};\boldsymbol{\theta}^i), \tag{2.10}$$

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

[4] 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.

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 (2.9). Block alternation can also provide faster alternatives to certain non-block alternating algorithm designs [14]. To apply block alternation meaningfully, Θ must be decomposable into the Cartesian product form (2.1) with $M > 1$. When this is not the case, one can sometimes find a subset $\Theta' \subset \Theta$ that does have this form, and which contains a solution to (1.1). One can then reformulate the problem by substituting Θ' for Θ.

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

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

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

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

to denote certain constrained balls.

Given vectors $x_j \in \mathbb{R}^d$ and real scalars α_j, $j = 1, \ldots, N$ for which $\sum_{j=1}^N \alpha_j = 1$, the form $\sum_{j=1}^N \alpha_j x_j$ is called an *affine combination* of these vectors. A set $G \in \mathbb{R}^d$ is called affine if it contains all affine combinations of its members. Given a set $G \subset \mathbb{R}^d$, the *affine hull* $\text{aff}\,(G)$ of G is defined as the smallest affine set containing G or, equivalently, the set of all affine combinations of elements in G. A point $x \in \mathbb{R}^d$ is said to lie in the *relative interior* $\text{ri}\,(G)$ if there exists some $r > 0$ such that $B^d(r, x) \cap \text{aff}\,(G) \subset G$. When $\text{aff}\,(G) = \mathbb{R}^d$, then $\text{ri}\,(G)$ is simply the interior of G. We denote the closure of G by $\text{cl}\,(G)$. Recall that $\text{cl}\,(G)$ is the smallest closed set containing G or, equivalently, the set of all limits of sequences of points in G. The notation ∂G will denote the relative boundary, $\text{cl}\,(G) \setminus \text{ri}\,(G)$.

A set $G \in \mathbb{R}^d$ is said to be *discrete* if for each $x \in G$, there exists an $r > 0$ such that $B^d(r, x) \cap G = \{x\}$. The points in G are then said to be *isolated*. A function $f : D \subset \mathbb{R}^d \rightarrow \mathbb{R}$ is said to be *connected* on a set $D_0 \subset D$ if (see [31, p. 98]), given any $x,\, y \in D_0$, there exists a continuous function $g : [0,\, 1] \rightarrow D_0$ 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 $g : [0,\, 1] \rightarrow C$ 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 [31, p. 99]) that a function is connected if and only if its level 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,[5]

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

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

(R1.1) The sequence $\{\boldsymbol{\theta}^i\}$ is contained in a compact (i.e., closed and bounded) subset of Θ.

(R2) For each i and $\boldsymbol{\xi} \in \Theta_{\mathcal{S}^i}$, the non-normalized directional derivative

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

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

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

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) For every i and $\bar{\boldsymbol{\theta}} \in \Theta_{\mathcal{S}^i}$,

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

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

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

[5] Recall that (3.1) is a more fundamental first order condition than the KKT conditions. Condition (3.1) is necessary for $\boldsymbol{\theta}^*$ to be a local minimizer of Φ and, if Φ is convex, sufficient for $\boldsymbol{\theta}^*$ to be a global minimizer (see Proposition 2.1.2 in [5, p. 194]).

(C3) 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) 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) $\lim_{i\to\infty} ||\boldsymbol{\theta}^{i+1} - \boldsymbol{\theta}^i|| = 0$.

(C5.1) 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 3.5(c).

(C6) 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)),

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

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 level sets $\mathrm{lev}_\tau\, \Phi \stackrel{\triangle}{=} \{\boldsymbol{\theta} \in \Theta \,:\, \Phi(\boldsymbol{\theta}) \leq \tau\}$ of Φ are closed, which is often a straightforward exercise. In the latter case, with $\tau_0 = \Phi(\boldsymbol{\theta}^0)$, the level set $\mathrm{lev}_{\tau_0}\, \Phi$ 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 $\mathrm{lev}_{\tau_0}\, \Phi$) is compact, then (R1.1) holds. The closure or compactness of level 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., [31, 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 [15].

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

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

3.3 More Preliminaries

We now give several lemmas that facilitate the analysis in this paper. Most of these lemmas are slight generalizations of existing results.

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 let $y \in D$.*

(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

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

(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

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

Proof. Assume first that the assumptions of part (a) hold. Since D is convex, the scalar function $f(y + t(x - y))$ is defined for $t \in [0,\, 1]$. Moreover, since f is continuously differentiable, then the directional derivative $\langle \nabla f(y + t(x - y)),\, x - y \rangle$ is Riemann integrable as a function of t in the interval $[0,\, 1]$. Thus, by the fundamental theorem of calculus,

$$\begin{aligned}
f(x) - f(y) &= \int_0^1 \langle \nabla f(y + t(x - y)),\, x - y \rangle \, \mathrm{d}t \\
&= \langle \nabla f(y),\, x - y \rangle + \int_0^1 \langle \nabla f(y + t(x - y)) - \nabla f(y),\, x - y \rangle \, \mathrm{d}t \\
&\leq \langle \nabla f(y),\, x - y \rangle + \gamma^+ ||x - y||^2 \int_0^1 t \, \mathrm{d}t \\
&= \langle \nabla f(y),\, x - y \rangle + \frac{1}{2}\gamma^+ ||x - y||^2.
\end{aligned}$$

Virtually identical manipulations, but with reversed inequalities, establish part (b). □

Remark 3.4 Results similar to Lemma 3.3 are often proved under slightly stronger assumptions (e.g., Proposition A.24 in [5, p. 667]).

Lemma 3.5 (Consequences of the existence 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, in addition, (C5.1) *holds then, $\lim_{i \to \infty} ||\boldsymbol{\theta}^i - \boldsymbol{\theta}^{i+1}|| = 0$.*

Proof.

(a) Let $\{\boldsymbol{\theta}^{i_k}\}$ be a subsequence converging to $\boldsymbol{\theta}^*$. The continuity of Φ then implies that $\Phi(\boldsymbol{\theta}^{i_k}) \to \Phi(\boldsymbol{\theta}^*)$. Since $\{\Phi(\boldsymbol{\theta}^i)\}$ is monotonically non-increasing, we conclude that $\{\Phi(\boldsymbol{\theta}^i)\} \searrow \Phi(\boldsymbol{\theta}^*)$.

(b) Immediate from part (a) and the uniqueness of the limit of $\{\Phi(\boldsymbol{\theta}^i)\}$.

(c) Since (C5.1) holds, then Lemma 3.3(b) applies with $f = \phi^i(\cdot;\boldsymbol{\theta}^i)$, $D = D^i(\boldsymbol{\theta}^i)$, $x = \boldsymbol{\theta}^i_{\mathcal{S}^i}$, and $y = \boldsymbol{\theta}^{i+1}_{\mathcal{S}^i}$,

$$\phi^i(\boldsymbol{\theta}^i_{\mathcal{S}^i};\boldsymbol{\theta}^i) - \phi^i(\boldsymbol{\theta}^{i+1}_{\mathcal{S}^i};\boldsymbol{\theta}^i) \geq \left\langle \nabla^{10}\phi^i(\boldsymbol{\theta}^{i+1}_{\mathcal{S}^i};\boldsymbol{\theta}^i),\, \boldsymbol{\theta}^i_{\mathcal{S}^i} - \boldsymbol{\theta}^{i+1}_{\mathcal{S}^i} \right\rangle + \frac{1}{2}\gamma^- ||\boldsymbol{\theta}^i_{\mathcal{S}^i} - \boldsymbol{\theta}^{i+1}_{\mathcal{S}^i}||^2 \tag{3.5}$$

for all i. Since $\phi^i(\cdot;\boldsymbol{\theta}^i)$ is convex with minimizer $\boldsymbol{\theta}^{i+1}$,

$$\left\langle \nabla^{10}\phi^i(\boldsymbol{\theta}^{i+1}_{\mathcal{S}^i};\boldsymbol{\theta}^i),\, \boldsymbol{\theta}^i_{\mathcal{S}^i} - \boldsymbol{\theta}^{i+1}_{\mathcal{S}^i} \right\rangle \geq 0.$$

In addition, due to (2.2),

$$\phi^i(\boldsymbol{\theta}^i_{\mathcal{S}^i};\boldsymbol{\theta}^i) - \phi^i(\boldsymbol{\theta}^{i+1}_{\mathcal{S}^i};\boldsymbol{\theta}^i) \leq \Phi(\boldsymbol{\theta}^i) - \Phi(\boldsymbol{\theta}^{i+1}).$$

Incorporating these observations into (3.5),

$$||\boldsymbol{\theta}^i_{\mathcal{S}^i} - \boldsymbol{\theta}^{i+1}_{\mathcal{S}^i}||^2 \leq \frac{2}{\gamma^-}(\Phi(\boldsymbol{\theta}^{i+1}) - \Phi(\boldsymbol{\theta}^i)),$$

and since $\boldsymbol{\theta}^i_{\tilde{\mathcal{S}}^i} = \boldsymbol{\theta}^{i+1}_{\tilde{\mathcal{S}}^i}$, this is equivalent to

$$||\boldsymbol{\theta}^i - \boldsymbol{\theta}^{i+1}||^2 \leq \frac{2}{\gamma^-}(\Phi(\boldsymbol{\theta}^{i+1}) - \Phi(\boldsymbol{\theta}^i)).$$

Due to part (a), the limit of the right hand side of this inequality is 0 and the result follows. □

Lemma 3.6 (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 $\mathcal{C}$.

Proof. Since $\{\boldsymbol{\theta}^i\}$ lies in a compact set, convergence is established by showing that S is a singleton. The fact that S contains only stationary points implies that $S \subset \mathcal{C}$. Therefore, in case (a) it is readily seen that S is a singleton. Alternatively, suppose that (b) is true. Then, since $S \subset \mathcal{C}$ and $\mathcal{C}$ is discrete, then likewise S is discrete. In addition, since $\mathcal{K}$ is bounded and (C5) holds, then S is also connected (see p.173 of [32]). Since S is both discrete and connected, it is a singleton. □

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 4.5). 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., [34, 15, 30]). 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. Note also that, in Theorem 4.1, the tangent majorants can be non-convex.

Theorem 4.1 (Asymptotically stationary: non-block alternating case) *Suppose that all $\mathcal{S}^i = \{1, \ldots, M\}$, that $\{\boldsymbol{\theta}^i\}$ is an MM sequence generated by* (2.10)*, 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,

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

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

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

and the scalar functions

$$h_k(t) \stackrel{\triangle}{=} \phi^{i_k}(\boldsymbol{\theta}^{i_k} + t\mathbf{s}^k; \boldsymbol{\theta}^{i_k}) - \left[\phi^{i_k}(\boldsymbol{\theta}^{i_k}; \boldsymbol{\theta}^{i_k}) - \Phi(\boldsymbol{\theta}^{i_k})\right], \qquad t \in [0,\ \bar{t}\,]. \tag{4.4}$$

Due to (R3) and (4.2), all h_k are well-defined on the common interval indicated. The next several inequalities follow from (2.10), (2.9), and Lemma 3.5(a), respectively,

$$h_k(t) \geq \phi^{i_k}(\boldsymbol{\theta}^{i_k+1}; \boldsymbol{\theta}^{i_k}) - \left[\phi^{i_k}(\boldsymbol{\theta}^{i_k}; \boldsymbol{\theta}^{i_k}) - \Phi(\boldsymbol{\theta}^{i_k})\right] \geq \Phi(\boldsymbol{\theta}^{i_k+1}) \tag{4.5}$$

$$\geq \Phi(\boldsymbol{\theta}^*). \tag{4.6}$$

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

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

Passing to the limit in k,

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

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 $\{\boldsymbol{\theta}^{i_k}\}$ 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 $t \in [0,\ \bar{t}\,]$,

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

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 $r_{\mathcal{Z}} \leq \bar{t}$ 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,

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

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

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

However, $\dot{h}_k(0) \leq c$ 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 $\dot{h}^*(0) \leq c < 0$, contradicting (4.9). □

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. From this example, one can see that the requirements of Theorem 4.1 are not excessive and give meaningful guidelines for the design of majorant generators.

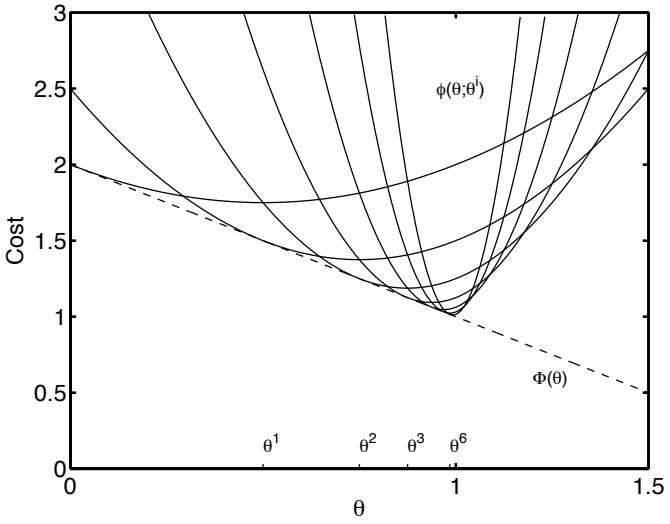

Figure 2: Illustration of Example 4.2. The MM sequence $\{\theta^i\}$ converges to a non-stationary point. This is possible since the conditions of Theorem 4.1 are not satisfied.

Example 4.2 Consider the case $\mathbb{R}^p = \mathbb{R}$, $\Theta = [0, 1.5]$, and $\Phi(\theta) = 2 - \theta$. Take $\theta^0 = 0$ and let $\{\theta^i\}$ be the sequence generated via (2.10) with

$$\phi^i(\theta; \bar{\theta}) = \phi(\theta; \bar{\theta}) \triangleq c(\bar{\theta})(\theta - \bar{\theta})^2 + \Phi(\theta)$$

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

It is immediate from the definition of $\phi(\theta;\bar{\theta})$ that every $\phi(\cdot;\bar{\theta})$ is a true tangent majorant. The resulting sequence of iterates $\{\theta^i\}$ and tangent majorants $\phi(\cdot;\theta^i)$ are depicted for several iterations in Figure 2. By induction, one can readily determine that $\theta^i = 1 - 2^{-i}$. Hence, $\{\theta^i\}$ 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 $\{c(\theta^i) = 2^i\}$ tend to infinity. Also, $\phi(\theta;\cdot)$ is discontinuous at $\bar{\theta} = 1$, so (C3) is not satisfied. Consequently, the hypothesis of Theorem 4.1 does not hold.

Remark 4.3 The kind of discontinuities exhibited in Example 4.2 can arise in EM majorant generators because of a discontinuity in the KL distance (see Note A.4).

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 4.4 (Asymptotic stationarity: block alternating case) *Suppose that* $\{\boldsymbol{\theta}^i\}$ *is an MM sequence generated by* (2.4) *and* (2.5) *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 $\boldsymbol{\theta}' \neq \boldsymbol{\theta}^* \in \Theta$ and an $m \in \{1, \ldots, M\}$, such that

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

and such that $\boldsymbol{\theta}'_{\tilde{m}} = \boldsymbol{\theta}^*_{\tilde{m}}, \forall \tilde{m} \neq m$. 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 $\{\boldsymbol{\theta}^{i_k + J_k}\}$, $J_k \leq J$ 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,

$$||\boldsymbol{\theta}' - \boldsymbol{\theta}^{i_k}|| \geq \min(r, ||\boldsymbol{\theta}' - \boldsymbol{\theta}^*||/2) \stackrel{\triangle}{=} \bar{t}.$$

and

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

for some $c < 0$. Now define

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

and

$$h_k(t) \stackrel{\triangle}{=} \phi^{(m)}(\boldsymbol{\theta}^{i_k}_{\mathcal{S}^{(m)}} + t\mathbf{s}^k; \boldsymbol{\theta}^{i_k}) - \left[\phi^{(m)}(\boldsymbol{\theta}^{i_k}_{\mathcal{S}^{(m)}}; \boldsymbol{\theta}^{i_k}) - \Phi(\boldsymbol{\theta}^{i_k})\right], \qquad t \in [0,\, \bar{t}\,].$$

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 4.5 (Convergence of MM sequences to isolated stationary points) *Suppose* $\{\boldsymbol{\theta}^i\}$ *is an MM sequence satisfying* (R1.1)*, as well as the conditions of either Theorem 4.1 or Theorem 4.4. 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 4.4. The result then follows from Lemma 3.6. □

Remark 4.6 The convergence analysis in [15] 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 [15] are indeed valid, even if the arguments are not. This follows from Theorem 4.5(a) above, which implies convergence under conditions no stronger than those assumed in [15].

5 The Capture Property of Connected Tangent Majorants

When Φ is non-convex, one often thinks of its graph as consisting of many *capture basins*, i.e., high dimensional analogues of valley-shaped regions, each containing a local minimum. In this section, we show that, if the tangent majorants are connected, the MM algorithm will remain confined to such a region. This property, which we call the *capture property* of MM, has a variety of consequences that we shall discuss.

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

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

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

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 capture 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, nor that Φ attain its maximum over G in ∂G. However, it is straightforward to show that any generalized

capture basin G must have the same dimension as Θ, in the sense that $\text{aff}\,(G) = \text{aff}\,(\Theta)$ (see Note A.5). Thus, for example, if $\Theta = \mathbb{R}^2$, no line segment inside Θ can constitute a generalized capture basin. This is consistent with common intuition.

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 capture 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. For the case where the tangent majorant takes arguments in $\mathbb{R}^p$ (i.e., excluding tangent majorants used for block alternation), Figure 3 shows how this result can be interpreted in terms of the tangent majorant level sets.

Proposition 5.3 *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 capture basin G. Suppose, further, that $\boldsymbol{\theta} \in \Theta$ satisfies*

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

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

Proof. It is sufficient to show that $\boldsymbol{\theta} \in G$. For taking any $\tilde{\boldsymbol{\theta}} \in \text{cl}\,(G) \cap \text{cl}\,(\Theta \setminus G)$, 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 $\mathbf{g} : [0,\ 1] \to \Theta$ with $\mathbf{g}(0) = \bar{\boldsymbol{\theta}}$, $\mathbf{g}(1) = \boldsymbol{\theta}$, and such that, for all $\alpha \in (0,\ 1)$, one has

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

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 $\boldsymbol{\psi} = \mathbf{g}(\alpha^*)$.

We now argue that $\boldsymbol{\psi} \in \text{cl}\,(G) \cap \text{cl}\,(\Theta \setminus G)$. Firstly, due to the definition of α^*, there must exist a sequence $0 \leq \hat{\alpha}_j \leq \alpha^*$, $j = 1, 2, \ldots$ with $\hat{\alpha}_j \to \alpha^*$. Since $\mathbf{g}(\hat{\alpha}_j) \to \boldsymbol{\psi}$, by continuity, and all $\mathbf{g}(\hat{\alpha}_j) \in G$, it follows that $\boldsymbol{\psi} \in \text{cl}\,(G)$. Secondly, the definition of α^* also implies that, if $\alpha^* < \alpha \leq 1$, then $\mathbf{g}(\alpha) \in \Theta \setminus G$. Together with the fact that $\mathbf{g}(1) = \boldsymbol{\theta} \in \Theta \setminus G$, it follows that there is a sequence $\alpha^* \leq \check{\alpha}_j \leq 1$, $j = 1, 2, \ldots$ with $\check{\alpha}_j \to \alpha^*$ and $\mathbf{g}(\check{\alpha}_j) \in \Theta \setminus G$. Since $\mathbf{g}(\check{\alpha}_j) \to \boldsymbol{\psi}$, we have that $\boldsymbol{\psi} \in \text{cl}\,(\Theta \setminus G)$ as well. We conclude that $\boldsymbol{\psi} \in \text{cl}\,(G) \cap \text{cl}\,(\Theta \setminus G)$ as claimed. 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 5.3, we obtain the following result as an immediate consequence. It articulates the capture property of MM for generalized capture basins.

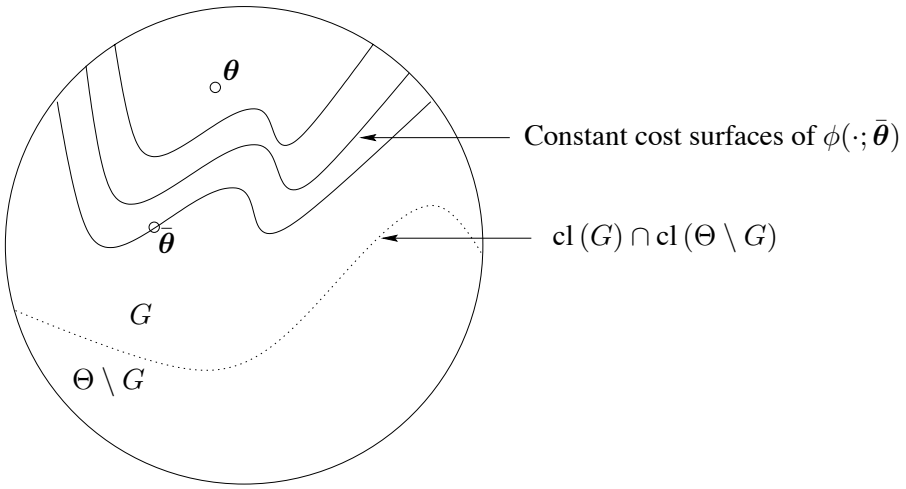

Figure 3: Illustration of the level set containment property of connected tangent majorants. Suppose that $\phi(\cdot;\bar{\boldsymbol{\theta}})$ takes arguments in $\mathbb{R}^p$ and satisfies the assumptions of Proposition 5.3. Then the proposition implies that the level sets of $\phi(\cdot;\bar{\boldsymbol{\theta}})$ of level $\phi(\bar{\boldsymbol{\theta}};\bar{\boldsymbol{\theta}})$ or less are strictly contained in G.

Theorem 5.4 (Capture property of MM and generalized capture basins) *Suppose that $\{\boldsymbol{\theta}^i\}$ is an MM sequence generated by* (2.4) *and* (2.5). *In addition, suppose that some iterate $\boldsymbol{\theta}^n$ is well-contained in a generalized capture 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 5.3 and an obvious induction argument. □

Remark 5.5 Note that Proposition 5.3 and Theorem 5.4 are fundamental properties in that they rely on (C1), but none of the regularity conditions described in Section 3.2. Also, using Proposition 5.3, one can obtain the same conclusions as in Theorem 5.4 if the sequence $\{\boldsymbol{\theta}^i\}$ merely satisfied $\phi^i(\boldsymbol{\theta}^{i+1}_{\mathcal{S}^i};\boldsymbol{\theta}^i) \leq \phi^i(\boldsymbol{\theta}^i_{\mathcal{S}^i};\boldsymbol{\theta}^i)$ rather than (2.4). This is relevant to practical situations, since one often does not obtain the exact minimizers in (2.4).

The capture property of MM is linked to the global information implicit in tangent majorants in general, and in connected tangent majorants in particular. The algorithm designer uses insights into the global shape of Φ to derive a function satisfying the defining property (2.2). Still more global information is needed to ensure that the tangent majorant is connected. This collective information allows the algorithm descent mechanism to respect the boundaries of a generalized capture basin, even though the location of these boundaries may not be known explicitly to the algorithm designer. Textbook-variety algorithms not assisted by such global information clearly will not imitate the capture property reliably. Such algorithms include derivative-based feasible direction methods (e.g., steepest descent or Newton's method), possibly combined with *ad hoc* constant step-size choices or numerical line-search operations.

Algorithms using constant step-sizes will clearly escape a generalized capture basin if the step-size is chosen too large in comparison to the size of this region. Avoiding such large choices of step-sizes therefore requires foreknowledge of the size of the surrounding generalized capture basin, a degree of global information no less than that inherent in MM. Proposition 1.2.5 in [5, p. 52] describes a capture property for gradient methods with constant step-sizes. However, the region of capture in that Proposition is smaller than the set of well-contained points and becomes smaller as the step-size is increased.

Common numerical line search methods (bisection, Armijo, ...) can likewise let the algorithm escape a generalized capture basin. This is because many points on the search line can satisfy the termination criteria of the line

search method and not all of these points are guaranteed to lie within the smallest surrounding generalized capture basin. Bisection, for example, can find any 1D stationary point on the search line and, for non-convex Φ, many such points may exist, some lying within the local generalized capture basin and some without. To ensure capture, one would need to restrict the search operations to the line segment intersecting the surrounding generalized capture basin. Here again, though, global information would be required to locate the boundaries of this line segment.

Our first application of the capture property is in deriving the following local version of Theorem 4.5.

Theorem 5.6 (Convergence of MM sequences to isolated stationary points (local form)) *In addition to the assumptions of Theorem 5.4, suppose that the conditions of either Theorem 4.1 or Theorem 4.4 are satisfied. Suppose further that G is bounded and either of the following are true*

(a) $\mathrm{cl}\,(G)$ *contains a single stationary point, or*

(b) Condition (C5) *holds and the set of stationary points in* $\mathrm{cl}\,(G)$ *is discrete.*

Then $\{\boldsymbol{\theta}^i\}$ converges to a stationary point $\boldsymbol{\theta}^* \in \mathrm{cl}\,(G)$.

Proof. Since G is bounded, it follows from Theorem 5.4 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 4.4. The conclusions of the Theorem then follow from Lemma 3.6. □

Naturally, an instance of part (a) of primary interest is the case where the region $\mathrm{cl}\,(G)$ contains a single stationary point which is also a global minimizer. For then, the theorem guarantees convergence to a solution of (1.1). Traditionally, local convergence results for minimization algorithms, such as Proposition 1.2.5 in [5, p. 52], assume that the region of convergence is unimodal and basin-like around the local minimizer. Contrary to Theorem 5.6, however, they do not guarantee that convergence will take place from any (well-contained) point within such a region. In a sense, therefore, Theorem 5.6 is the strongest kind of local convergence result, ensuring convergence over the largest possible region that one can expect.

Apart from its role in local convergence, the capture property makes MM an appropriate instrument for implementing the basin-probing steps in various non-convex minimization strategies. Perhaps the most standard strategy is is try to obtain, by heuristics or *ad hoc* methods, an initial point believed to reasonably approximate the desired solution and to hope that this point lies in a unimodal capture basin around the global minimizer. The strategy then tries to descend locally, within the capture basin, to reach the global minimizer. Figure 1 illustrates how the MM capture property facilitates this kind of basin-search. There, the MM sequence results from convex (and hence connected) tangent majorants. Consequently, the sequence is confined to the basin-like region in the interval $\{B,\, C]$ that, fortunately, contains the global minimizer.

If the graph of Φ is clustered with peaks and valleys, a single basin-probing step may not be sufficient. In this case, another standard strategy is to do several basin-searches, as outlined above, but using different initial points. The idea is to search locally around those points and to find the deepest basin. To implement this strategy in a *principled* way, it is highly desirable to do the basin-searches with an algorithm endowed with the capture property. Otherwise, the basin-searches could converge to any stationary point on the graph of Φ and one has no assurance that distinct basins will be probed.

A third example worth mentioning is a path-following method due to [6] called Graduated Non-Convexity (GNC). The GNC strategy employs a sequence of increasingly accurate approximations $\{F(\cdot, t_k)\}_{k=1}^K$ of $\Phi(\cdot)$, begining with a convex function $F(\cdot, t_1)$ that can be easily globally minimized, and ending with $F(\cdot, t_K) = \Phi(\cdot)$. By globally minimizing each $F(\cdot, t_k)$, a sequence $\{\boldsymbol{\theta}^*(t_k)\}_{k=1}^K$ is obtained which, one hopes, converges to the global minimum of Φ. Moreover, each minimization step is initialized with the result of the previous minimization so that

the $\{\boldsymbol{\theta}^*(t_k)\}$, one hopes, are obtained incrementally. In well-behaved circumstances, the initial point of each minimization step will lie in a capture basin containing the solution to the current minimization problem (see Figure 4). Therefore, an algorithm endowed with the capture property is desirable here.

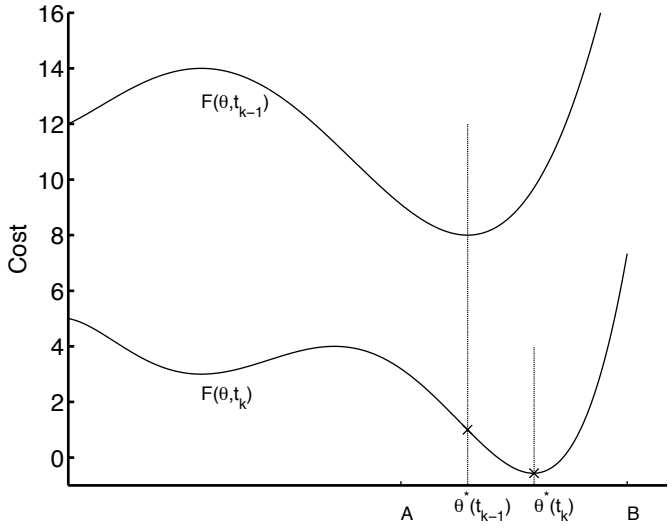

Figure 4: Illustration of the capture property as it pertains to the Graduated Non-Convexity (GNC) strategy.

For certain problems, it may be too expensive to employ strategies involving multiple basin-searches. In such cases, and if Φ is clustered with peaks and valleys, then MM with connected tangent majorants is hazardous, because the likelihood of getting trapped at a sub-optimal solution is high. This limitation is important to recognize, since convex tangent majorants are the most common type used. In such situations, it is worthwhile to consider algorithms not endowed with the capture property. For example, one may try to derive an MM algorithm with non-connected tangent majorants. However, we know of no instance where the absence of the capture property has been systematically exploited. Any success that such algorithms, be they MM or otherwise, have had in avoiding sub-optimal minima seems largely fortuitous.

6 Advancements in MM Applications in PET

The purpose of this section is to illustrate how recent work in statistical medical imaging has evolved beyond the framework of [15]. The work in [15] was motivated by minimization problems encountered in tomographical imaging and ideas for MM algorithms prevailing at that time. More recent work in the imaging field, however, has led to minimization problems of more advanced forms, and correspondingly more advanced ideas for MM algorithms. In the next two subsections, we survey some of these.

In a tomographic imaging system, photons pass through a patient, originating either from an external source or from a radioactive tracer substance injected into the patient's body. Detectors measure the quantity of photons passing through the body in different directions, resulting in a vector of independent Poisson measurements $\boldsymbol{y}$ with components y_i. The mean of the measurement vector is $\bar{\boldsymbol{y}}(\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is an unknown parameter vector containing information about the internal structure of the patient. A commonly considered method of estimating $\boldsymbol{\theta}$ (in particular, it is the focus of [15]) is penalized likelihood estimation, whereby an estimator $\hat{\boldsymbol{\theta}}$ is taken from the minimizers of

$$\Phi(\boldsymbol{\theta}) = L\left(\bar{\boldsymbol{y}}(\boldsymbol{\theta}), \boldsymbol{y}\right) + \rho(\boldsymbol{\theta}). \tag{6.1}$$

Here,

$$L\left(\bar{\boldsymbol{y}}\left(\boldsymbol{\theta}\right),\boldsymbol{y}\right)=\sum_{i=1}^{N}\left[\bar{y}_i\left(\boldsymbol{\theta}\right)-y_i\log\bar{y}_i\left(\boldsymbol{\theta}\right)\right]$$

is the negative of the loglikelihood function, ignoring irrelevant additive constants, while $\rho\left(\boldsymbol{\theta}\right)$ is a penalty function meant to discourage non-physical $\hat{\boldsymbol{\theta}}$.

Often, $\rho\left(\boldsymbol{\theta}\right)$ is convex and $L\left(\bar{\boldsymbol{y}}\left(\cdot\right),\boldsymbol{y}\right)$ is non-convex, as will be the case in the examples considered in this section. A common strategy for deriving MM algorithms is to find majorant generators of the form

$$\phi^i_{\mathrm{PL}}(\boldsymbol{\theta}_{\mathcal{S}^i};\bar{\boldsymbol{\theta}})=\phi^i(\boldsymbol{\theta}_{\mathcal{S}^i};\bar{\boldsymbol{\theta}})+\rho\left(\boldsymbol{\theta}_{\mathcal{S}^i},\bar{\boldsymbol{\theta}}_{\bar{\mathcal{S}}}\right)+c||\boldsymbol{\theta}_{\mathcal{S}^i}-\bar{\boldsymbol{\theta}}_{\mathcal{S}^i}||^2. \tag{6.2}$$

where $\phi^i(\cdot;\bar{\boldsymbol{\theta}})$ is a convex tangent majorant for $L\left(\bar{\boldsymbol{y}}\left(\cdot\right),\boldsymbol{y}\right)$ and $c>0$. In this case, every $\phi^i_{\mathrm{PL}}(\cdot;\bar{\boldsymbol{\theta}})$ is likewise convex, which facilitates minimization in the corresponding MM algorithm and also endows it with the capture property (see Section 5). The term $c||\boldsymbol{\theta}_{\mathcal{S}^i}-\bar{\boldsymbol{\theta}}_{\mathcal{S}^i}||^2$ is meant to ensure strong convexity and thus to help satisfy the conditions of Theorem 4.4. Ideally, one chooses c as small as possible, so that $\phi^i_{\mathrm{PL}}(\cdot;\bar{\boldsymbol{\theta}})$ approximates $\Phi(\cdot,\bar{\boldsymbol{\theta}}_{\bar{\mathcal{S}}^i})$ as accurately as possible.

Thus, the problem of deriving convex tangent majorants $\phi^i_{\mathrm{PL}}(\cdot;\bar{\boldsymbol{\theta}})$ for Φ reduces to deriving a tangent majorant $\phi^i(\cdot;\bar{\boldsymbol{\theta}})$ for $L\left(\bar{\boldsymbol{y}}(\cdot),\boldsymbol{y}\right)$. The following subsections discuss candidates for $\phi^i(\cdot;\bar{\boldsymbol{\theta}})$ for several recent models of $\bar{\boldsymbol{y}}\left(\boldsymbol{\theta}\right)$.

6.1 Joint Estimation of Attenuation and Tracer Concentration in PET

In Positron Emission Tomography (PET), the measurement vector $\boldsymbol{y}=(\boldsymbol{y}^T,\boldsymbol{y}^E)$ is the concatentation of two other measurement vectors $\boldsymbol{y}^T$ and $\boldsymbol{y}^E$. The measurement vector $\boldsymbol{y}^T$ is obtained from a transmission scan using an external photon source, while $\boldsymbol{y}^E$ results from photons emitted from radioactive tracer in the patient. The unknown parameter vector $\boldsymbol{\theta}=(\boldsymbol{\lambda},\boldsymbol{\mu})$ consists of vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$. The components of λ_j, j=1,...,J of $\boldsymbol{\lambda}$ represent the tracer concentration at different locations in the patient's body. The components μ_j, j=1,...,J of $\boldsymbol{\mu}$ represent the attenuation at different locations. Attenuation is a quantity that measures the tendency of matter at that location to absorb or deflect photons. Typically, the components of the statistical mean vector $\bar{\boldsymbol{y}}\left(\boldsymbol{\theta}\right)=\left(\bar{\boldsymbol{y}}^T\left(\boldsymbol{\mu}\right),\bar{\boldsymbol{y}}^E\left(\boldsymbol{\lambda},\boldsymbol{\mu}\right)\right)$ are modeled as follows

$$\bar{y}^T_i\left(\boldsymbol{\mu}\right)=b_i\exp\left(-\sum_{j=1}^{J}A_{ij}\mu_j\right)+r^T_i,\quad i=1,\ldots,N^T \tag{6.3}$$

$$\bar{y}^E_i\left(\boldsymbol{\lambda},\boldsymbol{\mu}\right)=\exp\left(-\sum_{j=1}^{J}A_{ij}\mu_j\right)\sum_{j=1}^{J}P_{ij}\lambda_j+r^E_i,\quad i=1,\ldots,N^E. \tag{6.4}$$

The quantities b_i, A_{ij}, and P_{ij} are known and non-negative. The quantities r^T_i and r^E_i are known and positive.

It is natural to choose ρ of the form

$$\rho\left(\boldsymbol{\lambda},\boldsymbol{\mu}\right)=\beta_{\mathrm{act}}R_{\mathrm{act}}\left(\boldsymbol{\lambda}\right)+\beta_{\mathrm{atten}}R_{\mathrm{atten}}\left(\mu\right), \tag{6.5}$$

where $R_{\mathrm{act}},R_{\mathrm{atten}}$ are measures of spatial roughness in $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ respectively and $\beta_{\mathrm{act}},\beta_{\mathrm{atten}}$ are non-negative weights chosen by the algorithm designer. Common choices of roughness measures are discussed in Section 3.2.1 in [2]. These choices are convex, but most of them are only once differentiable.

Because the total amount of tracer substance Λ_{tot} injected into the patient is known, it is natural to constrain $\boldsymbol{\lambda}$

to the simplex[6]

$$\Theta_{\boldsymbol{\lambda}} = \left\{ \boldsymbol{\lambda} : \sum_j \lambda_j \leq \Lambda_{\text{tot}},\, \lambda_j \geq 0,\, j = 1, \ldots, J \right\}. \tag{6.6}$$

Also, the physical attenuation values of substances in a patient's body are known to lie in some interval $[0,\ \mu_{\max}]$. Therefore, it is appropriate to constrain $\boldsymbol{\mu}$ to the box

$$\Theta_{\boldsymbol{\mu}} = \{ \boldsymbol{\mu} : 0 \leq \mu_j \leq \mu_{\max},\, j = 1, \ldots, J \}.$$

The overall feasible set is then $\Theta = \Theta_{\boldsymbol{\lambda}} \times \Theta_{\boldsymbol{\mu}}$.

In PET, the standard approach is to obtain an estimate $\hat{\boldsymbol{\mu}}$ to $\boldsymbol{\mu}$ based on $\boldsymbol{y}^T$ alone. Then, treating $\hat{\boldsymbol{\mu}}$ as the known value of $\boldsymbol{\mu}$, $\Phi(\boldsymbol{\lambda}, \hat{\boldsymbol{\mu}})$ as given by (6.1) simplifies to a convex function. However, the more statistically principled approach is to minimize $\Phi(\boldsymbol{\lambda}, \boldsymbol{\mu})$ simultaneously as a function of $(\boldsymbol{\lambda}, \boldsymbol{\mu})$. We now discuss the application of MM in the latter case.

A natural way to apply block alternating MM to this problem is to alternately iterate with respect to $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$, i.e., to use index sets $\mathcal{S}^i = i \bmod 2$. This allows us to use existing results to derive convex tangent majorants. The loglikelihood $L\left(\bar{\boldsymbol{y}}\left(\boldsymbol{\lambda}, \bar{\boldsymbol{\mu}}\right), \boldsymbol{y}\right)$ is convex with respect to $\boldsymbol{\lambda}$. Therefore, when iterating with respect to $\boldsymbol{\lambda}$, the convex tangent majorant that best approximates the loglikelihood is the loglikelihood itself,

$$\phi^i(\boldsymbol{\lambda}; \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\mu}}) = L\left(\bar{\boldsymbol{y}}\left(\boldsymbol{\lambda}, \bar{\boldsymbol{\mu}}\right), \boldsymbol{y}\right).$$

When iterating with respect to $\boldsymbol{\mu}$, one can use convex quadratic tangent majorants $\phi^i(\boldsymbol{\mu}; \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\mu}})$ derived in [2].

This minimization problem and the suggested MM algorithm design go beyond the scope of [15] in at least three ways. Firstly, the cost function Φ is non-convex. Secondly, the suggested tangent majorants are not generally twice differentiable when once-differentiable roughess measures are used in (6.5) and then incorporated into (6.2). Thirdly, since Θ_1 is a simplex, the feasible set Θ does not have a Cartesian product decomposition (2.1) with $p = M$.

In [20], we considered a slightly more advanced version of this problem and tested several kinds of MM algorithms, including a block alternating algorithm similar in flavor to that suggested here. This block alternating algorithm effectively reached the minimum cost by iteration $i = 3$. Minimization of each tangent majorant was accomplished iteratively (a large number of sub-iterations was performed) using the conjugate barrier method [4], although any convex programming algorithm could be used.

6.2 Joint Estimation of Tracer Concentration and Anatomical Motion in PET

The physical model considered in Subsection 6.1 assumes that $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ remain constant while the patient is being scanned. However, anatomical motion, e.g., cardiac or respiratory activity, cause these quantities to vary in the course of the scan. Neglecting motion leads to distorted estimates of $\boldsymbol{\lambda}$, particularly of the components λ_j near small features of interest, such as cancerous tumors. Recently, there has been much interest in correcting for motion (e.g., [16, 8, 27]) to reduce such distortion. To allow motion correction in a tomographic scan, one splits the scan into time frames $\tau = 0, 1, \ldots, T-1$ and takes a separate measurement $\boldsymbol{y}^\tau$ in each. Thus, the total measurement vector is $\boldsymbol{y} = (\boldsymbol{y}^0, \ldots, \boldsymbol{y}^{T-1})$.

To model the effect of motion on, say, tracer concentration, a natural approach is to represent the tracer concentrations in time frames $\tau = 1, \ldots, T-1$ as transformations of the tracer concentrations in frame $\tau = 0$. Let $\lambda\left(x, y, z, \tau\right)$ denote the tracer concentration at location (x, y, z) in time frame τ. A common class of transformations

[6] Simplex constraints were also applied to $\boldsymbol{\lambda}$, but with a slightly different motivation, in [3].

is obtained by approximating $\lambda(x,y,z,0)$ according to

$$\lambda(x,y,z,0) = \sum_k h_k(x,y,z)\lambda(x_k,y_k,z_k,0) \tag{6.7}$$

where (x_k,y_k,z_k), k =1,...,J are sample locations and h_k is an interpolation function. For $\tau > 0$, one then applies a nonlinear coordinate transformation to (6.7) of the form,

$$x \mapsto x + d^X(x,y,z|\boldsymbol{\alpha}_X^\tau) \tag{6.8}$$

$$y \mapsto y + d^Y(x,y,z|\boldsymbol{\alpha}_Y^\tau) \tag{6.9}$$

$$z \mapsto z + d^Z(x,y,z|\boldsymbol{\alpha}_Z^\tau). \tag{6.10}$$

Here $\boldsymbol{\alpha}_X^\tau, \boldsymbol{\alpha}_Y^\tau$, and $\boldsymbol{\alpha}_Z^\tau$ are unknown motion parameter vectors and, given $\ell \in \{X,Y,Z\}$ and basis functions $b_k^\ell : \mathbb{R}^3 \to \mathbb{R}$, k =1,...,K_ℓ

$$d^\ell(x,y,z|\boldsymbol{\alpha}_\ell^\tau) \stackrel{\triangle}{=} \sum_{k=1}^{K_\ell} \alpha_{\ell,k} b_k^\ell(x,y,z).$$

Finally, one samples the transformed continuous-space function at (x_j,y_j,z_j) to obtain the motion-altered value of $\lambda(x_j,y_j,z_j,0)$. The net result is that, for a given frame $\tau > 0$, the tracer concentration samples $\lambda(x_j,y_j,z_j,0)$ undergo a linear transformation parametrized by $\boldsymbol{\alpha}^\tau = (\boldsymbol{\alpha}_X^\tau, \boldsymbol{\alpha}_Y^\tau, \boldsymbol{\alpha}_Z^\tau) \in \mathbb{R}^{n_\tau}$,

$$\lambda(x_j,y_j,z_j,\tau) = \sum_k W_{jk}(\boldsymbol{\alpha}^\tau)\,\lambda(x_k,y_k,z_k,0)\,. \tag{6.11}$$

The coefficients $W_{jk}(\boldsymbol{\alpha}^\tau)$ have the form,

$$W_{jk}(\boldsymbol{\alpha}^\tau) = h_k\left(x_j + d^X(x_j,y_j,z_j|\boldsymbol{\alpha}_X^\tau), y_j + d^Y(x_j,y_j,z_j|\boldsymbol{\alpha}_Y^\tau), z_j + d^Z(x_j,y_j,z_j|\boldsymbol{\alpha}_Z^\tau)\right).$$

Thus, if the interpolators h_k are non-negative, then so are the $W_{jk}(\boldsymbol{\alpha}^\tau)$ and the transformation (6.11) preserves non-negativity of the tracer concentrations.

Let $\boldsymbol{\lambda}$ denote the parameter vector with components $\lambda_j = \lambda(x_j,y_j,z_j,0)$. In our recent work on this topic, we have assumed that $\boldsymbol{\mu}$ has been pre-determined and neglect the effect of its motion. This leads to a model for $\bar{\boldsymbol{y}}$ similar to (6.4), but in which the $\lambda_j = \lambda(x_j,y_j,z_j,0)$ are transformed according to (6.11) for $\tau > 0$,

$$\bar{y}_i^0(\boldsymbol{\lambda}) = \sum_{j=1}^J P_{ij}\lambda_j + r_i^0 \tag{6.12}$$

$$\bar{y}_i^\tau(\boldsymbol{\lambda},\boldsymbol{\alpha}^\tau) = \sum_{j=1}^J P_{ij} \sum_k W_{jk}(\boldsymbol{\alpha}^\tau)\,\lambda_k + r_i^\tau, \quad \tau = 1,\ldots,T-1. \tag{6.13}$$

Let $\boldsymbol{\alpha} = (\boldsymbol{\alpha}^1,\ldots,\boldsymbol{\alpha}^{T-1})$ denote the concatenation of all the motion parameter vectors for the different time frames. The above equations define the form of the negative loglikelihood $L(\bar{\boldsymbol{y}}(\cdot),\boldsymbol{y})$. Once the penalty function ρ is defined, the form of the cost function $\Phi(\boldsymbol{\lambda},\boldsymbol{\alpha})$ is fully specified. A reasonable choice for ρ is

$$\rho(\boldsymbol{\lambda},\boldsymbol{\alpha}) = \beta_{\rm act} R_{\rm act}(\boldsymbol{\lambda})$$

to penalize roughness in $\boldsymbol{\lambda}$.

To reflect physical reality, it is desirable that the transformations (6.8)-(6.10) be invertible. In Proposition 1 in [22, p. 60], it was shown that invertibility is ensured by imposing constraints on $\boldsymbol{\alpha}$ of the form,

$$\left|\sum_{k=1}^{K_\ell} \alpha_{\ell,k}^\tau \frac{\partial b_k^\ell(x_j,y_j,z_j)}{\partial l}\right| \le a_\ell \tag{6.14}$$

for all j, all $\ell \in \{X, Y, Z\}$, all $l \in \{x, y, z\}$ and all time frames τ. In addition, the transformed coordinates (6.8) and (6.9) need to stay in the scanner's cylindrical field of view (FOV). If the coordinate $(0, 0, 0)$ is the center of the FOV and $R_{\rm FOV}$ is its radius, this leads to constraints

$$\left(x_j + d^X(x_j, y_j, z_j|\boldsymbol{\alpha}_X^\tau)\right)^2 + \left(y_j + d^Y(x_j, y_j, z_j|\boldsymbol{\alpha}_Y^\tau)\right)^2 \leq R_{\rm FOV}^2 \qquad \forall 1 \leq j \leq J. \tag{6.15}$$

In light of (6.14), each $\boldsymbol{\alpha}_Z^\tau$ is therefore constrained to some polyhedral set Θ_Z^τ. Additionally, each $(\boldsymbol{\alpha}_X^\tau, \boldsymbol{\alpha}_Y^\tau)$ is constrained, in light of (6.14) and (6.15), to a convex non-polyhedral region Θ_{XY}^τ. The overall feasible set therefore has the Cartesian product form

$$\Theta = \Theta_{\boldsymbol{\lambda}} \times \underbrace{\left[\times_{\tau=1}^T \left(\Theta_{XY}^\tau \times \Theta_Z^\tau\right)\right]}_{\Theta_{\boldsymbol{\alpha}}}. \tag{6.16}$$

The set $\Theta_{\boldsymbol{\lambda}}$, as before, is given by (6.6). Similar to Section 6.1, $L\left(\bar{\boldsymbol{y}}\left(\boldsymbol{\lambda}, \boldsymbol{\alpha}\right), \boldsymbol{y}\right)$ is convex as a function of $\boldsymbol{\lambda}$, but non-convex as a function of the other variables. Together with (6.16), this motivates a block alternating MM algorithm design in which $\boldsymbol{\lambda}$ is updated with tangent majorants $\phi(\boldsymbol{\lambda}; \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}}) = L\left(\bar{\boldsymbol{y}}\left(\boldsymbol{\lambda}, \bar{\boldsymbol{\alpha}}\right), \boldsymbol{y}\right)$.

In [21], we proposed techniques to derive convex tangent majorants with respect to $\boldsymbol{\alpha}$. One technique recognized that it is relatively easy to find convex quadratic true tangent majorants $Q_{i\tau}^+(\cdot; \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}})$ to each $\bar{y}_i^\tau\left(\bar{\boldsymbol{\lambda}}, \cdot\right)$ and $Q_{i\tau}^-(\cdot; \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}})$ to $-\bar{y}_i^\tau\left(\bar{\boldsymbol{\lambda}}, \cdot\right)$, i.e.,

$$Q_{i\tau}^-(\cdot; \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}}) \underset{\mathbb{R}^{n_\tau}}{\overset{\bar{\boldsymbol{\alpha}}^\tau}{\prec}} \bar{y}_i^\tau\left(\bar{\boldsymbol{\lambda}}, \cdot\right) \underset{\mathbb{R}^{n_\tau}}{\overset{\bar{\boldsymbol{\alpha}}^\tau}{\prec}} Q_{i\tau}^+(\cdot; \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}}).$$

It then follows (see [21, Proposition 3.3]) that,

$$\phi(\boldsymbol{\alpha}; \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}}) \overset{\triangle}{=} \sum_{i,\tau} \left[Q_{i\tau}^+(\boldsymbol{\alpha}^\tau; \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}}) - y_i \log Q_{i\tau}^-(\boldsymbol{\alpha}^\tau; \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}})\right]$$

is convex and satisfies $\phi(\cdot; \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}}) \underset{D(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}})}{\overset{\bar{\boldsymbol{\alpha}}}{\succeq}} L\left(\bar{\boldsymbol{y}}\left(\bar{\boldsymbol{\lambda}}, \cdot\right), \boldsymbol{y}\right)$, where

$$D(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}}) \overset{\triangle}{=} \left\{\boldsymbol{\alpha} \,:\, Q_{i\tau}^-(\boldsymbol{\alpha}^\tau; \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}}) > 0,\, \forall i, \tau\right\}.$$

That is, $\phi(\cdot; \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}})$ is a true tangent majorant for $L\left(\bar{\boldsymbol{y}}\left(\bar{\boldsymbol{\lambda}}, \cdot\right), \boldsymbol{y}\right)$ on $D(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}})$.

This minimization problem and the suggested MM algorithm design go beyond the scope of [15] in ways similar to Section 6.1. The cost function is again non-convex and we continue to encounter more advanced constraints. In this case, some of the constraints like (6.15) are not even linear. In addition, we have an example of tangent majorants whose domains $D(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}})$ are strict subsets of Θ. However, one can show that any MM algorithm using these tangent majorants nevertheless satisfies the regularity conditions (R1)-(R3), as well as (C3). If $\boldsymbol{\lambda}$ and the $\boldsymbol{\alpha}^\tau$ are alternatingly updated in a way that satisfies (C4), then Theorem 4.4 will ensure asymptotic stationarity.

7 Toward Conditions for Convergence to Non-Isolated Stationary Points

In previous sections, we established convergence of MM sequences under the assumption that the stationary points of (1.1) are isolated. However, other investigators have found that certain MM algorithms are convergent even when a continuum of stationary points is present. This has been observed in certain applications of the EM algorithm (e.g., [7, 33]) and also coordinate descent [28], both of which, as noted earlier, are special cases of MM algorithms. It is natural to wonder, therefore, whether there are general conditions ensuring convergence when the stationary points are possibly not isolated. In this section, we discuss, and examine by way of examples, what conditions might be required.

It seems clear that (R1.1) must be hypothesized *a priori*. This is because an MM sequence can easily become unbounded if it is initialized in a region where the graph of Φ has a pathological shape. Taking a 1D example, suppose $\Phi(\theta) = \theta \exp(-\theta)$, $\Theta = \{\theta \in \mathbb{R} : \theta \geq 0\}$, and connected tangent majorants are used that satisfy the conditions of Theorem 4.1. Initializing with $\theta^0 > 1$ can only produce an unbounded sequence $\{\theta^i\}$. This must be the case since $G = (1,\ \infty)$ is a generalized capture basin where Φ has no stationary points. Thus, by Theorem 4.1 and Theorem 5.4, $\{\theta^i\}$ is confined to $(1,\ \infty)$ but has no limit points there. (Conversely, if $\theta^0 \in [0,\ 1)$, then $\{\theta^i\}$ must converge to the stationary point at $\theta = 0$.)

Another hypothesis that seems necessary for convergence is that the tangent majorants $\phi^i(\cdot; \boldsymbol{\theta}^i)$ have a uniform positive lower bound on their curvatures. Otherwise, they can become asymptotically flat in places, which makes oscillations possible. The next example demonstrates this.

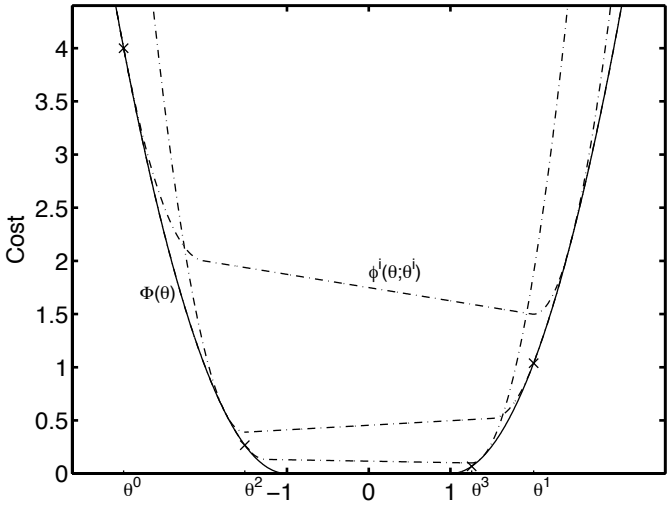

Figure 5: Illustration of Example 7.1. The sequence $\{\theta^i\}$ oscillates between points in the intervals $(-\infty,\ -1)$ and $(1,\ \infty)$, and therefore does not converge. The oscillations are possibly because the tangent majorants become asymptotically flat in the interval $[-1,\ 1]$.

Example 7.1 Suppose $p = 1$ and consider the cost function

$$\Phi(\theta) = \begin{cases} 0.25(|\theta| - 1)^2, & |\theta| \geq 1 \\ 0, & |\theta| < 1. \end{cases}$$

This cost function is piece-wise quadratic. In this example, we present an MM sequence $\{\theta^i\}$, obtained from tangent majorants that are also piece-wise quadratic, which, as in Figure 5, oscillates between the intervals $(-\infty,\ -1)$ and $(1,\ \infty)$. To construct the relevant tangent majorants, we first define, for any $d > 0$,

$$\begin{aligned}
\gamma_d &= 3 + \left(4 - \sqrt{\frac{17}{4}\, d^2 + 16\, d + 16}\right) d^{-1} \\
B_d &= 2d(\gamma_d - 1) \\
A_d &= 1 + 2(B_d - d)^2/d^2 \\
g_1(t, d) &= (t + 1)^2 \\
g_2(t, d) &= 2(t + d + 1)^2 - 2d(t + d + 1) + d^2 \\
g_3(t, d) &= (0.5d^2\gamma_d^2 - d^2\gamma_d + d^2) + B_d(t + d + 1 - 0.5d\gamma_d) \\
g_4(t, d) &= A_d(t - 1 - 0.5d)^2 + B_d(t - 1 - 0.5d) + 0.375d^2,
\end{aligned}$$

and

$$g(t;d) = \begin{cases} g_1(t,d) & t \leq -d-1 \\ g_2(t,d) & -d-1 < t \leq -d-1+0.5d\gamma_d \\ g_3(t,d) & -d-1+0.5d\gamma_d < t \leq 1+0.5d \\ g_4(t,d) & t > 1+0.5d. \end{cases}$$

One can verify the following facts about the foregoing definitions. Firstly, $0 \leq \gamma_d \leq 1$. Secondly, $g(\cdot;d)$ is continuously differentiable with a non-decreasing derivative (and hence is convex). Thirdly, $g(\cdot;d) \overset{d}{\underset{\mathbb{R}}{\succ}} \Phi(\cdot)$. Fourthly, the unique minimizer of $g(\cdot;d)$ is

$$t_d^{\min} = 1 + 0.5d - \frac{B_d}{2A_d}. \tag{7.1}$$

Define the following majorant generator

$$\phi(\theta;\bar{\theta}) = \begin{cases} \Phi(\theta), & |\bar{\theta}| \leq 1 \\ g(-\operatorname{sign}(\bar{\theta})\theta; |\bar{\theta}|-1), & |\bar{\theta}| > 1. \end{cases}$$

One can readily verify, from the aforementioned properties of f, that $\phi(\cdot;\cdot)$ satisfies the requisite properties of a majorant generator.

Let $\{\theta^i\}$ be the MM sequence produced by

$$\begin{aligned} \theta^0 &= -3 \\ \theta^{i+1} &= \operatorname*{argmin}_{t \in \mathbb{R}} \phi(t;\theta^i) \end{aligned}$$

The first few iterations of this sequence are shown in Figure 5. By considering (7.1) and the definition of ϕ, one can verify that $\theta^{i+1} > 1$ if $\theta^i < -1$ and vice-versa for all i. Thus, the sequence $\{\theta^i\}$ oscillates between the intervals $(-\infty,\ -1)$ and $(1,\ \infty)$ as claimed, and so cannot converge. This oscillatory behavior is possible precisely because the tangent majorants become progressively flatter in the interval $[-1,\ 1]$.

Because the algorithm is monotonic and Φ has compact level sets, (R1.1) holds, so $\{\theta^i\}$ has a bounded set of limit points. It is natural to wonder whether these limit points are stationary. If non-stationary limit points exist, it means that $\{\Phi(\theta^i)\}$ cannot converge to the global minimum value of Φ. In fact, though, one can easily verify that the assumptions of Theorem 4.1 are satisfied. In the present example, both condition sets (C6) and {(C2), (C3)} are met. Therefore, the limit points of $\{\theta^i\}$ are indeed stationary and consist of the two global minimizers $\{-1,\ 1\}$.

From the above discussion and example, boundedness of $\{\boldsymbol{\theta}^i\}$ and a lower bound on the tangent majorant curvatures seem to be minimum requirements for a convergence theory. For one-dimensional problems, i.e., when $p = 1$, we can show that these assumptions are also sufficient. This is formalized in the following theorem.

Theorem 7.2 *Suppose that $p = 1$ and that $\{\theta^i\}$ is an MM sequence generated by* (2.10). *Suppose, further, that* (R1.1) *and* (C5.1) *hold. Then $\{\theta^i\}$ converges to a point in Θ.*

Proof. Due to (R1.1), the sequence $\{\theta^i\}$ is bounded and has a compact set of limit points in Θ. Furthermore, since (C5.1) holds, then likewise (C5) holds, so that $\lim_{i\to\infty} ||\theta^{i+1} - \theta^i|| = 0$. A bounded sequence with this property has a connected set of limit points (see [32, p. 173]). In this context, where $\mathbb{R}^p = \mathbb{R}$, we can therefore conclude that the limit points of $\{\theta^i\}$ form a closed, bounded interval $[a,\ b] \subset \Theta$. It remains to show that $a = b$.

Aiming for a contradiction, suppose $a < b$, so that the limit points form an interval of positive length. Then, there exists some k such that $\theta^k \in (a,\ b)$. In addition, due to Lemma 3.5(b), $\Phi(\theta)$ is constant throughout $[a,\ b]$, and

consequently $\frac{d\Phi(\theta^k)}{d\theta} = 0$. By first order minimality conditions, Equation (2.9) can therefore be satisfied only if

$$\frac{d\phi^k(\theta;\theta^k)}{d\theta}\,|_{\theta=\theta^k} = 0. \tag{7.2}$$

Now due to (C5.1), $\phi^k(\cdot;\theta^k)$ is strongly convex. Thus, (7.2) implies that θ^k is the unique global minimizer of $\phi^k(\cdot;\theta^k)$. This, in turn, implies that $\theta^{k+1} = \theta^k$. Repeating this argument inductively, we can conclude that $\theta^i = \theta^k$ for all $i \geq k$. But then $\{\theta^i\}$ converges to θ^k, contradicting the assumption that the limit points constitute an interval of positive length. □

Corollary 7.3 *Suppose that $p = 1$ and that $\{\theta^i\}$ is an MM sequence generated by* (2.10). *Furthermore, suppose* (C5.1) *and the assumptions of Theorem 4.1 hold. Then $\{\theta^i\}$ converges to a stationary point in Θ.*

Proof. The assumptions of Theorem 7.2 hold, so $\{\theta^i\}$ converges to a feasible point. Since the assumptions of Theorem 4.1 holds as well, this limit is stationary. □

Remark 7.4 The assumptions of Theorem 7.2 but not Theorem 4.1 are satisfied in Example 4.2. The sequence $\{\theta^i\}$ therefore converges, but the limit is non-stationary.

It is common to obtain intuition about the behavior of multi-variable problems from single variable ones. Theorem 7.2, however, is one case where this can be misleading. As the next example shows, convergence may not occur if $p > 1$, even if the other assumptions of Theorem 7.2 are met.

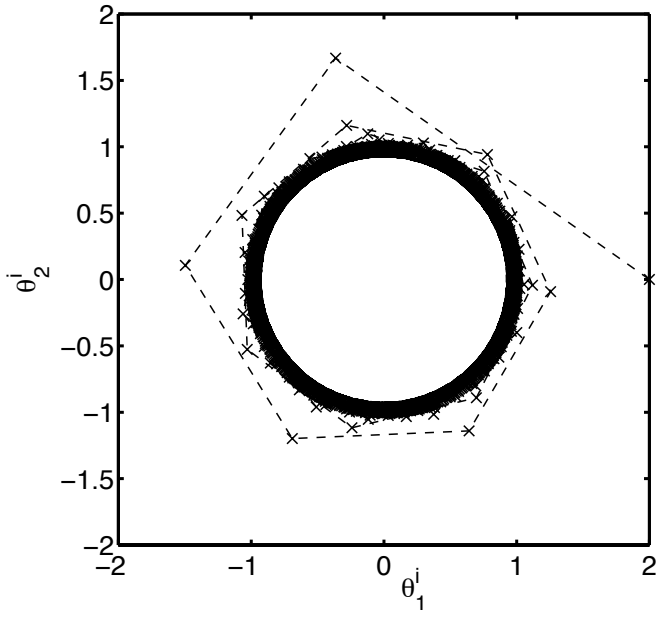

Figure 6: Illustration of Example 7.5. The sequence $\{\boldsymbol{\theta}^i\}$ spirals into an asymptotic orbit about the unit circle.

Example 7.5 Suppose $\mathbb{R}^p = \mathbb{R}^2$ and consider the quadratic cost function

$$\Phi(\boldsymbol{\theta}) = \frac{1}{2}\left[(\theta_1)^2 + (\theta_2)^2\right]$$

In this example, we present a non-convergent MM sequence $\{\boldsymbol{\theta}^i\}$, produced by tangent majorants that are also quadratic, which starts at the point $\boldsymbol{\theta}^0 = [2\ 0]^T$ and spirals in such a way that its limit points are the entire unit circle (see Figure 6).

We begin with a series of definitions:

$$
\begin{aligned}
r_i &= 1 + \sin\left(\frac{\pi}{2(i+1)}\right) \\
\alpha_i &= \begin{cases} \sum_{j=1}^{i} \frac{\pi}{4j}, & i \geq 1 \\ 0, & i = 0 \end{cases} \\
\boldsymbol{\xi}^i &= [r_i \cos\alpha_i,\ r_i \sin\alpha_i] \\
\lambda_i &= \frac{r_i \cos\left(\frac{\pi}{4(i+1)}\right)}{r_i \cos\left(\frac{\pi}{4(i+1)}\right) - r_{i+1}} \\
R_i &= \begin{bmatrix} \cos\alpha_{i+1} & -\sin\alpha_{i+1} \\ \sin\alpha_{i+1} & \cos\alpha_{i+1} \end{bmatrix} \\
\phi^i(\boldsymbol{\theta}; \bar{\boldsymbol{\theta}}) &= (\boldsymbol{\theta} - \boldsymbol{\xi}^{i+1}) R_i^T \begin{bmatrix} \lambda_i & 0 \\ 0 & 1 \end{bmatrix} R_i (\boldsymbol{\theta} - \boldsymbol{\xi}^{i+1})^T
\end{aligned}
$$

We now argue that $\{\boldsymbol{\theta}^i\} = \{\boldsymbol{\xi}^i\}$ is an MM sequence produced using $\{\phi^i(\cdot;\cdot)\}$. To do so, we show that the $\{\phi^i(\cdot;\cdot)\}$ satisfy (2.3) and that $\{\boldsymbol{\theta}^i\}$ satisfies (2.10).

One can verify, from the definition of $\phi^i(\cdot;\cdot)$, that (3.4) holds. It is now sufficient to show that $\lambda_i \geq \frac{1}{2}$ for all i. For if this is true, the first implication is that all the Hessians $\nabla^{20}\phi^i(\cdot;\boldsymbol{\theta}^i)$ are non-negative definite throughout $\mathbb{R}^2$. This, in turn, implies that the quadratics $\phi^i(\cdot;\boldsymbol{\theta}^i)$ are convex with minima at $\boldsymbol{\xi}^{i+1}$ so that the recursion (2.10) is satisfied by $\{\boldsymbol{\theta}^i\} = \{\boldsymbol{\xi}^i\}$. In addition, if $\lambda_i \geq \frac{1}{2}$, one can verify that $\nabla^{20}\phi^i - \nabla^2\Phi$ is non-negative definite. Together with (3.4), this means that each $\phi^i(\cdot;\cdot)$ satisfies (2.3).

From the definition of λ_i, one can verify that $\lambda_i \geq \frac{1}{2}$ if

$$
r_i \cos\left(\frac{\pi}{4(i+1)}\right) - r_{i+1} > 0 \tag{7.3}
$$

Using Taylor's theorem with remainder, one can derive the following bounds for arbitrary $0 \leq t \leq 1$,

$$
\begin{aligned}
1 + \sin\left(\frac{\pi t}{2}\right) &\geq 1 + \frac{\pi t}{2} - \frac{\pi^3 t^3}{48} \\
-\left(1 + \sin\left(\frac{\pi t}{2}\right)\right) &\geq -1 - \frac{\pi t}{2} \\
\cos\left(\frac{\pi t}{4}\right) &\geq 1 - \frac{\pi^2 t^2}{32}.
\end{aligned}
$$

Recalling the definition of r_i, the last three inequalities give lower bounds on the three trigonometric terms/factors on the LHS of (7.3). This leads to

$$
r_i \cos\left(\frac{\pi}{4(i+1)}\right) - r_{i+1} > \frac{P(i)}{Q(i)}, \tag{7.4}
$$

where $P(t)$ and $Q(t)$ are polynomials given by,

$$
\begin{aligned}
P(t) &= 1.2624\,t^4 + 3.6106\,t^3 + 2.1272\,t^2 - 1.3287\,t - 0.9085 \\
Q(t) &= (t+1)^5\,(t+2)\,.
\end{aligned}
$$

One can verify numerically that, for all $t \geq 1$, the rational function $\frac{P(t)}{Q(t)}$ has no real roots and is strictly positive. Hence, (7.3) follows from (7.4). We conclude that the $\{\phi^i(\cdot;\cdot)\}$ satisfy the requisite property (2.3) and that $\{\boldsymbol{\theta}^i\} = \{\boldsymbol{\xi}^i\}$ is the corresponding MM sequence.

From the definition of $\boldsymbol{\xi}^i$, one can see that $\{\boldsymbol{\theta}^i\}$ has the spiral, non-convergent trajectory of Figure 6. This is in spite of the fact that the conditions of Theorem 7.2 (apart from $p = 1$) are satisfied. The sequence is indeed bounded, again by direct inspection of the definition of $\boldsymbol{\xi}^i$. Moreover, the fact that $\lambda_i \geq \frac{1}{2}$ for all i implies that the tangent majorant curvatures are uniformly bounded from below in all directions. Hence, (C5.1) holds.

In the example, $\{\boldsymbol{\theta}^i\}$ does not converge even though it is a bounded sequence and the tangent majorant curvatures are bounded from below. These conditions are insufficient to prevent the increments $\{\boldsymbol{\theta}^{i+1} - \boldsymbol{\theta}^i\}$ from becoming asymptotically tangential. Thus, in the limit, the $\{\boldsymbol{\theta}^i\}$ move in a circular path about the origin. This observation suggests that the convergence of $\{\boldsymbol{\theta}^i\}$ might be ensured by preventing this kind of asymptotically tangential behavior. Preventing this behavior seems to require, at minimum, that *both* an upper and lower bound exist on the curvatures of the tangent majorants $\phi^i(\cdot; \boldsymbol{\theta}^i)$. In Example 7.5, asymptotic tangentiality comes about precisely because the $\{\lambda_i\}$, which determine the tangent majorant curvatures, become unbounded. We may explore these ideas further in future work.

8 Conclusion

We have generalized the analysis of MM given in [15] by relaxing the twice differentiability and convexity assumptions on the cost function Φ and tangent majorants $\{\phi^i(\cdot; \boldsymbol{\theta}^i)\}$. The analysis applies to any convex feasible set and allows the tangent majorant domains to be strict subsets of this set. We have also considered a more general version of the block alternation technique. Our analysis examined the asymptotic properties of such algorithms as well as the tendency of an MM algorithm to be captured in basin-like regions in the graph of Φ.

The asymptotic analysis addressed separately the cases where block alternation was used and where block alternation was not used. When block alternation is not used, asymptotic stationarity is assured if the sequence $\{\phi^i(\cdot; \cdot)\}$ consists of majorant generators chosen from a finite set (condition (C2)) and, moreover, that each $\phi^i(\cdot; \cdot)$ satisfy continuity conditions described in (C3). Alternatively, one can require that the tangent majorant curvatures be uniformly upper bounded in the manner described in (C6). The tangent majorants need not be convex. In the block alternating case, we dropped (C2) and added (C4) and (C5). However, these modified assumptions are no stronger than those considered previously in [15]. In these various cases, convergence results followed (see Theorem 4.5) under standard discreteness assumptions on the problem's stationary points.

In addition to the generality of our assumptions, our asymptotic analysis is structured in a way that imparts several additional theoretical insights, as compared to previous literature. Firstly, we found in Theorem 4.1 that, in the non-block alternating case, conditions like (C5.1), or even its weaker version (C5), need not be known to hold *a priori* to establish asymptotic stationarity. Conversely, in [15, Condition 5], a strong convexity condition similar to (C5.1) is incorporated throughout the analysis. Secondly, our analysis shows that asymptotic stationarity can be assured by curvature conditions like (C6), and not merely continuity conditions like (C3). Conversely, previous convergence analyses are based mainly on continuity conditions. An advantage of the curvature conditions, in the non-block alternating case at least, is that they allow more flexible iteration-dependent behavior to be used. Note that, in Theorem 4.1, the continuity condition (C3) is accompanied by (C2). No such restriction is necessary when (C6) holds. Thirdly, our analysis clarifies when asymptotic stationarity can occur – for algorithms using iteration-dependent majorant generators – even when convergence in norm may not. By contrast, the line of proof in [15] establishes convergence first, before investigating whether the limit is stationary.

In Section 5, we proved the capture property of MM algorithms that use connected (e.g., convex) tangent majorants, which is the most common practice. An implication of this property is that global minimizers attract the iterates over essentially the largest possible neighborhood. The capture property also makes MM a useful instrument in non-convex minimization strategies that rely on basin-probing steps. A negative implication of the capture property is that MM might have a higher tendency to get stuck in local minima than other algorithms. To mitigate this,

one must run the algorithm from multiple initial points, and this can be computationally expensive.

Finally, we made a preliminary examination of conditions that ensure the convergence of MM sequences in the presence of non-isolated stationary points. For 1D problems, convergence is assured provided that a uniform lower bound exists on the tangent majorant curvatures. However, for higher dimensional problems, this condition is insufficient, as was shown by way of a 2D example. Conditions that ensure convergence in higher dimensions were conjectured and may be considered further in sequel papers.

Tangent majorants that can be minimized analytically often provide poor approximations to Φ and result in slow MM algorithms. They also become difficult to devise as problems with more complicated constraints are considered. We therefore expect that MM algorithm implementations, in which tangent majorants are iteratively (and hence inexactly) minimized, will become prevalent. For this reason, sufficient decrease principles for MM is an important topic for future work. At minimum, we have shown that, if convex tangent majorants can be derived, then the MM algorithm will be endowed with the capture property even when the tangent majorants are inexactly minimized (see Remark 5.5). The derivation of convex tangent majorants to non-convex cost functions is a highly active and creative area. We have given several examples of this, in the context of positron emission tomography, in Section 6. Derivation techniques are also discussed in [17] for various applications.

A Notes and Extended Discussions

Note A.1 (Tangent majorants vs. true tangent majorants) A sequence $\{\boldsymbol{\theta}^i\}$ satisfies (2.4) and (2.5) even if global constants are added to the $\{\phi^i(\cdot;\boldsymbol{\theta}^i)\}$. Thus, the behavior of MM sequences can be studied irrespective of whether the $\{\phi^i(\cdot;\boldsymbol{\theta}^i)\}$ are true tangent majorants, or merely tangent majorants. The distinction becomes important, however, when deriving tangent majorants by composition of functions. For example, suppose $f:\mathbb{R}\to\mathbb{R}$ and $g:\mathbb{R}^p\to\mathbb{R}$ with f monotone increasing and suppose $q(\cdot;\bar{\boldsymbol{\theta}}) \overset{\bar{\boldsymbol{\theta}}}{\underset{\Theta}{\succeq}} g(\cdot)$. Then the montonicity of f implies

$$Q(\cdot;\bar{\boldsymbol{\theta}}) \stackrel{\triangle}{=} f(q(\cdot;\bar{\boldsymbol{\theta}})) \overset{\bar{\boldsymbol{\theta}}}{\underset{\Theta}{\succeq}} f(g(\cdot,\bar{\boldsymbol{\theta}}_{\bar{\mathcal{S}}})). \tag{A.1}$$

Thus, deriving the tangent majorant $Q(\cdot;\bar{\boldsymbol{\theta}})$ for $f(g(\cdot,\bar{\boldsymbol{\theta}}_{\bar{\mathcal{S}}}))$ is accomplished by finding a true tangent majorant $q(\cdot;\bar{\boldsymbol{\theta}})$ to g. Note that (A.1) is not necessarily true if $q(\cdot;\bar{\boldsymbol{\theta}})$ is a tangent majorant to g, but not a true tangent majorant.

Note A.2 (Implied derivative matching) Often, MM majorant generators are designed so that, for all i and $\bar{\boldsymbol{\theta}}\in\Theta$

$$\eta^i(\bar{\boldsymbol{\theta}};\boldsymbol{\xi}) = \left\langle \nabla_{\mathcal{S}^i}\Phi(\bar{\boldsymbol{\theta}}),\, \boldsymbol{\xi}-\bar{\boldsymbol{\theta}}_{\mathcal{S}^i}\right\rangle \qquad \forall \boldsymbol{\xi}\in\Theta_{\mathcal{S}^i}. \tag{A.2}$$

That is, the feasible directional derivatives of the tangent majorants $\{\phi^i(\cdot;\boldsymbol{\theta}^i)\}$ should match the cost function at any possible expansion point. This is a stronger version of (3.3), which requires that derivatives match only at the iterates $\{\boldsymbol{\theta}^i\}$.

It is interesting to note that Equation (A.2) is an implied condition whenever $\text{aff}\left(D^i(\bar{\boldsymbol{\theta}})\right) = \text{aff}\left(\Theta_{\mathcal{S}^i}\right)$ and $\bar{\boldsymbol{\theta}}_{\mathcal{S}^i}\in \text{ri}\left(D^i(\bar{\boldsymbol{\theta}})\right)$. This follows directly from the alternative definition of a majorant generator (2.3). The definition requires that the difference $\phi^i(\cdot;\boldsymbol{\theta}^i)-\Phi(\cdot)$ be minimized over $D^i(\boldsymbol{\theta}^i)$ at $\boldsymbol{\theta}^i$. By standard necessary optimality conditions for interior points, the derivatives of this difference must vanish in all feasible directions, i.e., all $\boldsymbol{\xi}-\bar{\boldsymbol{\theta}}_{\mathcal{S}^i}$ such that $\boldsymbol{\xi}\in\text{aff}\left(D^i(\bar{\boldsymbol{\theta}})\right) = \text{aff}\left(\Theta_{\mathcal{S}^i}\right)$. Since $\Theta_{\mathcal{S}^i}\subset\text{aff}\left(\Theta_{\mathcal{S}^i}\right)$, the derivatives must vanish in all directions $\boldsymbol{\xi}-\bar{\boldsymbol{\theta}}_{\mathcal{S}^i}$ such that $\boldsymbol{\xi}\in\Theta_{\mathcal{S}^i}$. This is precisely (A.2).

In some literature (e.g., Footnote 6 in [1]), it is claimed incorrectly that (A.2) is implied at *all* points in $D^i(\bar{\theta})$ and not just in its relative interior. However, since the above arguments rely on optimality conditions for interior points, one cannot expect it to remain true at boundary points. This is illustrated in the following simple example.

Example A.1 (Derivatives need not match at the boundaries) Consider the 1D cost function $\Phi(\theta) = \exp(-\theta) + \theta$ on $\Theta = \{\theta \geq 0\}$ and $\phi(\theta; 0) = \theta + 1$ with domain $D(0) = \Theta$ (see also Figure 7). One can verify that $\phi(\cdot; 0)$ satisfies the requisite property (2.2) of a tangent majorant at expansion point $\bar{\theta} = 0$. However, the feasible directional derivatives of Φ and $\phi(\cdot; 0)$ do not match at the expansion point.

Furthermore, minimizing $\phi(\cdot; 0)$ does not produce a new expansion point. Any MM algorithm that uses $\phi(\cdot; 0)$ gets stuck at the non-stationary point $\bar{\theta} = 0$. It is for this reason that derivative matching must be directly enforced by the algorithm designer (e.g., through (R2) or (A.2)) to avoid degenerate behavior.

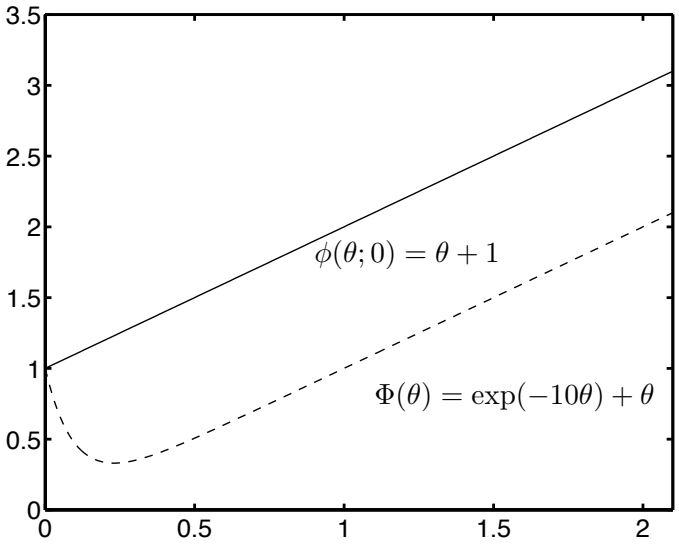

Figure 7: Illustration that, when the expansion point of a tangent majorant is at the boundary of its domain, its derivatives need not match the cost function at that point. It also shows how this can cause a corresponding MM algorithm to prematurely stop at a non-stationary point.

Note A.3 (Imposed gradient matching in non-solid sets) In existing MM literature, the stronger condition (R2.1) is customarily used to ensure (R2). Condition (R2.1) requires that the derivatives of the cost function and tangent majorants match in all directions, not just feasible ones. This might be excessive when $\Theta_{\mathcal{S}^i}$ is not a *solid* subset of $\mathbb{R}_{\mathcal{S}^i}$ (i.e., when $\mathbb{R}_{\mathcal{S}^i} \neq \text{aff}\,(\Theta_{\mathcal{S}^i})$), in which case the space of feasible directions in $\Theta_{\mathcal{S}^i}$ is smaller than $\mathbb{R}_{\mathcal{S}^i}$. A rudimentary example is obtained from Example 2.1 with $\mathcal{S} = 1$. In $\Theta_{\mathcal{S}} = \Theta_1$, the space of feasible directions at any point is a one-dimensional subset of $\mathbb{R}_{\mathcal{S}} = \mathbb{R}^3$. A less rudimentary example is encountered in applications to Positron Emission Tomography (PET) reconstruction. In PET, the sum of the measured data is, with high probability approximately equal to its mean. This motivates hyperplane (i.e., non-solid) constraints in addition to the usual positivity constraints. Hyperplane constraints were also imposed for this problem in [3], but with somewhat different motivation.

When dealing with non-solid $\Theta_{\mathcal{S}^i}$, it may be advantageous to impose (A.2). This condition is weaker than (R2.1) but stronger and perhaps more easy to verify, than (R2).

Note A.4 (EM as a special case of MM) As discussed in Section 1, the family of EM algorithms is a prominent special case of MM algorithms for minimizing the negative loglikelihood $\Phi(\boldsymbol{\theta}) = -\log \mathcal{P}(Y = y|\boldsymbol{\theta})$ of a measurement vector y. One develops an EM algorithm by inventing a joint distribution $\mathcal{P}(Y, Z|\boldsymbol{\theta})$ whose marginal with respect to Y coincides with the given measurement likelihood $\mathcal{P}(Y|\boldsymbol{\theta})$. In the EM literature, the artifical random vector Z is called a *complete data* vector. An MM algorithm is then constructed using the majorant generator,

$$\phi(\boldsymbol{\theta}; \bar{\boldsymbol{\theta}}) = \Phi(\boldsymbol{\theta}) + \text{KL}[\mathcal{P}(Z|Y = y, \bar{\boldsymbol{\theta}}) \,||\, \mathcal{P}(Z|Y = y, \boldsymbol{\theta})], \tag{A.3}$$

where $\mathrm{KL}[f \,||\, g]$ is the Kullback-Leibler (KL) divergence between two probability distributions $f(Z)$ and $g(Z)$,

$$\mathrm{KL}[f \,||\, g] = \int f(z) \log \frac{f(z)}{g(z)} \, \mathrm{d}z. \tag{A.4}$$

Often a joint distribution $\mathcal{P}(Y, Z|\boldsymbol{\theta})$ is constructed by finding a complete data vector Z such that $Y = h(Z)$, where h is a deterministic function. This was the family of constructions considered in [12], but generalizations have been proposed, e.g., [14, 26, 29]. Note that the converse, i.e., constructing complete data so that $Z = h(Y)$ is generally ineffective since (A.3) then reduces to $\phi(\boldsymbol{\theta}; \bar{\boldsymbol{\theta}}) = \Phi(\boldsymbol{\theta})$.

Although it is a classical example of MM, the EM design methodology has frequently resulted in slow algorithms (as discussed, for example, in [13]). This has motivated certain investigators (e.g., [1, 2]) to depart from the EM framework and look for more general types of majorant generators.

The kind of discontinuities exhibited in Example 4.2 can arise in EM majorant generators (A.3) because (see (A.4)), $\mathrm{KL}[f \,||\, g]$ can be discontinuous wherever $f(Z) = g(Z) = 0$. This occurs, for example, in the emission tomography EM application of [33]. When such discontinuities exist, the limit points of $\{\boldsymbol{\theta}^i\}$ may still be stationary. However, very situation-specific analysis may be required to establish this. For the EM algorithm of [33], such an analysis can be found, for example, in [7]. Of course, if the complete data vector is chosen so that $\mathcal{P}(Z|Y = y, \boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta}$, then such discontinuities are not present.

Note A.5 (The dimension of a generalized capture basin) Any generalized capture basin G must have the same dimension as Θ, in the sense that $\mathrm{aff}\,(G) = \mathrm{aff}\,(\Theta)$. Thus, for example, if $\Theta = \mathbb{R}^2$, no line segment inside Θ can constitute a generalized capture basin. We show this by contradiction.

Accordingly, assume instead that $\mathrm{aff}\,(G)$ is a strict subset of $\mathrm{aff}\,(\Theta)$. A consequence of this is that $\Theta \setminus \mathrm{aff}\,(G)$ is non-empty and contains a point that we denote $\boldsymbol{\xi}$. This must be true, since otherwise we would have $\Theta \subset \mathrm{aff}\,(G)$, from which it would follow that $\mathrm{aff}\,(\Theta) \subset \mathrm{aff}\,(G)$, contradicting the assumption that $\mathrm{aff}\,(G)$ is a strict subset of $\mathrm{aff}\,(\Theta)$. Now fix any point $\boldsymbol{\psi} \in G$ and any $\alpha \in (0,\ 1)$. Define

$$\boldsymbol{\theta}_\alpha \stackrel{\triangle}{=} \alpha\boldsymbol{\psi} + (1 - \alpha)\boldsymbol{\xi}. \tag{A.5}$$

Then $\boldsymbol{\theta}_\alpha \in \Theta$, due to the convexity of Θ, but $\boldsymbol{\theta}_\alpha \notin G$. If $\boldsymbol{\theta}_\alpha$ were an element of G, then rearranging (A.5) as follows,

$$\boldsymbol{\xi} = \frac{1}{1 - \alpha}\boldsymbol{\theta}_\alpha - \frac{\alpha}{1 - \alpha}\boldsymbol{\psi},$$

yields $\boldsymbol{\xi}$ as an affine combination of elements of G. However, this is impossible, since $\boldsymbol{\xi}$ was drawn from the set $\Theta \setminus \mathrm{aff}\,(G)$. We conclude that $\boldsymbol{\theta}_\alpha \in \Theta \setminus G$ for any fixed $\alpha \in (0,\ 1)$.

Now, letting α tend to 1 causes $\boldsymbol{\theta}_\alpha$ to tend to $\boldsymbol{\psi}$ and establishes that $\boldsymbol{\psi} \in \mathrm{cl}\,(\Theta \setminus G)$. Also, since $\boldsymbol{\psi}$ was drawn from G, we have that $\boldsymbol{\psi} \in \mathrm{cl}\,(G)$. It follows that $\boldsymbol{\psi} \in \mathrm{cl}\,(G) \cap \mathrm{cl}\,(\Theta \setminus G)$ and, because $\boldsymbol{\psi}$ was arbitrary, that $G = \mathrm{cl}\,(G) \cap \mathrm{cl}\,(\Theta \setminus G)$. However, no generalized capture basin G can satisfy $G = \mathrm{cl}\,(G) \cap \mathrm{cl}\,(\Theta \setminus G)$ because (5.1) would be violated by taking any $\boldsymbol{\theta} \in G$ and letting $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}$. This establishes a contradiction and proves that $\mathrm{aff}\,(G) = \mathrm{aff}\,(\Theta)$.

References

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

[2] 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.

[3] A. Ben-Tal, T. Margalit, and A. Nemirovski. The ordered subsets mirror descent optimization method with applications to tomography. *SIAM J. Optim.*, 12(1):79–108, 2001.

[4] A. Ben-Tal and A. Nemirovski. The conjugate barrier method for non-smooth convex optimization. Technical Report TR #5/99, Minerva Optimization Center, Technion - Israel Institute of Technology, 1999.

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

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

[7] C. Byrne. Likelihood maximization for list-mode emission tomographic image reconstruction. *IEEE Tr. Med. Im.*, 20(10):1084–92, October 2001.

[8] Z. Cao, D. R. Gilland, B. A. Mair, and R. J. Jaszczak. Three-dimensional motion estimation with image reconstruction for gated cardiac ECT. *IEEE Tr. Nuc. Sci.*, 50(3):384–8, June 2003.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] J. A. Fessler, N. H. Clinthorne, and W. L. Rogers. On complete data spaces for PET reconstruction algorithms. *IEEE Tr. Nuc. Sci.*, 40(4):1055–61, August 1993.

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

[15] 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.

[16] D. R. Gilland, B. A. Mair, J. E. Bowsher, and R. J. Jaszczak. Simultaneous reconstruction and motion estimation for gated cardiac ECT. *IEEE Tr. Nuc. Sci.*, 49(5):2344–9, October 2002.

[17] 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.

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

[19] H. M. Hudson and R. S. Larkin. Accelerated image reconstruction using ordered subsets of projection data. *IEEE Tr. Med. Imag.*, 13(4):601–9, December 1994.

[20] M. Jacobson and J. A. Fessler. Simultaneous estimation of attenuation and activity images using optimization transfer. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 4, pages 2085–9, 2001.

[21] 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.*, 2003. To appear. M16-8 http://ieeexplore.ieee.org/xpl/RecentCon.jsp?puNumber=9356.

[22] J. Kim. *Intensity based image registration using robust similarity measure and constrained optimization: applications for radiation therapy*. PhD thesis, Univ. of Michigan, Ann Arbor, MI, 48109-2122, Ann Arbor, MI., 2004.

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

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

[25] 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.

[26] C. H. Liu and Y. N. Wu. Parameter expansion scheme for data augmentation. *J. Am. Stat. Ass.*, 94(448):1264–74, December 1999.

[27] W. Lu and T. R. Mackie. Tomographic motion detection and correction directly in sinogram space. *Phys. Med. Biol.*, 47(8):1267–84, April 2002.

[28] Z. Q. Luo and P. Tseng. On the convergence of the coordinate descent method for convex differentiable minimization. *J. Optim. Theory Appl.*, 72(1):7–35, January 1992.

[29] X. L. Meng and D. van Dyk. The EM algorithm - An old folk song sung to a fast new tune. *J. Royal Stat. Soc. Ser. B*, 59(3):511–67, 1997.

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

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

[32] A. M. Ostrowski. *Solution of equations in Euclidian and Banach spaces*. Academic, New York, 1973.

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

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

[35] 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.

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