# Globally Convergent Image Reconstruction for Emission Tomography Using Relaxed Ordered Subsets Algorithms

Sangtae Ahn*, *Student Member, IEEE*, and Jeffrey A. Fessler, *Senior Member, IEEE*

***Abstract*—We present two types of globally convergent relaxed ordered subsets (OS) algorithms for penalized-likelihood image reconstruction in emission tomography: modified block sequential regularized expectation-maximization (BSREM) and relaxed OS separable paraboloidal surrogates (OS-SPS). The global convergence proof of the existing BSREM (De Pierro and Yamagishi, 2001) required a few *a posteriori* assumptions. By modifying the scaling functions of BSREM, we are able to prove the convergence of the modified BSREM under realistic assumptions. Our modification also makes stepsize selection more convenient. In addition, we introduce relaxation into the OS-SPS algorithm (Erdoğan and Fessler, 1999) that otherwise would converge to a limit cycle. We prove the global convergence of diagonally scaled incremental gradient methods of which the relaxed OS-SPS is a special case; main results of the proofs are from (Nedić and Bertsekas, 2001) and (Correa and Lemaréchal, 1993). Simulation results showed that both new algorithms achieve global convergence yet retain the fast initial convergence speed of conventional unrelaxed ordered subsets algorithms.**

***Index Terms*—Image reconstruction, maximum-likelihood estimation, positron emission tomography, single photon emission computed tomography.**

## I. Introduction

STATISTICAL image reconstruction methods have shown improved image quality over conventional filtered backprojection (FBP) methods (e.g., [5] for maximum-likelihood (ML) reconstruction in emission tomography, and [6] for the analysis of lesion detectability). They use accurate physical models, take the stochastic nature of noise into account, and easily enforce object constraints like nonnegativity. However, iterative algorithms for achieving ML or penalized-likelihood (PL) reconstruction require considerable computation per iteration; so there has been ongoing efforts to develop fast algorithms.

A class of ordered subsets (OS) algorithms, also known as block-iterative or incremental gradient methods, has shown significantly accelerated "convergence." The OS idea is to use only one subset (or block) of the measurement data for each update instead of the total data. Usually, cyclic passing through every subset constitutes one iteration.

The classical "algebraic reconstruction technique" (ART) [7], [8] can be considered to be a type of "OS" method in which each subset consists of a single measurement. However, most ART methods formulate the reconstruction problem as one of finding the solution to a system of equations that involves the imaging physics but not the measurement statistics. Some ART algorithms can be made to converge by introducing relaxation, but the limiting solution has a geometric interpretation in terms of distances to hyperplanes, rather than arising from statistical considerations [9]–[11]. Here, we focus on OS algorithms that are designed to maximize an objective function that captures the statistical properties of the measurements.

The OS principle was applied to the classical expectation-maximization (EM) algorithm [12]–[14] to yield several OS-EM variants. ML reconstruction algorithms include the OS-EM algorithm [15], the rescaled block-iterative EMML (RBI-EMML) algorithm [16], the row-action ML algorithm (RAMLA) [17], and the complete-data OSEM (C-OSEM) [18]. PL reconstruction algorithms include the block sequential regularized EM (BSREM) algorithm [1] (BSREM has RAMLA as a special unregularized case). The paraboloidal surrogates (PS) methods [19], [20] also adopted the OS idea to construct the OS separable paraboloidal surrogates (OS-SPS) [2], originally named the OS transmission (OSTR) algorithm in the context of transmission tomography.

The OS algorithms, including OS-EM, RBI-EMML, and OS-SPS, were successful in speeding up "convergence"; however, they are not *globally convergent*—not even locally convergent—in general. (An algorithm is said to be *globally convergent* if for any starting point the algorithm is guaranteed to generate a sequence of points converging to a solution [21, p. 182].) They usually exhibit limit-cycle like behavior. For each update, OS algorithms use an "approximate gradient" computed from only a part of the data. The approximation is quite reasonable far from a solution if the subset gradients are reasonably "balanced"; the algorithms show acceleration in the early iterations. However, OS algorithms, particularly with a constant stepsize, usually do not converge to a solution, e.g., a stationary point, since the gradient approximation is not exact. Fig. 1 illustrates this typical behavior of OS algorithms.

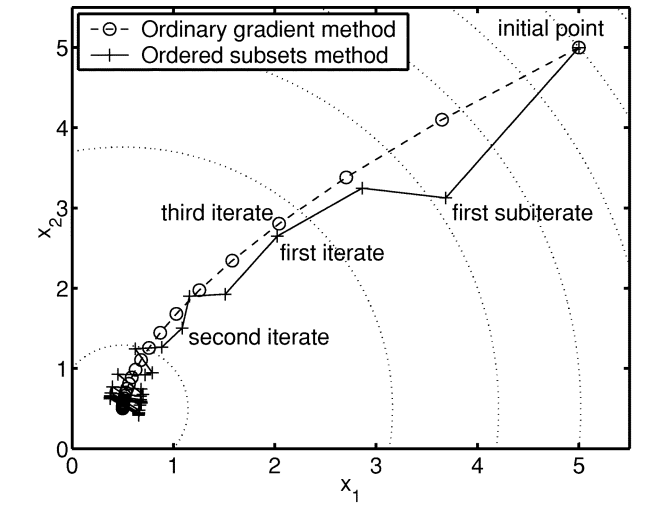
One method for making OS algorithms globally convergent is relaxation, i.e., using diminishing stepsizes. This modification comes from the intuition that the size of a limit cycle should be

Manuscript received May 15, 2002; revised November 26, 2002. This work was supported in part by the National Science Foundation (NSF) under Grant BES-9982349 and by the National Institutes of Health (NIH) under Grant CA-60711. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was R. Leahy. *Asterisk indicates corresponding author.*

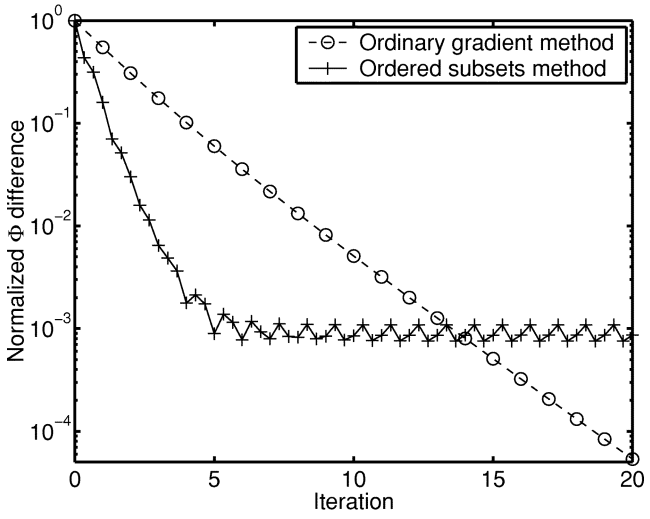
*S. Ahn is with the Electrical Engineering and Computer Science Department, University of Michigan, 4415 Electrical Engineering and Computer Science Building, 1301 Beal Avenue, Ann Arbor, MI 48109-2122 USA (e-mail: sangtaea@umich.edu).

J. A. Fessler is with the Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor, MI 48109-2122 USA.

Digital Object Identifier 10.1109/TMI.2003.812251




(a)




(b)

Fig. 1. Toy example of OS algorithms. (a) Trajectory of iterates of a (non-OS) gradient method with a constant stepsize and its OS version with three subsets. The optimal point is $\hat{x} = (0.5, 0.5)$ and the initial point is $x^0 = (5, 5)$. (b) Normalized $\Phi$ difference $(\Phi(\hat{x}) - \Phi(x^{n,m}))/(\Phi(\hat{x}) - \Phi(x^0))$ versus iteration number. For the OS method, each subiterate is denoted.

proportional to the stepsize. BSREM and RAMLA use diminishing relaxation parameters [1], [17]. De Pierro and Yamagishi [1] provided a global convergence proof for BSREM after imposing a few *a posteriori* assumptions: the convergence of the objective sequence, and the positivity and boundedness of each iterate. In this paper, we relax these assumptions by making some modifications to BSREM.

Kudo, Nakazawa, and Saito [22], [23] also used a relaxation scheme in their block-gradient method applied to penalized weighted least-squares image reconstruction for emission tomography; however, they ignored the nonnegativity constraint. Their method appears to be a special case of incremental gradient methods [3], [24], [25]. Nedić and Bertsekas analyzed the incremental gradient methods and obtained many useful results about their convergence properties [3], [24]. Observing that OS-SPS is a special case of diagonally scaled version of incremental gradient methods with a constant stepsize, in this paper we prove the global convergence of diagonally scaled incremental gradient methods with diminishing stepsizes, thereby establishing global convergence of relaxed OS-SPS.

An alternate method for ensuring convergence would be to run an OS algorithm for several iterations, then switch to a non-OS algorithm known to be globally convergent. In the same spirit, one could decrease the number of subsets over iterations, or continuously decrease parameterized incrementalism as in [27]. The incremental EM [28] can also be considered; this method achieves convergence by applying the incremental (OS) idea block-coordinatewise in an alternating maximization scheme [18], [29].

We focus on relaxed algorithms in this paper. We present two types of relaxed OS algorithms [30]: modified BSREM and relaxed OS-SPS, and we prove the global convergence of the algorithms. Both of them use diagonally scaled gradient ascent for each update to maximize a PL objective function. Although the main difference between these two methods is the form of scaling functions, the approaches of the global convergence proofs are quite different. These algorithms are parallelizable, i.e., able to update all pixels simultaneously and independently, so they are computationally convenient.

In Section II, we formulate the problem for emission tomography. In particular, we establish object constraints as a closed and bounded set instead of the usual unbounded nonnegative orthant. More importantly, we modify the PL objective function without changing the final solution, so that its gradients are Lipschitz continuous on the constraint *including the boundary*. This plays an essential role in subsequent convergence proofs. Section III defines our modified BSREM and relaxed OS-SPS algorithms. Section IV gives simulation results including discussion of relaxation parameters as related to convergence rate.

## II. Emission Tomography Problem

### A. PL Image Reconstruction

We focus on the linear Poisson statistical model that has been used extensively for emission computed tomography, including positron emission tomography (PET) or single photon emission computed tomography (SPECT), as well as for photon-limited optical applications like fluorescence confocal microscopy [31]. Assuming usual Poisson distributions, the measurement model[1] for emission scans is as follows:

$$y_i \sim \text{Poisson}\left\{\sum_{j=1}^{p} a_{ij}\lambda_j^{\text{true}} + r_i\right\}, \quad i = 1, 2, \ldots, N$$

where $y_i \geq 0$ is the number of photons counted in the $i$th bin, $\lambda_j^{\text{true}} \geq 0$ is the activity at the $j$th pixel, $r_i \geq 0$ is the mean number of background events such as scatters, random coincidences and background radiation, and $\boldsymbol{A} = \{a_{ij}\}$ is a system matrix (incorporating scanning time, detector efficiencies, attenuation, scan geometry, etc.) such that $a_{ij} \geq 0$. The goal is to estimate the unknown activity $\boldsymbol{\lambda}^{\text{true}} = [\lambda_1^{\text{true}}, \lambda_2^{\text{true}}, \ldots, \lambda_p^{\text{true}}]'$ based on the measurement $\boldsymbol{y} = [y_1, y_2, \ldots, y_N]'$ with $\boldsymbol{A}$ and $\boldsymbol{r} = [r_1, r_2, \ldots, r_N]'$ being known where $'$ denotes matrix transpose.

[1]For randoms-precorrected PET scans, a shifted Poisson model can be used [32]. An extension to that case is straightforward.

We assume that the sensitivity factors, $\sum_{i=1}^{N} a_{ij}$, are nonzero for all $j$, which is reasonable in practice.

The log-likelihood of $\boldsymbol{\lambda}$ given $\boldsymbol{y}$ can be written, ignoring constants independent of $\boldsymbol{\lambda}$, as follows:

$$L(\boldsymbol{\lambda}) = \sum_{i=1}^{N} h_i\left(l_i(\boldsymbol{\lambda})\right) \tag{1}$$

where $h_i(l) = y_i \log l - l$ and $l_i(\boldsymbol{\lambda}) = \sum_{j=1}^{p} a_{ij}\lambda_j + r_i$. The following properties of $h_i$ can be easily shown[2]

$$\text{(i)} \quad h_i(l) \le h_i(y_i), \quad \forall l \ge 0 \tag{2}$$

$$\text{(ii)} \quad h_i \text{ is monotone increasing on } [0, y_i] \text{ and monotone decreasing on } [y_i, \infty). \tag{3}$$

$$\text{(iii)} \quad h_i \text{ is concave on } [0, \infty). \tag{4}$$

For PL reconstruction, one must find a maximizer of the following objective function over its domain $\mathcal{D}$:

$$\Phi(\boldsymbol{\lambda}) = L(\boldsymbol{\lambda}) - R(\boldsymbol{\lambda}) \tag{5}$$

where

$$\begin{aligned}\mathcal{D} &\triangleq \left\{\boldsymbol{\lambda} \in \mathbb{R}_+^p : \Phi(\boldsymbol{\lambda}) \in \mathbb{R}\right\} \\ &= \left\{\boldsymbol{\lambda} \in \mathbb{R}_+^p : l_i(\boldsymbol{\lambda}) > 0 \text{ or } y_i = 0, \forall i\right\}\end{aligned}$$

with

$$\mathbb{R}_+^p = \{\boldsymbol{\lambda} \in \mathbb{R}^p : \lambda_j \ge 0,\ \forall j\}$$

and $R$ is a regularization term. The reason for taking the domain $\mathcal{D}$ instead of $\mathbb{R}_+^p$ is that the gradient of the log-likelihood is infinite on $\mathbb{R}_+^p \backslash \mathcal{D}$. The use of the feasible domain $\mathcal{D}$ facilitates subsequent analyses. Although the methods described here can be easily generalized, for simplicity, we assume that $R$ is the following type of roughness penalty function:

$$R(\boldsymbol{\lambda}) = \frac{\beta}{2} \sum_{j=1}^{p} \sum_{k \in \mathcal{N}_j} \omega_{jk} \psi(\lambda_j - \lambda_k) \tag{6}$$

where $\beta \ge 0$ is a regularization parameter that controls the smoothness of the reconstructed image, $\mathcal{N}_j$ denotes the neighborhood of the $j$th pixel, $\psi$ is a potential function, and $\omega_{jk} > 0$ is a weighting factor such that $\omega_{jk} = \omega_{kj}$. Viewing the pixels of an image as nodes of a graph with neighboring pixels (say, $\mathcal{N}_j$ for the $j$th pixel) connected by an edge, we assume that the graph is *connected* in the sense that it is always possible to find some sequence of edges leading from any pixel to any other pixel [33]. We assume that $\psi(x)$ is nondecreasing in $|x|$, convex, continuously differentiable, and symmetric, i.e., $\psi(x) = \psi(-x)$, and that $\psi(0) = 0$. Then, $R$ is nonnegative and convex. If $R = 0$, the problem becomes ML reconstruction.

*1) Existence and Uniqueness:* One can verify that the level set $\{\boldsymbol{\lambda} \in \mathcal{D} : \Phi(\boldsymbol{\lambda}) \ge \Phi(\mathbf{1})\}$ is compact (bounded and closed) where $\mathbf{1}$ is a column vector of ones, using the coerciveness[3] of $\Phi$ (i.e., $\lim_{\|\boldsymbol{\lambda}\| \to \infty} \Phi(\boldsymbol{\lambda}) = -\infty$) and the continuity of $\Phi$ on $\mathcal{D}$. Then, by the Weierstrass' Theorem [26, p. 654], there exists a (possibly nonunique) PL solution $\boldsymbol{\lambda}^* \in \mathcal{D}$ such that $\Phi(\boldsymbol{\lambda}^*) = \max_{\boldsymbol{\lambda} \in \mathcal{D}} \Phi(\boldsymbol{\lambda})$.

[2]For convenience, we adopt the convention that $\log 0 = -\infty$ and $0 \cdot \log 0 = 0$.

[3]This can be easily shown by the assumption of nonzero sensitivity factors.

If the objective function $\Phi$ is strictly concave on $\mathcal{D}$, then there exists a unique PL solution [26, p. 685], $\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda} \in \mathcal{D}} \Phi(\boldsymbol{\lambda})$. We assume strict concavity for proving convergence of the modified BSREM algorithms in Section III-A. However, we will allow a concave objective function (possibly having multiple solutions) for the relaxed OS-SPS algorithm, or for more general diagonally scaled incremental gradient methods, in Section III-B. The following Lemma (c.f. [34, Th. 1] and [33, Lemma 1]) provides a simple sufficient condition for the strict concavity of $\Phi$ with a strictly convex and twice differentiable potential function $\psi$. Such potential functions include the quadratic function $\psi(x) = x^2/2$ and many others suggested by Lange [33].

*Lemma 1:* If $\boldsymbol{y}'\boldsymbol{A}\mathbf{1} \ne 0$, then $\Phi$ in (5) [with (6) for $\beta > 0$] is strictly concave on $\mathcal{D}$ for any $\psi$ that is strictly convex and twice differentiable.

*Proof:* The (negative) Hessian of $\Phi$ can be computed as follows:

$$-\nabla^2 \Phi(\boldsymbol{\lambda}) = \boldsymbol{A}'\boldsymbol{W}(\boldsymbol{\lambda})\boldsymbol{A} + \nabla^2 R(\boldsymbol{\lambda})$$

with

$$\boldsymbol{W}(\boldsymbol{\lambda}) = \operatorname{diag}\left\{\frac{y_i}{l_i^2(\boldsymbol{\lambda})}\right\} \tag{7}$$

for $\boldsymbol{\lambda} \in \mathcal{D}$, where we interpret $y_i / l_i^2(\boldsymbol{\lambda})$ as 0 if $y_i = 0$. For any $\boldsymbol{x} \in \mathbb{R}^p$, using the symmetry of $\psi$ and $\omega_{jk}$, we obtain

$$\boldsymbol{x}'\nabla^2 R(\boldsymbol{\lambda})\boldsymbol{x} = \frac{\beta}{2} \sum_{j=1}^{p} \sum_{k \in \mathcal{N}_j} \omega_{jk} \ddot{\psi}(\lambda_j - \lambda_k)(x_j - x_k)^2.$$

Since $\ddot{\psi} > 0$ and the neighborhood system is connected by assumption, for $\beta > 0$, $\boldsymbol{x}'\nabla^2 R(\boldsymbol{\lambda})\boldsymbol{x} = 0$ only if $\boldsymbol{x} = \mathbf{0}$ or $\boldsymbol{x} = c\mathbf{1}$ for some $c \ne 0$. But $c\mathbf{1}'\boldsymbol{A}'\boldsymbol{W}(\boldsymbol{\lambda})\boldsymbol{A}c\mathbf{1} = c^2\|\boldsymbol{W}^{1/2}(\boldsymbol{\lambda})\boldsymbol{A}\mathbf{1}\|^2 \ne 0$ by assumption. So $\boldsymbol{x}'\nabla^2\Phi(\boldsymbol{\lambda})\boldsymbol{x} < 0$, $\forall \boldsymbol{x} \ne \mathbf{0}$. ■

Since $y_i$ and $a_{ij}$ are nonnegative, the assumption $\boldsymbol{y}'\boldsymbol{A}\mathbf{1} \ne 0$ is equivalent to $\boldsymbol{A}'\boldsymbol{y} \ne \mathbf{0}$. In other words, the backprojection of the data must be a nonzero image, which is reasonable in practice.

*2) Boundedness:* It is clear that a PL solution set

$$\Lambda^* \triangleq \left\{\boldsymbol{\lambda}^* \in \mathcal{D} : \Phi(\boldsymbol{\lambda}^*) \ge \Phi(\boldsymbol{\lambda}),\ \forall \boldsymbol{\lambda} \in \mathcal{D}\right\} \tag{8}$$

is bounded by the coerciveness of $\Phi$. In fact, for given data $\boldsymbol{y}$, one *can compute* an upper bound $U = U(\boldsymbol{y}) \in (0, \infty)$ on the elements of $\Lambda^*$ such that

$$\Lambda^* \subset \mathcal{B} \triangleq \{\boldsymbol{\lambda} \in \mathbb{R}^p : 0 \le \lambda_j \le U,\ \forall j\}. \tag{9}$$

See Appendix A for a method of determining $U$. Thus, one can search for a solution over the *bounded* set $\mathcal{B} \cap \mathcal{D}$ instead of over $\mathcal{D}$. This property helps ensure that the (scaled) gradient of the objective function is bounded on a set of interest, which is one of essential ingredients of our global convergence proofs. For example, the gradient of a quadratic penalty with $\psi(x) = x^2/2$ is not bounded on $\mathcal{D}$, whereas it is bounded on $\mathcal{B} \cap \mathcal{D}$.

*3) Differentiability:* The objective function $\Phi$ is not differentiable on the set

$$\mathcal{S} \triangleq \mathbb{R}_+^p \backslash \mathcal{D} = \left\{\boldsymbol{\lambda} \in \mathbb{R}_+^p : l_i(\boldsymbol{\lambda}) = 0 \text{ for some } i \in \mathcal{I}\right\}$$

where

$$\mathcal{I} \triangleq \{i = 1, 2, \ldots, N : r_i = 0 \text{ and } y_i > 0\}. \tag{10}$$

One can see that $\|\nabla\Phi(\boldsymbol{\lambda})\| = \infty$ for $\boldsymbol{\lambda} \in \mathcal{S}$. If a gradient-based algorithm took a point in $\mathcal{S}$, it would collapse. Note that $\mathcal{I} = \emptyset$ and, thus, $\mathcal{S} = \emptyset$ for the case of nonzero backgrounds, $r_i > 0$, $\forall i$. This means that zero backgrounds, $r_i = 0$, can be problematic for some gradient-based algorithms. The EM algorithm for ML reconstruction avoids this problem due to its intrinsic positivity; however, regularization complicates the situation. To circumvent the problem, we slightly modify the log-likelihood, yet without changing the final solution set. We replace the log-likelihood near the problematic region $\mathcal{S}$ with well-behaved functions, e.g., quadratic approximations. We consider the following modified objective function:

$$\tilde{\Phi}(\boldsymbol{\lambda}) = \sum_{i=1}^{N} \tilde{h}_i\left(l_i(\boldsymbol{\lambda})\right) - R(\boldsymbol{\lambda})$$

where

$$\tilde{h}_i(l) \triangleq \begin{cases} \frac{\ddot{h}_i(\epsilon)}{2}(l-\epsilon)^2 + \dot{h}_i(\epsilon)(l-\epsilon) + h_i(\epsilon), \\ \qquad\qquad \text{for} \quad l \le \epsilon \text{ and } i \in \mathcal{I} \\ h_i(l), \qquad \text{otherwise} \end{cases} \tag{11}$$

for some $\epsilon > 0$. The modified marginal log-likelihood $\tilde{h}_i$ is a strictly concave real-valued function defined on $\mathbb{R}$ for $i \in \mathcal{I}$. Note that $\Phi(\boldsymbol{\lambda}) = \tilde{\Phi}(\boldsymbol{\lambda})$ for $\mathcal{E} = \{\boldsymbol{\lambda} \in \mathcal{D} : l_i(\boldsymbol{\lambda}) > \epsilon,\ \forall i \in \mathcal{I}\}$ and that $\tilde{\Phi}$ is well defined on $\mathbb{R}_+^p$. The modified objective function $\tilde{\Phi}$ preserves the (strict) concavity of $\Phi$.[4] Remarkably, one *can compute* $\epsilon > 0$ such that

$$\Lambda^* = \tilde{\Lambda}^* \triangleq \left\{\boldsymbol{\lambda}^* \in \mathcal{B} : \tilde{\Phi}(\boldsymbol{\lambda}^*) \ge \tilde{\Phi}(\boldsymbol{\lambda}),\ \forall \boldsymbol{\lambda} \in \mathcal{B}\right\} \tag{12}$$

meaning that this modified objective function has the same maximizer(s) as the original. See Appendix B for a method of determining $\epsilon$. With such $\epsilon$ in Appendix B, the modified objective function $\tilde{\Phi}$ is real-valued on the compact set $\mathcal{B}$, and it has a nice property that its gradient $\nabla\tilde{\Phi}$ is Lipschitz continuous[5] on $\mathcal{B}$. We will, henceforth, take $\tilde{\Phi}$ as our objective function but revert to the notation $\Phi$ for simplicity; likewise, $h_i$ will denote $\tilde{h}_i$ for $i \in \mathcal{I}$. One should be cautioned that the $\epsilon$ provided by Appendix B could be too small to be practical in finite precision computers; nevertheless, at least we can proceed to develop theory. For the more physically realistic case,[6] where $r_i > 0$, we have $\mathcal{I} = \emptyset$ and we need not modify the objective function.

### B. OS Algorithms

Most iterative algorithms for finding a maximizer of a PL objective function use its gradient $\nabla\Phi$. For objective functions of the form (1), the gradients involve a sum over sinogram indices, i.e., backprojection. Many "parallelizable" algorithms—able to update all the pixels simultaneously—can be written in the following form:[7]

$$\lambda_j^{n+1} = \lambda_j^n + \alpha_n d_j(\boldsymbol{\lambda}^n)\frac{\partial}{\partial\lambda_j}\Phi(\boldsymbol{\lambda}^n), \quad j = 1, 2, \ldots, p \tag{13}$$

where $\alpha_n > 0$ is a relaxation parameter (or stepsize), and $d_j(\boldsymbol{\lambda})$ is a nonnegative scaling function. We call the nonnegative function $d_j(\boldsymbol{\lambda})$ a *scaling function* to emphasize that it *scales* the derivative. Likewise, in vector form

$$\boldsymbol{\lambda}^{n+1} = \boldsymbol{\lambda}^n + \alpha_n \boldsymbol{D}(\boldsymbol{\lambda}^n)\nabla\Phi(\boldsymbol{\lambda}^n) \tag{14}$$

we call the $p \times p$ matrix $\boldsymbol{D}(\boldsymbol{\lambda})$ a *scaling matrix* or simply a *scaling function*. The partial derivative of $\Phi$ is given by

$$\frac{\partial}{\partial\lambda_j}\Phi(\boldsymbol{\lambda}) = \sum_{i=1}^{N} a_{ij}\dot{h}_i\left(l_i(\boldsymbol{\lambda})\right) - \frac{\partial}{\partial\lambda_j}R(\boldsymbol{\lambda}). \tag{15}$$

For example, (13) becomes the ML-EM algorithm if we choose $\alpha_n = 1$ and $d_j(\boldsymbol{\lambda}) = \lambda_j / \sum_{i=1}^{N} a_{ij}$ with $R = 0$.

OS algorithms are obtained by replacing the sum $\sum_{i=1}^{N}$ in (15) with a sum $\sum_{i\in S_m}$ over a subset $S_m$ of $\{1, 2, \ldots, N\}$. Let $\{S_m\}_{m=1}^{M}$ be disjoint subsets of $\{1, 2, \ldots, N\}$ such that $\bigcup_{m=1}^{M} S_m = \{1, 2, \ldots, N\}$, and let

$$f_m(\boldsymbol{\lambda}) \triangleq \sum_{i\in S_m} h_i\left(l_i(\boldsymbol{\lambda})\right) - \gamma_m R(\boldsymbol{\lambda}) \tag{16}$$

be a subobjective function, resulting in

$$\Phi = \sum_m f_m \tag{17}$$

where the regularization term is included in one or more of the $f_m$'s by choosing $\gamma_m \ge 0$ and $\sum_m \gamma_m = 1$. (Typically, we choose $\gamma_m = 1/M$.) Suppose that the following "subset gradient balance" conditions hold:

$$\nabla f_1(\boldsymbol{\lambda}) \cong \nabla f_2(\boldsymbol{\lambda}) \cong \cdots \cong \nabla f_M(\boldsymbol{\lambda}) \tag{18}$$

for $\boldsymbol{\lambda}$ far from the solution set or, equivalently

$$\nabla\Phi(\boldsymbol{\lambda}) \cong M\nabla f_m(\boldsymbol{\lambda}), \quad \forall m. \tag{19}$$

Then, an OS version of (13) is obtained by substituting $M(\partial/\partial\lambda_j)f_m(\boldsymbol{\lambda})$ for $(\partial/\partial\lambda_j)\Phi(\boldsymbol{\lambda})$, as follows[8]:

$$\lambda_j^{n,m} = \lambda_j^{n,m-1} + \alpha_n d_j(\boldsymbol{\lambda}^{n,m-1})\frac{\partial}{\partial\lambda_j}f_m(\boldsymbol{\lambda}^{n,m-1}) \tag{20}$$

for $m = 1, 2, \ldots, M$ where the factor $M$ is absorbed into $d_j$ (or $\alpha_n$), and we use the convention that

$$\lambda_j^{n,0} = \lambda_j^n \text{ and } \lambda_j^{n+1} = \lambda_j^{n,M}.$$

We refer to each update in (20) as the $m$th subiteration of the $n$th iteration. In the tomography context, the partition $\{S_m\}_{m=1}^{M}$

[4]For strict concavity, Lemma 1 still applies to $\bar{\Phi}$. If $\boldsymbol{\lambda} \in \mathbb{R}_+^p \backslash \mathcal{E}$, then its corresponding diagonal element of $\boldsymbol{W}(\boldsymbol{\lambda})$ in (7) would change to $-\ddot{h}_i(\epsilon) = y_i/\epsilon^2$; which leads to the same conclusion.

[5]A function $f$ is called *Lipschitz continuous* on $D$ if there exists some $L > 0$ such that $\|f(\boldsymbol{x}) - f(\boldsymbol{y})\| \le L\|\boldsymbol{x} - \boldsymbol{y}\|$ for all $\boldsymbol{x}, \boldsymbol{y} \in D$. A differentiable function is Lipschitz continuous if its derivatives are bounded. Conversely, the derivatives of a Lipschitz continuous function are bounded when they exist. Therefore, Lipschitz continuity conditions on the gradients of a function imply that the curvatures of the function, if any, are bounded.

[6]Any PET scan will have nonzero randoms and any real SPECT scan will be contaminated by a scattered component and by a nonzero (but possibly quite small) component from background radiation.

[7]Although, for some algorithms, we need to enforce nonnegativity each iteration, we ignore this detail in this section to simplify explanation of OS principles. We do consider this important detail in the convergence proofs, however.

[8]One could use a relaxation sequence $\alpha_{n,m}$ which depends on $m$. In this case, for global convergence, the variations of $\alpha_{n,m}$ over each cycle must be sufficiently small asymptotically (as $n$ goes to $\infty$). For example, see [25]. However, to avoid undue complexity in convergence analysis, we focus on relaxation parameters that are held constant during each iteration, as is widely used [1], [3], [17], [22].

is naturally chosen so that projections within one subset correspond to projections with downsampled projection angles. It is desirable to order the subsets such that projections corresponding to one subset are as "perpendicular" as possible to previously used angles at each subiteration [8]. This strategy has a long history; Hamaker and Solomon [35] analyzed quantitatively the relationship between the convergence rate of ART and ordering in terms of the angles between the null spaces of each projection.

Fig. 1 illustrates the behavior of an OS algorithm for a toy example with the following objective function:

$$\Phi(\boldsymbol{x}) = \sum_{i=1}^{3} \left( -\frac{1}{2}\boldsymbol{x}'\boldsymbol{Q}_i\boldsymbol{x} + \boldsymbol{b}_i'\boldsymbol{x} \right)$$

where $\boldsymbol{Q}_1 = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$, $\boldsymbol{Q}_2 = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$, $\boldsymbol{Q}_3 = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$, $\boldsymbol{b}_1 = \begin{bmatrix} 1.25 \\ 2.5 \end{bmatrix}$, $\boldsymbol{b}_2 = \begin{bmatrix} -1.25 \\ 0.25 \end{bmatrix}$, $\boldsymbol{b}_3 = \begin{bmatrix} 3 \\ -0.75 \end{bmatrix}$, and the maximizer is $\hat{\boldsymbol{x}} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$ We compare an ordinary gradient ascent method

$$\boldsymbol{x}^{n+1} = \boldsymbol{x}^n + \alpha \nabla \Phi(\boldsymbol{x}^n)$$

where $\alpha = .05$, and its OS version with three subsets

$$\boldsymbol{x}^{n,m} = \boldsymbol{x}^{n,m-1} + 3\alpha \nabla f_m(\boldsymbol{x}^{n,m-1}) \text{ for } m = 1, 2, 3$$

where $f_m = -(1/2)\boldsymbol{x}'\boldsymbol{Q}_m\boldsymbol{x} + \boldsymbol{b}_m'\boldsymbol{x}$. As can be seen in the figure, the OS algorithm is about three times faster initially far from the optimal point, but it converges to a limit cycle.

OS algorithms have been successful in speeding up "convergence." However, they generally exhibit limit cycle behavior particularly with a constant stepsize $\alpha_n = \alpha$. Although it is hard to prove the existence of such a limit cycle, one can expect that a set of limit points $\{\boldsymbol{\lambda}^{*,m}\}_{m=1}^{M}$ of a sequence $\{\boldsymbol{\lambda}^{n,m}\}$ generated by (20), if any, would satisfy

$$\lambda_j^{*,m} = \lambda_j^{*,m-1} + \alpha d_j(\boldsymbol{\lambda}^{*,m-1}) \frac{\partial}{\partial \lambda_j} f_m(\boldsymbol{\lambda}^{*,m-1}), \quad \forall m$$
$$\lambda_j^{*,M} = \lambda_j^{*,0}.$$

These conditions generally differ from the true optimality conditions, e.g., $(\partial/\partial\lambda_j)\Phi(\boldsymbol{\lambda}^*) = \sum_{m=1}^{M} (\partial/\partial\lambda_j) f_m(\boldsymbol{\lambda}^*) = 0$ for unconstrained optimization. One may need to use a diminishing stepsize such that $\lim_{n\to\infty} \alpha_n = 0$ to suppress the limit cycle. Even if an algorithm with such relaxation converges to some $\boldsymbol{\lambda}^*$, we must still ensure that the limit $\boldsymbol{\lambda}^*$ is a solution that belongs to $\Lambda^*$. Section III describes appropriate choices of $d_j(\cdot)$ and $\alpha_n$ that ensure global convergence.

## III. Globally Convergent OS Algorithms

The preceding section focused on the properties of the PL objective function $\Phi$ for the specific application of emission tomography. We now turn to the computational problem of maximizing such objective functions. The algorithms described in this section (and the accompanying convergence proofs in the appendices) are applicable to a broad family of objective functions that have the same general properties as the emission tomography case considered in Section II. Specifically, the properties that we exploit are the following: 1) $\Phi$ is concave (or strictly concave) and differentiable; 2) its maximizers lie in a bounded set defined by $0 \le x_j \le U$, where $U$ is a computable upper bound; and 3) $\Phi$ has the summation form (17), where each $f_m$ is concave. In addition, in the convergence proofs we assume that the gradients of the $f_m$ functions are Lipschitz continuous. Collectively, these are fairly unrestrictive assumptions so the algorithms should have broad applicability.

To achieve the goal of maximizing $\Phi$ over $\mathcal{B}$, we present two types of relaxed OS algorithms that are globally convergent: modified BSREM methods and diagonally scaled incremental gradient methods of which relaxed OS-SPS is a special case. For both of these OS algorithms, we use the subobjective functions given in (16). The main difference is in the form of $d_j(\cdot)$ in (13).

### A. Modified BSREM

De Pierro and Yamagishi [1] presented the BSREM algorithm and proved its global convergence under the following assumptions: the sequence $\{\boldsymbol{\lambda}^n\}$ generated by the algorithm is positive and bounded; and the objective sequence $\{\Phi(\boldsymbol{\lambda}^n)\}$ converges. These conditions are not automatically ensured by the form of the original BSREM. We eliminate those assumptions in our convergence analysis by modifying the $d_j(\cdot)$ functions.

The basic idea of the modification is to ensure that all iterates lie in the interior of the constraint set $\mathcal{B}$ by choosing suitable scaling functions $d_j(\cdot)$ and relaxation parameters $\alpha_n$. For EM-like algorithms including BSREM, we observe that using the form $d_j(\boldsymbol{\lambda}) = (\text{some term}) \times \lambda_j$ can help each iterate keep positivity, i.e., avoid crossing the lower boundary $\lambda_j = 0$. We enforce the upper bound $U$ similarly. Consider the following algorithm called modified BSREM-I in vector notation:

$$\boldsymbol{\lambda}^{n,m} = \boldsymbol{\lambda}^{n,m-1} + \alpha_n \boldsymbol{D}(\boldsymbol{\lambda}^{n,m-1}) \nabla f_m(\boldsymbol{\lambda}^{n,m-1}) \qquad (21)$$

for $m = 1, 2, \ldots, M$, where $\alpha_n > 0$ and $\boldsymbol{D}(\boldsymbol{\lambda}) = \text{diag}\{d_j(\boldsymbol{\lambda})\}$ with

$$d_j(\boldsymbol{\lambda}) \triangleq \begin{cases} \frac{\lambda_j}{p_j}, & \text{for } 0 \le \lambda_j < \frac{U}{2} \\ \frac{(U-\lambda_j)}{p_j}, & \text{for } \frac{U}{2} \le \lambda_j \le U \end{cases} \qquad (22)$$

for some $p_j > 0$. (The original BSREM used $d_j(\boldsymbol{\lambda}) = \lambda_j$.)

The convergence analysis of this type of algorithm for a strictly concave objective function is given in Appendix C. The first part (Lemma 2) of the analysis states that if (i) the relaxation sequence is bounded by a sufficiently small value and (ii) the starting point belongs to the interior of $\mathcal{B}$, then the iterates generated by (21) automatically stay in the interior of $\mathcal{B}$. The second part (Lemma 3–5) is about convergence: the iterates generated by (21) converge to the solution $\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}\in\mathcal{B}} \Phi(\boldsymbol{\lambda})$ if (iii) $\sum_{n=0}^{\infty} \alpha_n = \infty$, (iv) $\sum_{n=0}^{\infty} \alpha_n^2 < \infty$, and (v) $\boldsymbol{\lambda}^{n,m} \in \text{Int}\, \mathcal{B}$, $\forall n, m$, where Int $\mathcal{B}$ denotes the interior of $\mathcal{B}$. But the first part says that (v) is guaranteed if (i) and (ii) hold. So, combining two parts, one can conclude (Theorem 1 and Corollary 2) that the modified BSREM-I is globally convergent if (i)–(iv) hold.

A practical and critical issue is how small the relaxation parameter should be in (i) for ensuring (v). If an iterate hits the boundary, then all subsequent iterates remain stuck at the boundary because the scaling function is zero on the boundary. As shown in Lemma 2, one may compute a bound ensuring (v) and use relaxation parameters smaller than the bound.

However, a conservatively small bound will adversely affect convergence rate. So the convergence theorem for BSREM-I still leaves users with practical dilemmas. To overcome these limitations of BSREM-I, we propose to add the following step after (21) for each update:

$$\boldsymbol{\lambda}^{n,m} = \begin{cases} \mathcal{P}_{\mathcal{T}}(\boldsymbol{\lambda}^{n,m}), & \text{for } \boldsymbol{\lambda}^{n,m} \notin \text{ Int } \mathcal{B} \\ \boldsymbol{\lambda}^{n,m}, & \text{otherwise} \end{cases} \tag{23}$$

where $\mathcal{P}_{\mathcal{T}}(\boldsymbol{\lambda})$ is the projection[9] of $\boldsymbol{\lambda} \in \mathbb{R}^p$ onto $\mathcal{T} \triangleq \{\boldsymbol{\lambda} \in \mathbb{R}^p : t \le \lambda_j \le U - t, \forall j\}$ for some small $t > 0$. Consider this modified algorithm (21) with (23), called modified BSREM-II, and suppose that conditions (iii) and (iv) hold. Then, (v) is always satisfied by (23) regardless of whether (i) and/or (ii) hold. Since (iii) implies [36, p. 70] that $\lim_{n\to\infty} \alpha_n = 0$, there exists $N \in \mathbb{N}$ such that $\alpha_n$ satisfies (i) for $n \ge N$. Treating $\boldsymbol{\lambda}^N \in$ Int $\mathcal{B}$ as a "new" starting point, one can see that the iterates after $N$ iterations never hit the boundary by the first part of the analysis mentioned in the previous paragraph. This implies that the step (23) becomes vacuous and in subsequent iterations the modified BSREM-II becomes equivalent to the modified BSREM-I. So by the second part of the analysis the modified BSREM-II is globally convergent. The addition of step (23) removes the conditions (i) and (ii) while retaining global convergence.

In (22), any $p_j > 0$ can be used for global convergence. But we want to choose $p_j$ such that stepsize selection becomes convenient, akin to the appropriateness of a unity stepsize in Newton's methods due to the scaling by the Hessian's inverse. Motivated by the EM algorithm for emission tomography, a reasonable choice for $p_j$ is

$$p_j = \sum_{i=1}^{N} \frac{a_{ij}}{M}. \tag{24}$$

If $M = 1$ (one subset), $\alpha_n = 1$ (unrelaxed), and $R = 0$ (unregularized), then (21) with (24) reduces to ML-EM except the term $U - \lambda_j$ in (22). Although (24) ignores the regularization term, it seems to work well for the regularized case unless the regularization term is too large compared with the log-likelihood part. This is verified experimentally in Section IV.

If we take larger and larger $U$, then $\mathcal{B} \to \mathbb{R}^p_+$ and $d_j(\boldsymbol{\lambda}) \to \lambda_j / p_j$. So the modified BSREM should behave quite similarly to the original BSREM for large $U$ in practice except for our scaling by $p_j$. The upper bound $U$ seems to be more important for convergence analysis than for practical implementation.

### *B. Diagonally Scaled Incremental Gradient Method*

As an alternative to the BSREM methods, we consider next a family of OS algorithms with *constant* scaling functions $d_j(\cdot) = d_j$ as follows:

$$\boldsymbol{\lambda}^{n,m} = \mathcal{P}_{\mathcal{B}}\left(\boldsymbol{\lambda}^{n,m-1} + \alpha_n \boldsymbol{D} \nabla f_m(\boldsymbol{\lambda}^{n,m-1})\right) \tag{25}$$

for $m = 1, 2, \ldots, M$ where $\alpha_n > 0$ and $\boldsymbol{D} = \text{diag}\{d_j\}$ with $d_j > 0$, $\forall j$, and $\mathcal{P}_{\mathcal{B}}(\boldsymbol{\lambda})$ is the projection[10] of $\boldsymbol{\lambda} \in \mathbb{R}^p$ onto $\mathcal{B}$. We call these algorithms diagonally scaled incremental gradient methods since if we choose $\boldsymbol{D} = \boldsymbol{I}$, the algorithm (25) becomes an incremental gradient method [3]. Appendix D presents the convergence analysis of this type of algorithm for a concave objective function (possibly having multiple solutions). The iterates generated by (25) converge to a maximizer if $\sum_{n=0}^{\infty} \alpha_n = \infty$ and $\sum_{n=0}^{\infty} \alpha_n^2 < \infty$ as shown in Theorem 2 and Corollary 3. The global convergence holds regardless of $\boldsymbol{D}$ as long as it is diagonal with positive elements.

A practical issue is how to choose $\boldsymbol{D}$ for fast convergence rate and easy stepsize selection. Fortunately, some hints are given by observing that the OS-SPS method, which showed fairly fast convergence [2], is a special case of (25). In particular, (25) becomes quadratically penalized OS-SPS for a likelihood of the form (1) if $\alpha_n = 1$ and the scaling constants are chosen as follows:

$$d_j = M \left( \sum_{i=1}^{N} a_{ij} a_i w_i + 2\beta \sum_{k \in \mathcal{N}_j} \omega_{jk} \right)^{-1}, \quad \forall j \tag{26}$$

where $a_i \triangleq \sum_{j=1}^{p} a_{ij}$, $M$ is the number of subsets and

$$w_i = \begin{cases} -\ddot{h}_i(y_i), & \text{for } y_i > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Paraboloidal surrogates (PS) methods from which OS-SPS is derived are optimization techniques that, for each iteration, optimize computationally tractable paraboloidal surrogates instead of original objective functions. Those surrogates are characterized by their curvatures; one can optimize such curvatures that ensure monotonicity and fast convergence under certain conditions [20]. In OS version, OS-SPS, an accelerated convergence speed is obtained at the expense of convergence; in this case, the curvatures can be precomputed [2]. The terms in the parenthesis in (26) come from "precomputed curvatures" which are the approximated constant curvatures of separable paraboloidal surrogates [2]. For nonquadratic penalties, the second term in the parenthesis of (26) could be substituted with the curvatures of the penalty function at an initial point or at a uniform image. Although OS-SPS is not globally convergent in general, by allowing a diminishing stepsize, we obtain a relaxed OS-SPS that is readily shown to be globally convergent as a special member of the family (25). Interestingly, whereas the original PS methods [19] for emission tomography required $r_i > 0$ for monotonicity and convergence, we eliminate this requirement here by the modification (11) of the PL.

One of required conditions for the global convergence proofs of diagonally scaled incremental gradient methods is the boundedness of $\nabla f_m$ on $\mathcal{B}$. If the gradient $\nabla R$ of the penalty part is bounded on $\mathbb{R}^p_+$, then we can take $\mathcal{B} = \mathbb{R}^p_+$ while retaining

[9]For a Hilbert space $\mathcal{H}$, a projection $\mathcal{P}_K(\boldsymbol{x})$ of $\boldsymbol{x} \in \mathcal{H}$ onto a nonempty closed convex subset $K \subset \mathcal{H}$ is defined by $\mathcal{P}_K(\boldsymbol{x}) = \arg\min_{\boldsymbol{y} \in K} \|\boldsymbol{x} - \boldsymbol{y}\|$. Here, the projection $\mathcal{P}_{\mathcal{T}}(\boldsymbol{\lambda})$ is easily calculated componentwise as $[\mathcal{P}_{\mathcal{T}}(\boldsymbol{\lambda})]_j := t$ for $\lambda_j < t$, $[\mathcal{P}_{\mathcal{T}}(\boldsymbol{\lambda})]_j := U - t$ for $\lambda_j > U - t$, and $[\mathcal{P}_{\mathcal{T}}(\boldsymbol{\lambda})]_j := \lambda_j$ otherwise. So (23) can be written componentwise as $\lambda_j^{n,m} := t$ for $\lambda_j^{n,m} < 0$, $\lambda_j^{n,m} := U - t$ for $\lambda_j^{n,m} > U$, and $\lambda_j^{n,m} := \lambda_j^{n,m}$ otherwise.

[10]The projection is readily computed componentwise as $[\mathcal{P}_{\mathcal{B}}(\boldsymbol{\lambda})]_j = \text{median}\{0, \lambda_j, U\}$.

global convergence since the gradient of the (modified) log-likelihood is bounded on $\mathbb{R}_+^p$. Such penalties include the Huber penalty

$$\psi(x) = \begin{cases} \frac{x^2}{2}, & \text{for } |x| \le \delta \\ \delta|x| - \frac{\delta^2}{2}, & \text{otherwise} \end{cases}$$

for some $\delta > 0$.

### C. Regularization Into OS Algorithms

There are two typical ways of distributing the regularization term into subobjective functions, i.e., how to choose $\gamma_m$ in (16). One way is to include regularization in every $f_m$ as in [2]

$$\gamma_m = \frac{|S_m|}{N}, \quad \forall m \tag{27}$$

where $|S_m|$ is the number of elements in $S_m$. ($\gamma_m = 1/M$ for equally sized subsets.) Another way is to take the regularization term as a separate subobjective function as in [1]

$$\gamma_m = 0 \text{ for } m = 1, 2, \ldots, M, \text{ and } \gamma_{M+1} = 1 \tag{28}$$

where we have $(M + 1)$ subobjective functions and take $S_{M+1} = \emptyset$. Both cases satisfy the condition $\Phi = \sum_m f_m$. However, the convergence *rates* of the two choices can differ if the regularization parameter $\beta$ is not so small. Recalling the motivations of OS algorithms, (18) and (19), one can expect that (27) will yield faster convergence since (28) may cause poor "subset gradient balance." In other words, the amplitude of a limit cycle that is supposed to be suppressed by relaxation is larger for (28) due to significant dissimilarities between the subobjective functions. On the other hand, (27) requires more computation since the gradient of the regularization part should be computed every subiteration. This additional computational cost is proportional to the number of subsets; however, it is usually relatively small compared with the computation of the log-likelihood part. In experiments not shown, we have observed that the choice (27) usually makes algorithms faster and more stable, so we focus on (27) in Section IV. Nevertheless, our convergence results apply to any choices for the $\gamma_m$'s.

### D. Subiteration-Independent Scaling Matrices are Essential

Both algorithms, (21) and (25), belong to the class (20), where the functions $d_j(\cdot)$ are independent of subiteration index $m$. Classical OS-EM does not belong to this class. As pointed out by Browne and De Pierro [17], OS-EM in general does not converge to a solution even if relaxed. We generalize their argument. One could write a more general form of OS algorithms by allowing different scaling matrices over subiterations

$$\boldsymbol{\lambda}^{n,m} = \boldsymbol{\lambda}^{n,m-1} + \alpha_n \boldsymbol{D}_m(\boldsymbol{\lambda}^{n,m-1}) \nabla f_m(\boldsymbol{\lambda}^{n,m-1}), \ \forall m \tag{29}$$

where $\alpha_n > 0$, $\forall n$ and $\boldsymbol{D}_m(\boldsymbol{\lambda})$ is some nonnegative definite diagonal matrix (function). When we choose $\alpha_n = 1$ and $\boldsymbol{D}_m(\boldsymbol{\lambda}) = \text{diag}\{\lambda_j / \sum_{i \in S_m} a_{ij}\}$, the algorithm (29) becomes OS-EM for $R = 0$. Now consider a relaxed version by assuming $\lim_{n\to\infty} \alpha_n = 0$ and[11] $\sum_{n=0}^{\infty} \alpha_n = \infty$. Following [17], one can write the following expression for $\boldsymbol{\lambda}^{n+1}$:

$$\begin{aligned} \boldsymbol{\lambda}^{n+1} &= \boldsymbol{\lambda}^n + \alpha_n \sum_{m=1}^{M} \boldsymbol{D}_m(\boldsymbol{\lambda}^{n,m-1}) \nabla f_m(\boldsymbol{\lambda}^{n,m-1}) \\ &= \boldsymbol{\lambda}^0 + \sum_{k=0}^{n} \alpha_k \sum_{m=1}^{M} \boldsymbol{D}_m(\boldsymbol{\lambda}^{k,m-1}) \nabla f_m(\boldsymbol{\lambda}^{k,m-1}). \end{aligned}$$

Now suppose that the sequence $\{\boldsymbol{\lambda}^{n,m}\}$ generated by (29) converges to some $\boldsymbol{\lambda}^*$. Assuming that $\boldsymbol{D}_m \nabla f_m$ is continuous, we have

$$\lim_{k\to\infty} \boldsymbol{D}_m(\boldsymbol{\lambda}^{k,m-1}) \nabla f_m(\boldsymbol{\lambda}^{k,m-1}) = \boldsymbol{D}_m(\boldsymbol{\lambda}^*) \nabla f_m(\boldsymbol{\lambda}^*).$$

If $\sum_{m=1}^{M} \boldsymbol{D}_m(\boldsymbol{\lambda}^*) \nabla f_m(\boldsymbol{\lambda}^*) \neq \mathbf{0}$, then $\{\boldsymbol{\lambda}^n\}$ diverges since $\sum_{n=0}^{\infty} \alpha_n = \infty$. So it must be the case that:

$$\sum_{m=1}^{M} \boldsymbol{D}_m(\boldsymbol{\lambda}^*) \nabla f_m(\boldsymbol{\lambda}^*) = \mathbf{0}. \tag{30}$$

However, if the $\boldsymbol{D}_m$'s are different, then (30) is generally different from the true optimality conditions, e.g., $\nabla \Phi(\boldsymbol{\lambda}^*) = \sum_{m=1}^{M} \nabla f_m(\boldsymbol{\lambda}^*) = \mathbf{0}$ for unconstrained optimization. So, in general, OS algorithms with subiteration-dependent scaling matrices, including OS-EM and RBI-EM [16], do not converge *to the desired optimum point* even if they become convergent due to relaxation.

## IV. Results

In this paper, we focused on global convergence analysis. The outline of modified BSREM and relaxed OS-SPS algorithms for a Poisson PL in emission tomography are summarized in Table I and II. In addition to those conditions in Table II, for a general objective function, modified BSREM requires that $\Phi$ is strictly concave, and $\nabla f_m(\boldsymbol{\lambda})$ and $\boldsymbol{D}(\boldsymbol{\lambda}) \nabla f_m(\boldsymbol{\lambda})$ are Lipschitz continuous on $\mathcal{B}$. Diagonally scaled incremental gradient methods including relaxed OS-SPS require that $\nabla f_m$ is bounded on $\mathcal{B}$ and $f_m$ is concave. Local convergence rate analysis will be future work. A critical issue in practice will be how to determine relaxation parameters to get close to a solution within a few iterations. We focus on modified BSREM-II rather than modified BSREM-I in this section. The sufficient conditions on a relaxation sequence for global convergence are the following: $\sum_{n=0}^{\infty} \alpha_n = \infty$ and $\sum_{n=0}^{\infty} \alpha_n^2 < \infty$. One may try to optimize a finite number of relaxation parameters by training [1], [8], [17] if a reasonable training set is given for a particular task. Such relaxation parameters might not seem to satisfy those conditions. However, it may not be relevant since those conditions are *sufficient* and, moreover, *asymptotic*.

One simple choice of relaxation parameters satisfying those conditions is

$$\alpha_n = \frac{\alpha_0}{\gamma n + 1}, \ \forall n \tag{31}$$

[11] If we take a diminishing stepsize ($\lim_{n\to\infty} \alpha = 0$), we need the assumption: $\sum_{n=0}^{\infty} \alpha_n = \infty$. Suppose that $\sum_{n=0}^{\infty} \alpha_n < \infty$. Since $\|\boldsymbol{\lambda}^{n+1} - \boldsymbol{\lambda}^n\| = O(\alpha_n)$ (by assuming that $\boldsymbol{D}_m \nabla f_m$ is bounded), we will never get to the optimum point if an initial point is sufficiently far from it.

TABLE I
ALGORITHM OUTLINE FOR THE ALGORITHMS PRESENTED IN THIS PAPER

Compute a bound $U$ on a solution by (32) in Appendix A.
Compute $\epsilon$ by (35) in Appendix B if $\mathcal{I} \neq \emptyset$, that is, $r_i = 0$ but $y_i > 0$ for some $i$.
Precompute $p_j = \sum_{i=1}^{N} a_{ij}/M$ for modified BSREM,
or precompute $d_j^{\mathrm{P}}$ for relaxed OS-SPS. Use (26) for quadratic penalty.
**for** each iteration $n = 1, \ldots,$niter
  **for** each subset $m = 1, \ldots, M$
    $\hat{l}_i = \sum_{j=1}^{p} a_{ij}\hat{\lambda}_j + r_i$ for $i \in S_m$
    $\dot{h}_i = \begin{cases} \dot{h}_i(\epsilon) + \ddot{h}_i(\epsilon)(\hat{l}_i - \epsilon) & \text{for } i \in \mathcal{I} \text{ and } \hat{l}_i \leq \epsilon, \text{ where } h_i(l) = y_i \log l - l \\ (y_i/\hat{l}_i) - 1 & \text{otherwise} \end{cases}$
    $\boldsymbol{\lambda}^{\mathrm{old}} = \hat{\boldsymbol{\lambda}}$
    **for** $j = 1, \ldots, p$
      $\dot{\Phi}_j = \sum_{i \in S_m} a_{ij}\dot{h}_i - \beta \sum_{k \in \mathcal{N}_j} \omega_{jk}\dot{\psi}(\lambda_j^{\mathrm{old}} - \lambda_k^{\mathrm{old}})/M$
      Update $\hat{\lambda}_j$. (See Table II.)
    **end**
  **end**
**end**

TABLE II
COMPARISON OF ALGORITHMS

| Algorithm | Update in Table I | Sufficient conditions for convergence |
|---|---|---|
| Modified BSREM-I | $d_j^{\mathrm{B}} = \begin{cases} \hat{\lambda}_j/p_j \text{ for } \hat{\lambda}_j < U/2 \\ (U - \hat{\lambda}_j)/p_j \text{ for } \hat{\lambda}_j \geq U/2 \end{cases}$ $\hat{\lambda}_j := \hat{\lambda}_j + \alpha_n d_j^{\mathrm{B}}\dot{\Phi}_j$ | (i) $\sum_n \alpha_n = \infty$ (ii) $\sum_n \alpha_n^2 < \infty$ (iii) $\alpha_n$ is sufficiently small (vi) $\hat{\boldsymbol{\lambda}}^{\mathrm{Initial}} \in \mathrm{Int}\,\mathcal{B}$ or, instead of (iii) and (iv), (v) All iterates lie in the interior of $\mathcal{B}$ |
| Modified BSREM-II | $\hat{\lambda}_j := \hat{\lambda}_j + \alpha_n d_j^{\mathrm{B}}\dot{\Phi}_j$ same as above $\hat{\lambda}_j := \begin{cases} t \text{ if } \hat{\lambda}_j \leq 0 \\ U - t \text{ if } \hat{\lambda}_j \geq U \end{cases}$ | $\sum_n \alpha_n = \infty$ and $\sum_n \alpha_n^2 < \infty$ |
| Relaxed OS-SPS | $\hat{\lambda}_j := \hat{\lambda}_j + \alpha_n d_j^{\mathrm{P}}\dot{\Phi}_j$ $\hat{\lambda}_j := \begin{cases} 0 \text{ if } \hat{\lambda}_j \leq 0 \\ U \text{ if } \hat{\lambda}_j \geq U \end{cases}$ | $\sum_n \alpha_n = \infty$ and $\sum_n \alpha_n^2 < \infty$ |

$t$ is a small value, say, $0.001 \max_j \hat{\lambda}_j^{\mathrm{FBP}}$.

for $\gamma > 0$ and $\alpha_0 > 0$. We run simulations using these simple relaxation parameters. Our goal here is not to try to find the best relaxation but to get some insight into the effects of relaxation parameters on convergence rate through some experiments. By design, our modified BSREM and relaxed OS-SPS are properly scaled, meaning that even a constant $\alpha_n = 1$ works fairly well. So we could obtain reasonably good results by setting $\alpha_0 = 1$ and tuning experimentally only $\gamma$.

We performed image reconstruction using two-dimensional SPECT simulation data generated with the Shepp–Logan digital phantom. The projection space was 128 radial bins with 3.6 mm ray spacing and 120 angles over 360°, and the reconstructed images were 128 × 128 with 3.6 mm pixel size. The distance from the center of rotation to the detector plane was 288 mm. The system matrix $\boldsymbol{A}$ was generated by ASPIRE 3.0 [37] and it assumed a Gaussian shaped point spread function with the following model for the depth-dependent full-width at half-maximum (FWHM):

$$\mathrm{FWHM} = \sqrt{(0.086\,805\,6 \cdot z)^2 + (3\ \mathrm{mm})^2}$$

where $z$ is the distance from a pixel's center to the detector. We did not consider attenuation in this simulation. The total counts were $5{\times}10^5$, and $r_i$ corresponded to a uniform field of 10% of background events, a very crude approximation of the effects of scatter. We regularized the log-likelihood using the first-order quadratic penalty $\psi(x) = x^2/2$ with $\beta = 1.5$, and we took a FBP reconstruction as a starting image for PL reconstruction. Because the relaxed OS algorithms are additive updates, the scaling of the initial image can affect the initial convergence rate, so we implemented the FBP algorithm carefully with respect to the global scale factor. In contrast, the classical ML-EM and OS-EM methods for emission tomography are multiplicative, so the initial scaling is unimportant.

Fig. 2 compares two non-OS algorithms: SPS with optimum curvature [20] and De Pierro's modified EM [38]; and two unrelaxed OS algorithms: unrelaxed OS-SPS and unrelaxed modified BSREM with $\alpha_n = 1$ and with 8 subsets and 40 subsets. The OS algorithms initially increase the objective function much faster than the non-OS ones, but they get stuck at suboptimal points. The figure shows the normalized $\Phi$ difference $(\Phi(\hat{\boldsymbol{\lambda}}) - \Phi(\boldsymbol{\lambda}^n))/(\Phi(\hat{\boldsymbol{\lambda}}) - \Phi(\boldsymbol{\lambda}^0))$ versus iteration number where $\hat{\boldsymbol{\lambda}}$ is the solution estimated by 5000 iterations of De Pierro's modified EM, a globally convergent method [38]. One can see that the scaling factors (22) with (24), and

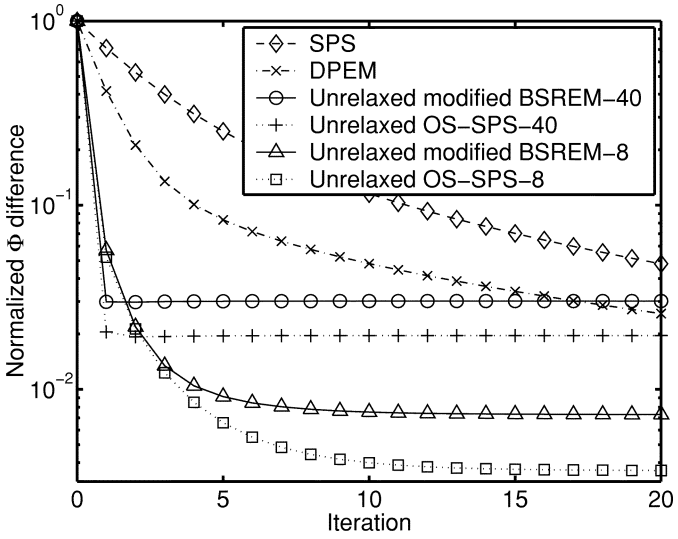



Fig. 2. Comparison of normalized $\Phi$ difference $(\Phi(\hat{\boldsymbol{\lambda}}) - \Phi(\boldsymbol{\lambda}^n))/(\Phi(\hat{\boldsymbol{\lambda}}) - \Phi(\boldsymbol{\lambda}^0))$ versus iteration number for non-OS algorithms including SPS and De Pierro's modified EM denoted by DPEM; and unrelaxed (i.e., constant stepsize) OS algorithms including unrelaxed OS-SPS and unrelaxed modified BSREM (with 8 and 40 subsets).

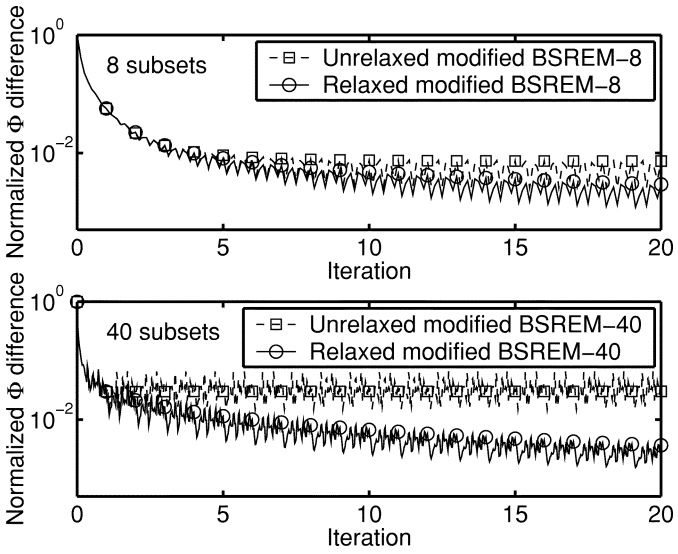



Fig. 3. Comparison of normalized $\Phi$ difference $(\Phi(\hat{\boldsymbol{\lambda}}) - \Phi(\boldsymbol{\lambda}^{n,m}))/(\Phi(\hat{\boldsymbol{\lambda}}) - \Phi(\boldsymbol{\lambda}^0))$ versus iteration number for unrelaxed modified BSREM and relaxed modified BSREM with 8 and 40 subsets. For relaxed modified BSREM-8 (top) and relaxed modified BSREM-40 (bottom), $\alpha_n = 1/((1/15)n + 1)$ and $\alpha_n = 1/(n+1)$ are used, respectively. This figure shows every subiterate.

(26) for the OS algorithms are reasonable since the stepsize of unity worked fairly well. For both unrelaxed OS-SPS and unrelaxed modified BSREM, using more subsets accelerated "convergence" but made the algorithms reach a limit cycle earlier. Roughly speaking, in early iterations more subsets are desirable but in later iterations fewer subsets would be preferable in the unrelaxed case.

Now, we see how relaxation improves convergence. Fig. 3 compares unrelaxed modified BSREM and relaxed modified BSREM. As can be seen in the figure, the unrelaxed modified BSREM algorithms converged to a limit cycle, whereas the relaxed ones showed better performance in increasing the objective function by suppressing the amplitude of the cycle (note

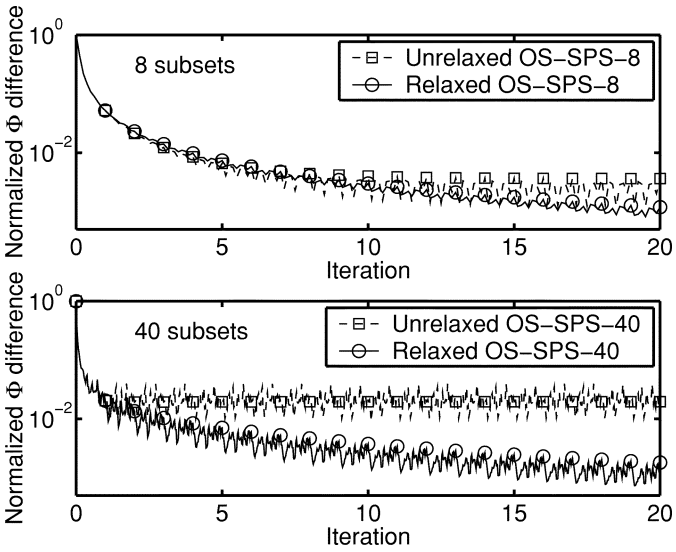



Fig. 4. Comparison of normalized $\Phi$ difference $(\Phi(\hat{\boldsymbol{\lambda}}) - \Phi(\boldsymbol{\lambda}^{n,m}))/(\Phi(\hat{\boldsymbol{\lambda}}) - \Phi(\boldsymbol{\lambda}^0))$ versus iteration number for unrelaxed OS-SPS and relaxed OS-SPS with 8 and 40 subsets. For relaxed OS-SPS-8 (top) and relaxed OS-SPS-40 (bottom), $\alpha_n = 1/((1/5)n + 1)$ and $\alpha_n = 1/(n+1)$ are used, respectively. This figure shows every subiterate.

the logarithmic scale). We chose $\alpha_n = 1/((1/15)n + 1)$ for relaxed modified BSREM-8 and $\alpha_n = 1/(n+1)$ for relaxed modified BSREM-40. In this experiment, the second part of the scaling function in (22) was never invoked due to the very large bound $U$ used; the scaling matrix we used was effectively the same as that of original BSREM except for $p_j$. Fig. 4 shows results for relaxed OS-SPS that are similar to those for modified BSREM. We chose $\alpha_n = 1/((1/5)n+1)$ for relaxed OS-SPS-8, and $\alpha_n = 1/(n+1)$ for relaxed OS-SPS-40. Fig. 5 summarizes Fig. 3 and Fig. 4. We also plotted distance to the solution $\|\boldsymbol{\lambda}^n - \hat{\boldsymbol{\lambda}}\|$ versus iteration number; although not shown in this paper, the plots showed similar results. The reconstructed images are shown in Fig. 6.

We observed, from experiments with relaxation parameters, that applying relaxation (less than unity) before an algorithm reaches a limit cycle far from the optimum point does not improve convergence rate because it slows down the algorithm's progress toward the optimum point. Apparently, relaxation is most helpful when an algorithm is nearing a limit cycle. Generally speaking, rapidly diminishing stepsizes are preferable for an algorithm using many subsets since such algorithms tend to reach a limit cycle quickly. But relaxation should be applied gradually in cases where it takes many iterations for an algorithm to reach a limit cycle, e.g., unregularized ML reconstruction or when few subsets are used.

## V. Conclusion

We presented two types of globally convergent relaxed OS algorithms: modified BSREM and relaxed OS-SPS which differ in their scaling functions $d_j(\cdot)$. We proved global convergence of both algorithms without *a posteriori* assumptions. A natural subsequent question is about convergence rate. This is related to how to determine the relaxation parameters. For relaxation parameters, we showed through experiments that relaxation

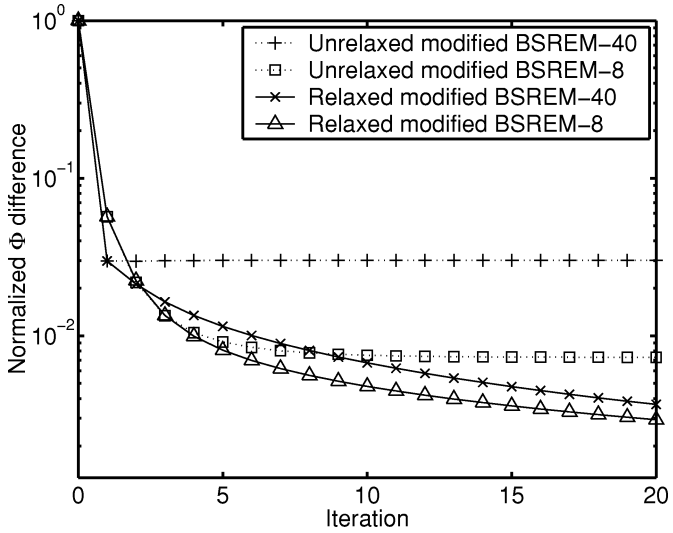



Fig. 5. Comparison of normalized $\Phi$ difference $(\Phi(\hat{\boldsymbol{\lambda}}) - \Phi(\boldsymbol{\lambda}^n))/(\Phi(\hat{\boldsymbol{\lambda}}) - \Phi(\boldsymbol{\lambda}^0))$ versus iteration number for unrelaxed OS algorithms and relaxed ones. (a) Unrelaxed modified BSREM and relaxed modified BSREM. This figure is the same as Fig. 3 except that it shows only each iterate. (b) Unrelaxed OS-SPS and relaxed OS-SPS. This figure is the same as Fig. 4 except that it shows only each iterate.

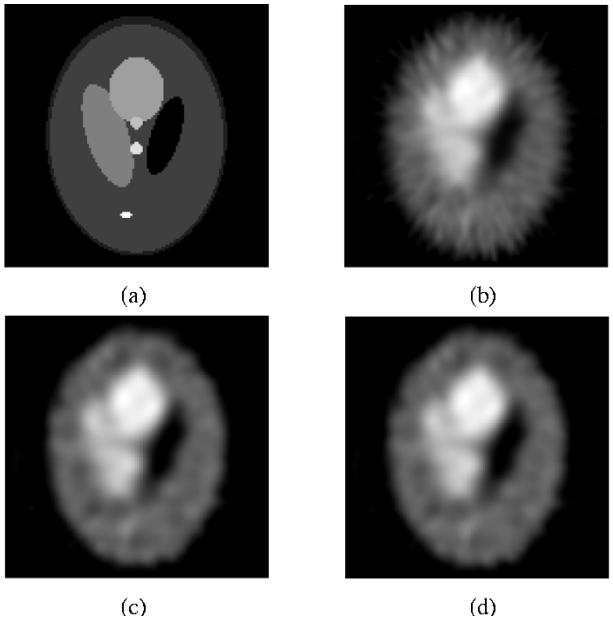



Fig. 6. (a) Shepp-Logan digital phantom (true image). (b) FBP reconstruction (starting image). (c) PL reconstruction using 20 iterations of relaxed modified BSREM with eight subsets. (d) PL reconstruction using 20 iterations of relaxed OS-SPS with eight subsets.

improves the OS algorithms convergence rates when the algorithms are approaching a limit cycle. Hopefully, future work on quantitative convergence rate analysis will provide more useful rules for determining relaxation parameters, perhaps adaptively.

The practical question of whether it is preferable to achieve convergence by using relaxation or by reducing the number of subsets with iteration remains open, and may simply be a matter of preference. When iterative algorithms become implemented in special purpose hardware, the consistent data flow provided by the relaxation approach may be beneficial.

In this paper, we have not tried to evaluate the relative merits of modified BSREM and relaxed OS-SPS. Both algorithms are globally convergent, and simulation results showed that appropriate relaxation accelerates convergence similarly for both of them. Finding better scaling functions in terms of convergence speed and computational efficiency could also be interesting future work.

The algorithms presented in this paper are easily adapted to transmission tomography for zero backgrounds $(r_i = 0)$. However, for a nonzero background case, the PL objective function can become nonconcave [20]. It will also be interesting future work to investigate whether the relaxed OS algorithms can be proved to converge to local maxima in nonconcave cases.

## APPENDIX A

In this Appendix, we construct an upper bound $U$ that makes (9) hold. Define

$$U \triangleq \max_i \left\{ \frac{y_i}{\min\limits_{j:a_{ij}\neq 0} a_{ij}} \right\}. \tag{32}$$

Suppose $\boldsymbol{\lambda}$ is a vector in $\mathcal{D}$ for which the set of "too large" elements $\mathcal{J} = \{j = 1,\ldots,p : \lambda j > U\}$ is nonempty. Define $\tilde{\boldsymbol{\lambda}}$ by clipping as $\tilde{\lambda}_j = \begin{cases} U, & j \in \mathcal{J} \\ \lambda_j, & j \notin \mathcal{J} \end{cases}$. It suffices to show that $\Phi(\boldsymbol{\lambda}) < \Phi(\tilde{\boldsymbol{\lambda}}) \leq \max_{\mathcal{D}} \Phi$. First, note that, for each $i$, if there exists $j_i \in \mathcal{J}$ such that $a_{ij_i} > 0$, then

$$\begin{aligned} l_i(\boldsymbol{\lambda}) &= a_{ij_i}\lambda_{j_i} + \sum_{j\neq j_i} a_{ij}\lambda_j + r_i \\ &> a_{ij_i}U + \sum_{j\neq j_i} a_{ij}\tilde{\lambda}_j + r_i = l_i(\tilde{\boldsymbol{\lambda}}) \qquad (33) \\ &\geq a_{ij_i}U \geq y_i \qquad (34) \end{aligned}$$

where in (33) we used the fact that $\lambda_j \geq \tilde{\lambda}_j$, $\forall j$ and that $\lambda_{j_i} > U = \tilde{\lambda}_{j_i}$, and (34) is due to our construction (32) of $U$. So $h_i(l_i(\boldsymbol{\lambda})) < h_i(l_i(\tilde{\boldsymbol{\lambda}}))$ by (3). Second, if such

$j_i$ does not exist for some $i$, then $h_i(l_i(\boldsymbol{\lambda})) = h_i(l_i(\tilde{\boldsymbol{\lambda}}))$ since $l_i(\boldsymbol{\lambda}) = l_i(\tilde{\boldsymbol{\lambda}})$. Third, one can verify that there exists some $i$ for which such $j_i$ exists by the assumption of nonzero sensitivity factors. Combining these, we have $L(\boldsymbol{\lambda}) = \sum_i^N h_i(l_i(\boldsymbol{\lambda})) < \sum_i^N h_i(l_i(\tilde{\boldsymbol{\lambda}})) = L(\tilde{\boldsymbol{\lambda}})$. One can also show that "clipping" all elements of $\boldsymbol{\lambda}$ greater than $U$ will always decrease the roughness penalty $R$ in (6) due to our assumption that the potential function $\psi(x)$ is nondecreasing in $|x|$. Now, we have established that $\Phi(\boldsymbol{\lambda}) = L(\boldsymbol{\lambda}) - R(\boldsymbol{\lambda}) < L(\tilde{\boldsymbol{\lambda}}) - R(\tilde{\boldsymbol{\lambda}}) = \Phi(\tilde{\boldsymbol{\lambda}})$.

One can also construct an upper bound for a broader family of penalty functions more general than those based on differences of neighboring pixels with a nondecreasing potential function, although not shown in this paper.

## Appendix B

In this Appendix, we determine $\epsilon > 0$ such that (12) holds. Pick any $\boldsymbol{\nu} \in \mathcal{D}$, e.g., $\boldsymbol{\nu} = \mathbf{1}$. Define

$$\epsilon \triangleq \min_{i \in \mathcal{I}} \left\{ y_i,\ \exp\left( \frac{\Phi(\boldsymbol{\nu}) - \sum_{k \neq i} h_k(y_k)}{y_i} \right) \right\} \tag{35}$$

where $\mathcal{I}$ was defined in (10). For $i \in \mathcal{I}$, the condition $\epsilon \leq y_i$ implies that the modified marginal log-likelihood $\tilde{h}_i$ defined by (11) satisfies (2)–(4). The second inequality implied by (35) ensures that $y_i \log \epsilon \leq \Phi(\boldsymbol{\nu}) - \sum_{k \neq i} h_k(y_k)$ for $i \in \mathcal{I}$, which is used below.

First, we show that $\Lambda^* \subset \mathcal{E} = \{\boldsymbol{\lambda} \in \mathcal{D} : l_i(\boldsymbol{\lambda}) > \epsilon, \forall i \in \mathcal{I}\}$, where $\Lambda^*$ was defined in (8). Suppose that $\mathcal{I} \neq \emptyset$ (a nontrivial case) and that $\boldsymbol{\mu} \in \mathcal{D} \backslash \mathcal{E}$, i.e., $l_i(\boldsymbol{\mu}) \leq \epsilon$ for some $i \in \mathcal{I}$. Then, one can obtain

$$\begin{aligned}
\Phi(\boldsymbol{\mu}) &= \sum_{k=1}^{N} h_k\left(l_k(\boldsymbol{\mu})\right) - R(\boldsymbol{\mu}) \\
&\leq h_i\left(l_i(\boldsymbol{\mu})\right) + \sum_{k \neq i} h_k(y_k) && (36) \\
&\leq h_i(\epsilon) + \sum_{k \neq i} h_k(y_k) && (37) \\
&< y_i \log \epsilon + \sum_{k \neq i} h_k(y_k) && (38) \\
&\leq \Phi(\boldsymbol{\nu}) \leq \sup_{\boldsymbol{\lambda} \in \mathcal{D}} \Phi(\boldsymbol{\lambda}). && (39)
\end{aligned}$$

where (36) is a consequence of (2) and the assumption that $R$ is nonnegative; in (37) we used (3) with the fact that $l_i(\boldsymbol{\mu}) \leq \epsilon \leq y_i$; (38) is from the definition, $h_i(l) = y_i \log l - l$; and (39) is a consequence of (35). This implies that $\boldsymbol{\mu} \notin \Lambda^*$ for $\boldsymbol{\mu} \in \mathcal{D} \backslash \mathcal{E}$; that is, $\Lambda^* \subset \mathcal{E}$.

Similarly, one can verify that $\tilde{\Lambda}^{**} \triangleq \{\boldsymbol{\lambda}^* \in \mathbb{R}_+^p : \tilde{\Phi}(\boldsymbol{\lambda}^*) \geq \tilde{\Phi}(\boldsymbol{\lambda}),\ \forall \boldsymbol{\lambda} \in \mathbb{R}_+^p\} \subset \mathcal{E}$. But since $\Phi(\boldsymbol{\lambda}) = \tilde{\Phi}(\boldsymbol{\lambda})$ for $\boldsymbol{\lambda} \in \mathcal{E}$, we have $\tilde{\Lambda}^{**} = \Lambda^*$. Now since $\Lambda^* \subset \mathcal{B}$ by Appendix A, we have $\tilde{\Lambda}^{**} = \tilde{\Lambda}^{**} \cap \mathcal{B} = \tilde{\Lambda}^*$, where $\tilde{\Lambda}^*$ was defined in (12).

## Appendix C

In this Appendix, we prove that the modified BSREM-I (21) with (22) is globally convergent. The required assumptions on the objective function are the following: $\Phi(\boldsymbol{\lambda})$ is strictly concave on $\mathcal{B}$; and $\nabla f_m(\boldsymbol{\lambda})$ and $\boldsymbol{D}(\boldsymbol{\lambda}) \nabla f_m(\boldsymbol{\lambda})$ are Lipschitz continuous (and, thus, bounded) on $\mathcal{B}$. They are satisfied by our (modified) Poisson PL.

*Lemma 2:* Suppose that $\{\boldsymbol{\lambda}^{n,m}\}$ is a sequence generated by (21) with $\boldsymbol{\lambda}^0 \in \text{Int } \mathcal{B}$. Then, there exists $\alpha_0 > 0$ such that if $0 < \alpha_n \leq \alpha_0$, $\forall n$, then $\boldsymbol{\lambda}^{n,m} \in \text{Int } \mathcal{B}$, $\forall n, m$.

*Proof:* Since $(\partial / \partial \lambda_j) f_m(\boldsymbol{\lambda})$ is bounded over $\mathcal{B}$ for all $j$ and $m$, one can choose $\alpha_0 > 0$ such that

$$\alpha_0 \left| \frac{1}{p_j} \frac{\partial}{\partial \lambda_j} f_m(\boldsymbol{\lambda}) \right| < 1, \quad \forall \boldsymbol{\lambda} \in \mathcal{B} \text{ and } \forall j, m.$$

Suppose that $0 < \alpha_n \leq \alpha_0$, $\forall n$ and $\lambda^{n,m-1} \in \text{Int } \mathcal{B}$. If $0 < \lambda_j^{n,m-1} < U/2$, one can show that $0 < \lambda_j^{n,m} < U$, using the following expression for $\lambda_j^{n,m}$:

$$\begin{aligned}
\lambda_j^{n,m} &= \lambda_j^{n,m-1} + \alpha_n \frac{\lambda_j^{n,m-1}}{p_j} \frac{\partial}{\partial \lambda_j} f_m(\boldsymbol{\lambda}^{n,m-1}) \\
&= \lambda_j^{n,m-1} \left( 1 + \alpha_n \frac{1}{p_j} \frac{\partial}{\partial \lambda_j} f_m(\boldsymbol{\lambda}^{n,m-1}) \right).
\end{aligned}$$

If $U/2 \leq \lambda_j^{n,m-1} < U$, one can also show that $0 < \lambda_j^{n,m} < U$, using the following expression for $\lambda_j^{n,m}$:

$$U - \lambda_j^{n,m} = \left( U - \lambda_j^{n,m-1} \right) \left( 1 - \alpha_n \frac{1}{p_j} \frac{\partial}{\partial \lambda_j} f_m(\boldsymbol{\lambda}^{n,m-1}) \right).$$

This implies that $\lambda^{n,m} \in \text{Int } \mathcal{B}$. ∎

*Lemma 3:* Suppose that $\{\boldsymbol{\lambda}^n\}$ is a sequence generated by (21) with $\alpha_n > 0$ such that $\sum_{n=0}^{\infty} \alpha_n = \infty$ and $\sum_{n=0}^{\infty} \alpha_n^2 < \infty$. If $\boldsymbol{\lambda}^{n,m} \in \mathcal{B}, \forall n, m$, then $\{\Phi(\boldsymbol{\lambda}^n)\}$ converges in $\mathbb{R}$ and there exists a limit point $\boldsymbol{\lambda}^* \in \mathcal{B}$ of $\{\boldsymbol{\lambda}^n\}$ such that $\boldsymbol{D}(\boldsymbol{\lambda}^*) \nabla \Phi(\boldsymbol{\lambda}^*) = \mathbf{0}$

*Proof:* Using the definition of the sequence $\{\boldsymbol{\lambda}^n\}$, we have

$$\begin{aligned}
\boldsymbol{\lambda}^{n+1} &= \boldsymbol{\lambda}^n + \alpha_n \sum_{m=1}^{M} \boldsymbol{D}(\boldsymbol{\lambda}^{n,m-1}) \nabla f_m(\boldsymbol{\lambda}^{n,m-1}) \\
&= \boldsymbol{\lambda}^n + \alpha_n \boldsymbol{D}(\boldsymbol{\lambda}^n) \nabla \Phi(\boldsymbol{\lambda}^n) \\
&\quad + \alpha_n \sum_{m=1}^{M} \Big( \boldsymbol{D}(\boldsymbol{\lambda}^{n,m-1}) \nabla f_m(\boldsymbol{\lambda}^{n,m-1}) \\
&\qquad - \boldsymbol{D}(\boldsymbol{\lambda}^n) \nabla f_m(\boldsymbol{\lambda}^n) \Big) \\
&= \boldsymbol{\lambda}^n + \alpha_n \boldsymbol{D}(\boldsymbol{\lambda}^n) \nabla \Phi(\boldsymbol{\lambda}^n) + O\left(\alpha_n^2\right) && (40)
\end{aligned}$$

where the last equality is obtained as follows by Lipschitz continuity and boundedness of $\boldsymbol{D}(\boldsymbol{\lambda}) \nabla f_m(\boldsymbol{\lambda})$ on $\mathcal{B}$. In particular, for some positive $L \in \mathbb{R}$, we have

$$\begin{aligned}
&\left\| \sum_{m=1}^{M} \Big( \boldsymbol{D}(\boldsymbol{\lambda}^{n,m-1}) \nabla f_m(\boldsymbol{\lambda}^{n,m-1}) - \boldsymbol{D}(\boldsymbol{\lambda}^n) \nabla f_m(\boldsymbol{\lambda}^n) \Big) \right\| \\
&\leq \sum_{m=1}^{M} \left\| \boldsymbol{D}(\boldsymbol{\lambda}^{n,m-1}) \nabla f_m(\boldsymbol{\lambda}^{n,m-1}) - \boldsymbol{D}(\boldsymbol{\lambda}^n) \nabla f_m(\boldsymbol{\lambda}^n) \right\| \\
&\leq L \sum_{m=1}^{M} \left\| \boldsymbol{\lambda}^{n,m-1} - \boldsymbol{\lambda}^n \right\| \\
&\leq \alpha_n L \sum_{m=1}^{M} \sum_{k=1}^{m-1} \left\| \boldsymbol{D}(\boldsymbol{\lambda}^{n,k-1}) \nabla f_k(\boldsymbol{\lambda}^{n,k-1}) \right\| \\
&\leq \alpha_n L M^2 \max_{m, \boldsymbol{\lambda} \in B} \left\| \boldsymbol{D}(\boldsymbol{\lambda}) \nabla f_m(\boldsymbol{\lambda}) \right\|.
\end{aligned}$$

Now consider the objective sequence $\{\Phi(\boldsymbol{\lambda}^n)\}$. Since $\nabla\Phi(\boldsymbol{\lambda})$ is Lipschitz continuous on $\mathcal{B}$, we have [39, p. 6]

$$\Phi(\boldsymbol{\lambda}^{n+1}) = \Phi(\boldsymbol{\lambda}^n) + \nabla'\Phi(\boldsymbol{\lambda}^n)(\boldsymbol{\lambda}^{n+1} - \boldsymbol{\lambda}^n) + O\Big(\|\boldsymbol{\lambda}^{n+1} - \boldsymbol{\lambda}^n\|^2\Big). \quad (41)$$

Using (40) and (41), for large $n$, we establish

$$\Phi(\boldsymbol{\lambda}^{n+1}) = \Phi(\boldsymbol{\lambda}^n) + \alpha_n \nabla'\Phi(\boldsymbol{\lambda}^n)\boldsymbol{D}(\boldsymbol{\lambda}^n)\nabla\Phi(\boldsymbol{\lambda}^n) + O\big(\alpha_n^2\big). \quad (42)$$

Now, in view of (i) $\sum_{n=0}^{\infty}\alpha_n^2 < \infty$; (ii) the boundedness of $\Phi(\boldsymbol{\lambda})$ on $\mathcal{B}$; and (iii) the nonnegative definiteness of $\boldsymbol{D}(\boldsymbol{\lambda}^n)$, using (42), one can show that

$$\sum_{n=l}^{k}\alpha_n\nabla'\Phi(\boldsymbol{\lambda}^n)\boldsymbol{D}(\boldsymbol{\lambda}^n)\nabla\Phi(\boldsymbol{\lambda}^n) < q < \infty, \quad \forall k > l$$

for some $q \in \mathbb{R}$ and some large $l$. This implies that $\sum_{n=0}^{\infty}\alpha_n\nabla'\Phi(\boldsymbol{\lambda}^n)\boldsymbol{D}(\boldsymbol{\lambda}^n)\nabla\Phi(\boldsymbol{\lambda}^n) < \infty$. Given any $\epsilon > 0$, suppose that there exists $k$ such that $\nabla'\Phi(\boldsymbol{\lambda}^n)\boldsymbol{D}(\boldsymbol{\lambda}^n)\nabla\Phi(\boldsymbol{\lambda}^n) > \epsilon$, $\forall n > k$. Then, since $\sum_{n=0}^{\infty}\alpha_n = \infty$, we have $\sum_{n=0}^{\infty}\alpha_n\nabla'\Phi(\boldsymbol{\lambda}^n)\boldsymbol{D}(\boldsymbol{\lambda}^n)\nabla\Phi(\boldsymbol{\lambda}^n) = \infty$, which is a contradiction. So it must be the case that there exists a subsequence $\{\boldsymbol{\lambda}^{n_k}\}$ of $\{\boldsymbol{\lambda}^n\}$ such that $\lim_{k\to\infty}\boldsymbol{\lambda}^{n_k} = \boldsymbol{\lambda}^* \in \mathcal{B}$ with $\nabla'\Phi(\boldsymbol{\lambda}^*)\boldsymbol{D}(\boldsymbol{\lambda}^*)\nabla\Phi(\boldsymbol{\lambda}^*) = 0$, i.e, $\boldsymbol{D}(\boldsymbol{\lambda}^*)\nabla\Phi(\boldsymbol{\lambda}^*) = \boldsymbol{0}$ since $\boldsymbol{D}(\boldsymbol{\lambda}^*)$ is a nonnegative definite diagonal matrix.

On the other hand, from (42), one can show that $\{\Phi(\boldsymbol{\lambda}^n)\}$ is a Cauchy sequence in $\mathbb{R}$ in view of $\sum_{n=0}^{\infty}\alpha_n^2 < \infty$ and $\sum_{n=0}^{\infty}\alpha_n\nabla'\Phi(\boldsymbol{\lambda}^n)\boldsymbol{D}(\boldsymbol{\lambda}^n)\nabla\Phi(\boldsymbol{\lambda}^n) < \infty$. This implies that $\{\Phi(\boldsymbol{\lambda}^n)\}$ converges [36, p. 46]. ■

*Lemma 4:* Suppose that $\{\boldsymbol{\lambda}^n\}$ is a sequence generated by (21) with $\alpha_n > 0$ such that $\lim_{n\to\infty}\alpha_n = 0$. If $\boldsymbol{\lambda}^{n,m} \in \mathcal{B}$, $\forall n, m$, then $\lim_{n\to\infty}(\boldsymbol{\lambda}^{n,m} - \boldsymbol{\lambda}^n) = \boldsymbol{0}$, $\forall m$.

*Proof:* Since $\boldsymbol{D}(\boldsymbol{\lambda})\nabla f_m(\boldsymbol{\lambda})$ is bounded on $\mathcal{B}$, $\forall m$, using $\lim_{n\to\infty}\alpha_n = 0$, we have

$$\boldsymbol{\lambda}^{n,m} - \boldsymbol{\lambda}^n = \alpha_n\sum_{k=1}^{m}\boldsymbol{D}(\boldsymbol{\lambda}^{n,k-1})\nabla f_k(\boldsymbol{\lambda}^{n,k-1}) \to \boldsymbol{0}$$

as $n \to \infty$. ■

*Corollary:* $\lim_{n\to\infty}(\boldsymbol{\lambda}^{n+1} - \boldsymbol{\lambda}^n) = \boldsymbol{0}$.

*Lemma 5:* The limit point $\boldsymbol{\lambda}^* \in \mathcal{B}$ in Lemma 3 such that $\boldsymbol{D}(\boldsymbol{\lambda}^*)\nabla\Phi(\boldsymbol{\lambda}^*) = \boldsymbol{0}$ is a maximizer of $\Phi(\boldsymbol{\lambda})$ over $\mathcal{B}$ if $\boldsymbol{\lambda}^{n,m} \in \text{Int}\,\mathcal{B}$, $\forall n, m$.

*Proof:* We extend the proof of [17, Prop. 3]. It is clear that $(\partial/\partial\lambda_j)\Phi(\boldsymbol{\lambda}^*) = 0$ if $0 < \lambda_j^* < U$. Considering the optimality conditions [39, p. 203], we need to prove that $\lambda_j^* = 0$ implies $(\partial/\partial\lambda_j)\Phi(\boldsymbol{\lambda}^*) \le 0$, and $\lambda_j^* = U$ implies $(\partial/\partial\lambda_j)\Phi(\boldsymbol{\lambda}^*) \ge 0$. Define $\mathcal{J} = \mathcal{J}_1 \cup \mathcal{J}_2$ where $\mathcal{J}_1 = \{j = 1,\ldots,p : \lambda_j^* = 0$ and $(\partial/\partial\lambda_j)\Phi(\boldsymbol{\lambda}^*) > 0\}$ and $\mathcal{J}_2 = \{j = 1,\ldots,p : \lambda_j^* = U$ and $(\partial/\partial\lambda_j)\Phi(\boldsymbol{\lambda}^*) < 0\}$. We show that $\mathcal{J} = \emptyset$.

Since $\nabla\Phi$ is continuous on $\mathcal{B}$, there exists $0 < \delta < U/2$ such that if $\boldsymbol{\lambda} \in \mathcal{B}_\delta$, then $(\partial/\partial\lambda_j)\Phi(\boldsymbol{\lambda}) > 0$, $\forall j \in \mathcal{J}_1$ and $(\partial/\partial\lambda_j)\Phi(\boldsymbol{\lambda}) < 0$, $\forall j \in \mathcal{J}_2$, where $\mathcal{B}_\delta = \{\boldsymbol{\lambda} \in \mathcal{B} : \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^*\| < \delta\}$.

Suppose that $\boldsymbol{\lambda}^n \in \mathcal{B}_\delta$ where $n$ is sufficiently large. Then, using Lemma 4, we have $\boldsymbol{\lambda}^{n,m} \in \mathcal{B}_\delta$, $\forall m$ since $n$ is large. For $j \in \mathcal{J}_2$, since $d_j(\boldsymbol{\lambda}^{n,m}) = (U - \lambda_j^{n,m})/p_j$, $\forall m$, one can show

$$U - \lambda_j^{n,m} = (U - \lambda_j^{n,m-1})\left(1 - \frac{\alpha_n}{p_j}\frac{\partial}{\partial\lambda_j}f_m(\boldsymbol{\lambda}^{n,m-1})\right).$$

Then, using the boundedness and Lipschitz continuity of $\nabla f_m$, we have

$$\begin{aligned} U - \lambda_j^{n+1} &= (U - \lambda_j^n)\prod_{m=1}^{M}\left(1 - \frac{\alpha_n}{p_j}\frac{\partial}{\partial\lambda_j}f_m(\boldsymbol{\lambda}^{n,m-1})\right) \\ &= (U - \lambda_j^n)\left(1 - \frac{\alpha_n}{p_j}\sum_{m=1}^{M}\frac{\partial}{\partial\lambda_j}f_m(\boldsymbol{\lambda}^{n,m-1}) + O\big(\alpha_n^2\big)\right) \\ &= (U - \lambda_j^n)\left(1 - \frac{\alpha_n}{p_j}\frac{\partial}{\partial\lambda_j}\Phi(\boldsymbol{\lambda}^n) + O\big(\alpha_n^2\big)\right). \end{aligned}$$

Now, we have $U - \lambda_j^{n+1} > U - \lambda_j^n$, i.e., $\lambda_j^{n+1} < \lambda_j^n$ since $(\partial/\partial\lambda_j)\Phi(\boldsymbol{\lambda}^n) < 0$. Similarly, one can show that $\lambda_j^{n+1} > \lambda_j^n$ for $j \in \mathcal{J}_1$.

Let $\{\boldsymbol{\lambda}^{n_k}\}$ be a subsequence of $\{\boldsymbol{\lambda}^n\}$ such that $\lim_{k\to\infty}\boldsymbol{\lambda}^{n_k} = \boldsymbol{\lambda}^*$. Let $t_k = \max\{q < n_k : \boldsymbol{\lambda}^q \notin \mathcal{B}_\delta\}$. If $\boldsymbol{\lambda}^q \in \mathcal{B}_\delta$, $\forall q < n_k$ for some $k$, set $t_k = 0$. Then, $\{t_k\}$ is a monotone increasing sequence of nonnegative integers such that $\boldsymbol{\lambda}^q \in \mathcal{B}_\delta$ for $t_k + 1 \le q \le n_k$ for large $k$. Suppose that $\lim_{k\to\infty} t_k = t < \infty$, i.e., $\boldsymbol{\lambda}^n$ stays in $\mathcal{B}_\delta$ for large $n$. Then, $\lambda_j^k > \lambda_j^l > 0$, $\forall k > l$, $\forall j \in \mathcal{J}_1$ and $\lambda_j^k < \lambda_j^l < U$, $\forall k > l$, $\forall j \in \mathcal{J}_2$ for some large $l$. This is a contradiction since we have assumed that $\{\boldsymbol{\lambda}^n\}$ has a limit point $\boldsymbol{\lambda}^*$ such that $\lambda_j^* = 0$, $\forall j \in \mathcal{J}_1$ and $\lambda_j^* = U$, $\forall j \in \mathcal{J}_2$. So it must be the case that $\lim_{k\to\infty} t_k = \infty$. Now we have $\lambda_j^{n_k} > \lambda_j^{t_k+1} \ge 0$, $\forall j \in \mathcal{J}_1$ and $\lambda_j^{n_k} < \lambda_j^{t_k+1} \le U$, $\forall j \in \mathcal{J}_2$ for large $k$. Since $\lim_{k\to\infty}\lambda_j^{n_k} = 0$, $\forall j \in \mathcal{J}_1$ and $\lim_{k\to\infty}\lambda_j^{n_k} = U$, $\forall j \in \mathcal{J}_2$, we have $\lim_{k\to\infty}\lambda_j^{t_k+1} = 0$, $\forall j \in \mathcal{J}_1$ and $\lim_{k\to\infty}\lambda_j^{t_k+1} = U$, $\forall j \in \mathcal{J}_2$. By Corollary 1, $\lim_{k\to\infty}\lambda_j^{t_k} = 0$, $\forall j \in \mathcal{J}_1$ and $\lim_{k\to\infty}\lambda_j^{t_k} = U$, $\forall j \in \mathcal{J}_2$. Now one can construct a subsequence $\{\boldsymbol{\lambda}^{t_{k_l}}\}$ of $\{\boldsymbol{\lambda}^{t_k}\}$, which is also a subsequence of $\{\boldsymbol{\lambda}^n\}$, such that $\lim_{l\to\infty}\boldsymbol{\lambda}^{t_{k_l}} = \boldsymbol{\lambda}^{**}$ with $\lambda_j^{**} = 0$, $\forall j \in \mathcal{J}_1$ and $\lambda_j^{**} = U$, $\forall j \in \mathcal{J}_2$ but $\boldsymbol{\lambda}^{**} \ne \boldsymbol{\lambda}^*$ (since $\boldsymbol{\lambda}^{t_k} \notin \mathcal{B}_\delta$ and, thus, $\boldsymbol{\lambda}^{**} \notin \mathcal{B}_\delta$). Then, $\Phi(\boldsymbol{\lambda}^*) = \Phi(\boldsymbol{\lambda}^{**})$ by Lemma 3. We have two different maximizers $\boldsymbol{\lambda}^*$ and $\boldsymbol{\lambda}^{**}$ of $\Phi$ over $\{\boldsymbol{\lambda} \in \mathcal{B} : \lambda_j = 0$, $\forall j \in \mathcal{J}_1$ and $\lambda_j = U$, $\forall j \in \mathcal{J}_2\}$. This is a contradiction since $\Phi$ is strictly concave. So it must be the case that $\mathcal{J} = \emptyset$. ■

*Theorem 1:* A sequence $\{\boldsymbol{\lambda}^n\}$ generated by (21), with sufficiently small $\alpha_n > 0$ such that $\sum_{n=0}^{\infty}\alpha_n = \infty$ and $\sum_{n=0}^{\infty}\alpha_n^2 < \infty$, converges to $\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}\in\mathcal{B}}\Phi(\boldsymbol{\lambda})$.

*Proof:* By Lemmas 2, 3, and 5, the maximizer $\hat{\boldsymbol{\lambda}}$ is a limit point of $\{\boldsymbol{\lambda}^n\}$. Suppose that $\boldsymbol{\lambda}^{**}$ is a limit point of $\{\boldsymbol{\lambda}^n\}$. Then, $\Phi(\boldsymbol{\lambda}^{**}) = \Phi(\hat{\boldsymbol{\lambda}})$ by Lemma 3. This implies that $\boldsymbol{\lambda}^{**}$ is also a maximizer. By the uniqueness of the maximizer, $\boldsymbol{\lambda}^{**} = \hat{\boldsymbol{\lambda}}$. So $\{\boldsymbol{\lambda}^n\}$ has a unique limit point $\hat{\boldsymbol{\lambda}}$. This implies that the bounded sequence $\{\boldsymbol{\lambda}^n\}$ converges to $\hat{\boldsymbol{\lambda}}$ by [26, Prop. A.5, p. 652]. ■

*Corollary 2:* $\lim_{n\to\infty}\boldsymbol{\lambda}^{n,m} = \hat{\boldsymbol{\lambda}}$, $\forall m$.

*Proof:* Use Lemma 4 and Theorem 1. ■

## APPENDIX D

In this Appendix, we prove the global convergence of the diagonally scaled incremental gradient method (25). The required assumptions on the objective function are the following: $\nabla f_m$ is bounded on $\mathcal{B}$ and $f_m$ is concave. They are satisfied by our (modified) Poisson PL. Define a norm $\|\cdot\|_{\boldsymbol{D}^{-1}}$ on $\mathbb{R}^p$

by $\|\boldsymbol{\lambda}\|_{\boldsymbol{D}^{-1}} = (\boldsymbol{\lambda}'\boldsymbol{D}^{-1}\boldsymbol{\lambda})^{1/2}$ for $\boldsymbol{\lambda} \in \mathbb{R}^p$. Suppose that $\Phi^* = \sup_{\boldsymbol{\lambda}\in\mathcal{B}} \Phi(\boldsymbol{\lambda})$.

*Lemma 6:* Let $\{\boldsymbol{\lambda}^n\}$ be a sequence generated by (25). Then, for any $\boldsymbol{\lambda} \in \mathcal{B}$, one can show

$$\|\boldsymbol{\lambda}^{n+1}-\boldsymbol{\lambda}\|^2_{\boldsymbol{D}^{-1}} \le \|\boldsymbol{\lambda}^{n}-\boldsymbol{\lambda}\|^2_{\boldsymbol{D}^{-1}} - 2\alpha_n\left(\Phi(\boldsymbol{\lambda})-\Phi(\boldsymbol{\lambda}^n)\right) + \alpha_n^2 C$$

for all $n$ and some $C > 0$.

*Proof:* One can verify that the algorithm (25) is equivalent to the following:

$$\boldsymbol{x}^{n,m} = \mathcal{P}_{\mathcal{B}'}\left(\boldsymbol{x}^{n,m-1} + \alpha_n \nabla g_m(\boldsymbol{x}^{n,m-1})\right)$$

for $m = 1, 2, \ldots, M$, where $\boldsymbol{x}^{n,m} = \boldsymbol{D}^{1/2}\boldsymbol{\lambda}^{n,m}$, $g_m(\boldsymbol{x}) = f_m(\boldsymbol{D}^{1/2}\boldsymbol{x})$, and $\mathcal{B}' = \{\boldsymbol{x} \in \mathbb{R}^p : 0 \le x_j \le U d_j^{-1/2}\}$. Then, use [3, Lemma 2.1]. ∎

*Lemma 7:* Suppose that $\{\boldsymbol{\lambda}^n\}$ is a sequence generated by (25) with $\alpha_n > 0$ such that $\lim_{n\to\infty}\alpha_n = 0$ and $\sum_{n=0}^{\infty}\alpha_n = \infty$. Then, $\limsup_{n\to\infty}\Phi(\boldsymbol{\lambda}^n) = \Phi^*$.

*Proof:* The proof is due to [4, Prop. 1.2]. Assume for contradiction that there are $\delta > 0$, $N \in \mathbb{N}$, and $\boldsymbol{\nu} \in \mathcal{B}$ such that $\Phi(\boldsymbol{\nu}) > \Phi(\lambda^n) + \delta$ for all $n \ge N$. Since $\lim_{n\to\infty}\alpha_n = 0$, one can assume that $N$ is so large that $\alpha_n C < \delta$ where $C > 0$ is a constant from Lemma 6. Using Lemma 6, one obtains

$$\begin{aligned}\|\boldsymbol{\lambda}^{n+1}-\boldsymbol{\nu}\|^2_{\boldsymbol{D}^{-1}} &\le \|\boldsymbol{\lambda}^{n}-\boldsymbol{\nu}\|^2_{\boldsymbol{D}^{-1}} + \alpha_n(\alpha_n C - 2\delta)\\ &\le \|\boldsymbol{\lambda}^{n}-\boldsymbol{\nu}\|^2_{\boldsymbol{D}^{-1}} - \alpha_n\delta\end{aligned}$$

for all $n \ge N$. Summing up, this gives

$$0 \le \|\boldsymbol{\lambda}^{n}-\boldsymbol{\nu}\|^2_{\boldsymbol{D}^{-1}} \le \|\boldsymbol{\lambda}^{N}-\boldsymbol{\nu}\|^2_{\boldsymbol{D}^{-1}} - \delta\sum_{k=N}^{n-1}\alpha_n$$

for all $n > N$. This is a contradiction since $\sum_{n=0}^{\infty}\alpha_n = \infty$. ∎

*Theorem 2:* Let $\{\boldsymbol{\lambda}^n\}$ be the sequence generated by (25) with $\alpha_n > 0$ such that $\sum_{n=0}^{\infty}\alpha_n = \infty$ and $\sum_{n=0}^{\infty}\alpha_n^2 < \infty$. Then, $\{\boldsymbol{\lambda}^n\}$ converges to some $\boldsymbol{\lambda}^* \in \Lambda^* = \{\boldsymbol{\nu} \in \mathcal{B} : \Phi(\boldsymbol{\nu}) \ge \Phi(\boldsymbol{\lambda}), \forall \boldsymbol{\lambda} \in \mathcal{B}\}$.

*Proof:* Using Lemma 6 with some $\boldsymbol{\nu} \in \Lambda^*$, we have

$$\begin{aligned}\|\boldsymbol{\lambda}^{n+1}-\boldsymbol{\nu}\|^2_{\boldsymbol{D}^{-1}} \le \|\boldsymbol{\lambda}^{0}-\boldsymbol{\nu}\|^2_{\boldsymbol{D}^{-1}}\\ -2\sum_{k=0}^{n}\alpha_k\left(\Phi^*-\Phi(\boldsymbol{\lambda}^k)\right) + \sum_{k=0}^{n}\alpha_k^2 C \quad (43)\end{aligned}$$

for all $n$. Since $\sum_{n=0}^{\infty}\alpha_n^2 < \infty$, we have

$$2\sum_{k=0}^{n}\alpha_k\left(\Phi^*-\Phi(\boldsymbol{\lambda}^k)\right) \le \|\boldsymbol{\lambda}^{0}-\boldsymbol{\nu}\|^2_{\boldsymbol{D}^{-1}} + \sum_{k=0}^{n}\alpha_k^2 C < q < \infty$$

for all $n$ and some $q$ where $C$ is a constant from Lemma 6. This implies that $\sum_{k=0}^{\infty}\alpha_k(\Phi^* - \Phi(\boldsymbol{\lambda}^k)) < \infty$ since $\Phi^* - \Phi(\boldsymbol{\lambda}^k) \ge 0, \forall k$. Therefore, (43) implies that $\{\boldsymbol{\lambda}^n\}$ is bounded. By Lemma 7, there exists a subsequence $\{\boldsymbol{\lambda}^{n_k}\}$ of $\{\boldsymbol{\lambda}^n\}$ such that $\lim_{k\to\infty}\Phi(\boldsymbol{\lambda}^{n_k}) = \Phi^*$. Since $\{\boldsymbol{\lambda}^{n_k}\}$ is bounded, there exists a subsequence $\{\boldsymbol{\lambda}^{n_{k_l}}\}$ of $\{\boldsymbol{\lambda}^{n_k}\}$ such that $\{\boldsymbol{\lambda}^{n_{k_l}}\}$ converges to some $\boldsymbol{\lambda}^* \in \mathcal{B}$ [26, p. 652]. By the continuity of $\Phi$, we have $\Phi(\boldsymbol{\lambda}^*) = \Phi^*$, that is, $\boldsymbol{\lambda}^* \in \Lambda^*$. We have obtained a limit point $\boldsymbol{\lambda}^* \in \Lambda^*$ of $\{\boldsymbol{\lambda}^n\}$. Now, we follow the line of the proof of [4, Prop. 1.3]. For any $\delta > 0$, take $N \in \mathbb{N}$ such that $\|\boldsymbol{\lambda}^N - \boldsymbol{\lambda}^*\|^2_{\boldsymbol{D}^{-1}} \le \delta/2$ and $\sum_{k=N}^{\infty}(-2\alpha_k(\Phi^* - \Phi(\boldsymbol{\lambda}^k)) + \alpha_k^2 C) \le \delta/2$. Using Lemma 6, one obtains

$$\begin{aligned}\|\boldsymbol{\lambda}^{n+1}-\boldsymbol{\lambda}^*\|^2_{\boldsymbol{D}^{-1}} &\le \|\boldsymbol{\lambda}^{N}-\boldsymbol{\lambda}^*\|^2_{\boldsymbol{D}^{-1}}\\ &\quad + \sum_{k=N}^{n}\left(-2\alpha_k\left(\Phi^*-\Phi(\boldsymbol{\lambda}^k)\right) + \alpha_k^2 C\right) \le \delta\end{aligned}$$

for all $n \ge N$. ∎

*Corollary 3:* $\lim_{n\to\infty}\boldsymbol{\lambda}^{n,m} = \boldsymbol{\lambda}^* \in \Lambda^*, \forall m$.

*Proof:* Use $\lim_{n\to\infty}\alpha_n = 0$ with the assumption that $\nabla f_m$ is bounded on $\mathcal{B}$. ∎

## REFERENCES


[1] A. R. De Pierro and M. E. B. Yamagishi, "Fast EM-like methods for maximum '*a posteriori*' estimates in emission tomography," *IEEE Trans. Med. Imag.*, vol. 20, pp. 280–288, Apr. 2001.

[2] H. Erdoğan and J. A. Fessler, "Ordered subsets algorithms for transmission tomography," *Phys. Med. Biol.*, vol. 44, no. 11, pp. 2835–2851, Nov. 1999.

[3] A. Nedić and D. P. Bertsekas, "Incremental subgradient methods for nondifferentiable optimization," *SIAM J. Optim.*, vol. 12, no. 1, pp. 109–138, 2001.

[4] R. Correa and C. Lemaréchal, "Convergence of some algorithms for convex minimization," *Math. Program.*, vol. 62, pp. 261–275, 1993.

[5] L. A. Shepp, Y. Vardi, J. B. Ra, S. K. Hilal, and Z. H. Cho, "Maximum likelihood PET with real data," *IEEE Trans. Nucl. Sci.*, vol. NS–31, pp. 910–913, Apr. 1984.

[6] J. Qi and R. H. Huesman, "Theoretical study of lesion detectability of MAP reconstruction using computer observers," *IEEE Trans. Med. Imag.*, vol. 20, pp. 815–822, Aug. 2001.

[7] R. Gordon, R. Bender, and G. T. Herman, "Algebraic reconstruction techniques (ART) for the three-dimensional electron microscopy and X-ray photography," *J. Theor. Biol.*, vol. 29, pp. 471–481, 1970.

[8] G. T. Herman and L. B. Meyer, "Algebraic reconstruction techniques can be made computationally efficient," *IEEE Trans. Med. Imag.*, vol. 12, pp. 600–609, Sept. 1993.

[9] Y. Censor, P. P. B. Eggermont, and D. Gordon, "Strong underrelaxation in Kaczmarz's method for inconsistent systems," *Numerische Mathematik*, vol. 41, pp. 83–92, 1983.

[10] Y. Censor, D. Gordon, and R. Gordon, "Component averaging: an efficient iterative parallel algorithm for large and sparse unstructured problems," *Parallel Computing*, vol. 27, pp. 777–808, 2001.

[11] Y. Censor and S. A. Zenios, *Parallel Optimization: Theory, Algorithms, and Applications*. New York: Oxford Univ. Press, 1997.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. Ser. B*, vol. 39, no. 1, pp. 1–38, 1977.

[13] L. A. Shepp and Y. Vardi, "Maximum likelihood reconstruction for emission tomography," *IEEE Trans. Med. Imag.*, vol. MI–1, pp. 113–122, Oct. 1982.

[14] K. Lange and R. Carson, "EM reconstruction algorithms for emission and transmission tomography," *J. Comput. Assist. Tomogr.*, vol. 8, no. 2, pp. 306–316, Apr. 1984.

[15] H. M. Hudson and R. S. Larkin, "Accelerated image reconstruction using ordered subsets of projection data," *IEEE Trans. Med. Imag.*, vol. 13, pp. 601–609, Dec. 1994.

[16] C. L. Byrne, "Accelerating the EMML algorithm and related iterative algorithms by rescaled block-iterative methods," *IEEE Trans. Image Processing*, vol. 7, pp. 100–109, Jan. 1998.

[17] J. A. Browne and A. R. De Pierro, "A row-action alternative to the EM algorithm for maximizing likelihoods in emission tomography," *IEEE Trans. Med. Imag.*, vol. 15, pp. 687–699, Oct. 1996.

[18] I. T. Hsiao, A. Rangarajan, and G. Gindi, "A provably convergent OS-EM like reconstruction algorithm for emission tomography," in *Proc. SPIE 4684, Medical Imaging 2002: Image Proc.*, 2002, pp. 10–19.

[19] J. A. Fessler and H. Erdogan, "A paraboloidal surrogates algorithm for convergent penalized-likelihood emission image reconstruction," in *Proc. IEEE Nuclear Science Symp. Medical Imaging Conf.*, vol. 2, 1998, pp. 1132–1135.

[20] H. Erdoğan and J. A. Fessler, "Monotonic algorithms for transmission tomography," *IEEE Trans. Med. Imag.*, vol. 18, pp. 801–814, Sept. 1999.

[21] D. G. Luenberger, *Linear and Nonlinear Programming*, 2nd ed. Reading, MA: Addison-Wesley, 1984.
[22] H. Kudo, H. Nakazawa, and T. Saito, "Convergent block-iterative method for general convex cost functions," in *Proc. 1999 Int. Mtg. Fully 3D Im. Recon. in Rad. Nuc. Med.*, 1999, pp. 247–250.
[23] ——, "Block-gradient method for image reconstruction in emission tomography," *Trans. IEICE*, vol. J83-D-II, no. 1, pp. 63–73, Jan. 2000. In Japanese.
[24] A. Nedić and D. Bertsekas, "Convergence rate of incremental subgradient algorithms," in *Stochastic Optimization: Algorithms and Applications*, S. P. Uryasev and P. M. Pardalos, Eds. Norwell, MA: Kluwer, 2000, pp. 263–304.
[25] V. M. Kibardin, "Decomposition into functions in the minimization problem," *Automat. Remote Control*, vol. 40, pp. 1311–1323, 1980.
[26] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA: Athena Scientific, 1999.
[27] ——, "A new class of incremental gradient methods for least squares problems," *SIAM J. Optim.*, vol. 7, no. 4, pp. 913–926, Nov. 1997.
[28] R. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse and other variants," in *Learning in Graphical Models*, M. I. Jordan, Ed. Dordrencht: Kluwer, 1998, pp. 355–368.
[29] A. J. R. Gunawardana, "The information geometry of EM variants for speech and image processing," Ph.D. dissertation, The Johns Hopkins Univ., Baltimore, MD, Apr. 1999.
[30] S. Ahn and J. A. Fessler, "Globally convergent ordered subsets algorithms: application to tomography," in *Proc. IEEE Nuclear Science Symp. Medical Imaging Conf.*, vol. 2, 2001, pp. 1064–1068.
[31] S. Sotthivirat and J. A. Fessler, "Relaxed ordered-subsets algorithm for penalized-likelihood image restoration," *J. Opt. Soc. Amer. A*, vol. 20, no. 3, pp. 439–449, Mar. 2003.
[32] M. Yavuz and J. A. Fessler, "Statistical image reconstruction methods for randoms-precorrected PET scans," *Med. Imag. Anal.*, vol. 2, no. 4, pp. 369–378, 1998.
[33] K. Lange, "Convergence of EM image reconstruction algorithms with Gibbs smoothing," *IEEE Trans. Med. Imag.*, vol. 9, pp. 439–446, Dec. 1990.
[34] J. A. Fessler, "Penalized weighted least-squares image reconstruction for positron emission tomography," *IEEE Trans. Med. Imag.*, vol. 13, pp. 290–300, June 1994.
[35] C. Hamaker and D. C. Solmon, "The angles between the null spaces of x rays," *J. Math. Anal. Appl.*, vol. 62, pp. 1–23, 1978.
[36] K. A. Ross, *Elementary Analysis: The Theory of Calculus*. New York: Springer-Verlag, 1980.
[37] J. A. Fessler. (1995, July) ASPIRE 3.0 User's Guide: A Sparse Iterative Reconstruction Library. Commun. Signal Proc. Lab., Dept. Elec. Eng. Comput. Sci., Univ. Michigan, Ann Arbor, MI, Tech. Rep. 293. [Online]Available: http:www.eecs.umich.edu/~fessler
[38] A. R. De Pierro, "A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography," *IEEE Trans. Med. Imag.*, vol. 14, pp. 132–137, Mar. 1995.
[39] B. T. Polyak, *Introduction to Optimization*. New York: Optimization Software, 1987.