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 algorithms for penalized-likelihood image reconstruction in emission tomography: modified BSREM (block sequential regularized expectation maximization) and relaxed OS-SPS (ordered subsets separable paraboloidal surrogates). The global convergence proof of the existing BSREM [1] 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 [2] 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 [3] and [4]. 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 reconstruction in emission tomography, and [6] for the analysis of lesion detectibility). 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 maximum likelihood (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 ordered subsets 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 “ordered subsets” 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 ordered subsets principle was applied to the classical expectation maximization (EM) algorithm [12–14] to yield several ordered subsets expectation maximization variants. ML reconstruction algorithms include the ordered subsets expectation maximization (OS-EM) algorithm [15], the row-action maximum likelihood (RAMLA) [17], and the complete-data OSEM (C-OSEM) [18]. PL reconstruction algorithms include the block sequential regularized expectation maximization (BSREM) algorithm [1] (BSREM has RAMLA as a special unregularized case). The paraboloidal surrogates (PS) methods [19, 20] also adopted the ordered subsets idea to construct the ordered subsets separable paraboloidal surrogates (OS-SPS) [2], originally named the ordered subsets 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 proportional to the stepsize. BSREM and RAMLA use diminishing relaxation parameters [1, 17]. De Pierro and Yamag-
ishii [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 (ordered subsets) 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 penalized-likelihood 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. Penalized-Likelihood 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 \(A = \{a_{ij}\}\) 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 \(\lambda^{\text{true}} = [\lambda_1^{\text{true}}, \lambda_2^{\text{true}}, \ldots, \lambda_p^{\text{true}}]^T\) based on the measurement \(y = [y_1, y_2, \ldots, y_N]^T\) with \(A\) and \(r = [r_1, r_2, \ldots, r_N]^T\) being known where \(^1\) denotes matrix transpose. We assume that the sensitivity factors, \(\sum_{i=1}^{N} a_{ij}\), are non-zero for all \(j\), which is reasonable in practice.

The log-likelihood of \(\lambda\) given \(y\) can be written, ignoring constants independent of \(\lambda\), as follows:

\[
L(\lambda) = \sum_{i=1}^{N} h_i(l_i(\lambda))
\]

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

\[
\begin{alignat}{2}
& (i) \quad h_i(l) \leq h_i(y_i), & \forall l \geq 0, \\
& (ii) \quad h_i \text{ is monotone increasing on } [0, y_i], & \text{and monotone decreasing on } [y_i, \infty), \\
& (iii) \quad h_i \text{ is concave on } [0, \infty),
\end{alignat}
\]

For penalized-likelihood (PL) reconstruction, one must find a maximizer of the following objective function over its domain \(D\):

\[
\Phi(\lambda) = L(\lambda) - R(\lambda)
\]

where

\[
D \triangleq \{\lambda \in \mathbb{R}_+^p : \Phi(\lambda) \in \mathbb{R}\}
\]

\[
= \{\lambda \in \mathbb{R}_+^p : l_i(\lambda) > 0 \text{ or } y_i = 0, \forall i\}
\]

with

\[
\mathbb{R}_+^p = \{\lambda \in \mathbb{R}^p : \lambda_j \geq 0, \forall j\},
\]

and \(R\) is a regularization term. The reason for taking the domain \(D\) instead of \(\mathbb{R}_+^p\) is that the gradient of the log-likelihood is infinite on \(\mathbb{R}_+^p \setminus D\). The use of the feasible domain \(D\) 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:

\[
R(\lambda) = \frac{\beta}{2} \sum_{j=1}^{p} \sum_{k \in N_j} \omega_{jk} (\lambda_j - \lambda_k)
\]

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

\(^2\) For convenience, we adopt the convention that \(\log 0 = -\infty\) and \(0 \cdot \log 0 = 0\).
where $\beta \geq 0$ is a regularization parameter that controls the smoothness of the reconstructed image, $N_j$ 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, $N_j$ 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 maximum likelihood (ML) reconstruction.

### A.1 Existence and Uniqueness

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

If the objective function $\Phi$ is strictly concave on $D$, then there exists a unique PL solution [26, p. 685], $\lambda = \arg \max_{\lambda \in D} \Phi(\lambda)$. 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 (cf. Theorem 1 of [34] and Lemma 1 of [33]) 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 $y' A 1 \neq 0$, then $\Phi$ in (5) (with (6) for $\beta > 0$) is strictly concave on $D$ for any $\psi$ that is strictly convex and twice differentiable.

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

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

with

$$W(\lambda) = \text{diag}\left\{ y_i \frac{2}{\ell_i^2(\lambda)} \right\}$$

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

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

Since $\hat{\psi} > 0$ and the neighborhood system is connected by assumption, for $\beta > 0$, $x' \nabla^2 R(\lambda) x = 0$ only if $x = 0$ or $x = c 1$ for some $c \neq 0$. But $c 1' A' W(\lambda) A c 1 = c^2 \|W^{1/2}(\lambda) A 1\|^2$ is not 0 by assumption. So $x' \nabla^2 \Phi(\lambda) x < 0, \forall x \neq 0$.

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

### A.2 Boundedness

It is clear that a PL solution set

$$\Lambda^* = \{ \lambda^* \in D : \Phi(\lambda^*) \geq \Phi(\lambda), \forall \lambda \in D \}$$

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

$$\Lambda^* \subset B = \{ \lambda \in \mathbb{R}^p : 0 \leq \lambda_j \leq U, \forall j \}.$$

See Appendix A for a method of determining $U$. Thus, one can search for a solution over the bounded set $B \cap D$ instead of over $D$. 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 $D$, whereas it is bounded on $B \cap D$.

### A.3 Differentiability

The objective function $\Phi$ is not differentiable on the set

$$S = \mathbb{R}^p \setminus D = \{ \lambda \in \mathbb{R}^p : \ell_i(\lambda) = 0 \text{ for some } i \in I \}$$

where

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

One can see that $\| \nabla \Phi(\lambda) \| = \infty$ for $\lambda \in S$. If a gradient-based algorithm took a point in $S$, it would collapse. Note that $I = \emptyset$ and thus $S = \emptyset$ 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 $S$ with well-behaved functions, e.g., quadratic approximations. We consider the following modified objective function:

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

where

$$\tilde{h}_i(l) = \begin{cases} \frac{h_i(\epsilon)(l - \epsilon)^2 + h_i(\epsilon)(l - \epsilon) + h_i(\epsilon)}{\epsilon^2} & \text{for } l \leq \epsilon \text{ and } i \in I \\ h_i(l) & \text{otherwise} \end{cases}$$

for some $\epsilon > 0$. The modified marginal log-likelihood $\tilde{h}_i$ is a strictly concave real-valued function defined on $\mathbb{R}$ for $i \in I$.

Note that $\Phi(\lambda) = \tilde{\Phi}(\lambda)$ for $\lambda \in \{ \lambda \in D : \ell_i(\lambda) > \epsilon, \forall i \in I \}$ and that $\Phi$ 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

\[
\Lambda^* = \tilde{\Lambda}^* = \left\{ \lambda^* \in B : \tilde{\Phi}(\lambda^*) \geq \Phi(\lambda), \forall \lambda \in B \right\},
\]

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 \( B \), and it has a nice property that its gradient \( \nabla \tilde{\Phi} \) is Lipschitz continuous\(^5\) on \( B \). We will henceforth take \( \Phi \) as our objective function but revert to the notation \( \tilde{\Phi} \) for simplicity; likewise, \( h_i \) will denote \( h_i \) for \( i \in I \). 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 \( I = \emptyset \) and we need not modify the objective function.

\section*{B. Ordered Subsets (OS) Algorithms}

Most iterative algorithms for finding a maximizer of a penalized-likelihood 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:\(^7\)

\[
\lambda_j^{n+1} = \lambda_j^n + \alpha_n d_j(\lambda^n) \frac{\partial}{\partial \lambda_j} \tilde{\Phi}(\lambda^n), \quad j = 1, 2, \ldots, p,
\]

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

\[
\lambda^{n+1} = \lambda^n + \alpha_n D(\lambda^n) \nabla \tilde{\Phi}(\lambda^n),
\]

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

\[
\frac{\partial}{\partial \lambda_j} \Phi(\lambda) = \sum_{i=1}^N a_{ij} h_i(\lambda_j(\lambda)) - \frac{\partial}{\partial \lambda_j} R(\lambda).
\]

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

Ordered subsets (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

\[
f_m(\lambda) = \sum_{i \in S_m} h_i(\lambda) - \gamma_m R(\lambda)
\]

be a sub-objective function, resulting in

\[
\Phi = \sum_m f_m,
\]

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

\[
\nabla f_1(\lambda) \approx \nabla f_2(\lambda) \approx \cdots \approx \nabla f_M(\lambda)
\]

for \( \lambda \) far from the solution set, or, equivalently,

\[
\nabla \Phi(\lambda) \approx M \nabla f_m(\lambda), \quad \forall m.
\]

Then ordered subsets version of (13) is obtained by substituting \( M \frac{\partial}{\partial \lambda_j} f_m(\lambda) \) for \( \frac{\partial}{\partial \lambda_j} \Phi(\lambda) \), as follows:\(^8\)

\[
\lambda_j^{n,m} = \lambda_j^{n,m-1} + \alpha_n d_j(\lambda^{n,m-1}) \frac{\partial}{\partial \lambda_j} f_m(\lambda^{n,m-1}),
\]

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

\[
\lambda_j^{n,0} = \lambda_j^n \quad \text{and} \quad \lambda_j^{n+1} = \lambda_j^{n,M}.
\]

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

\[
\Phi(x) = \sum_{i=1}^3 \left( -\frac{1}{2} x' Q_i x + b_i' x \right)
\]

where \( Q_1 = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}, \; Q_2 = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}, \; Q_3 = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 1 & 2.5 \end{bmatrix}, \; b_1 = \begin{bmatrix} 1.25 \\ 2 \end{bmatrix}, \; b_2 = \begin{bmatrix} -1.25 \\ -1 \end{bmatrix}, \; b_3 = \begin{bmatrix} 0.25 \end{bmatrix}.

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].
where $\alpha = 0.05$, and its ordered subsets version with 3 subsets:

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

and proved its global convergence under the following assumptions: the sequence $\{\lambda^n\}$ generated by the algorithm is positive and bounded; and the objective sequence $\{\Phi(\lambda^n)\}$ 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 $B$ 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 $d_j(\lambda) = \lambda_j$ 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:

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

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

$$d_j(\lambda) = \begin{cases} 
\frac{\lambda_j}{p_j} & \text{for } 0 \leq \lambda_j < \frac{U}{2}, \\
(U - \lambda_j)/p_j & \text{for } \frac{U}{2} \leq \lambda_j \leq U
\end{cases}$$

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

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 $B$, then the iterates generated by (21) automatically stay in the interior of $B$. The second part (Lemma 3–5) is about convergence: the iterates generated by (21) converge to the solution $\lambda = \text{arg max}_{\lambda \in \mathbb{R}^M} \Phi(\lambda)$ if (iii) $\sum_{n=0}^{\infty} \alpha_n = \infty$, (iv) $\sum_{n=0}^{\infty} \alpha_n^2 < \infty$, and (v) $\lambda^{n,m} \in \text{Int } B$, $\forall n, m$ where $\text{Int } B$ denotes the interior of $B$. 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, 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:

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

where $\mathcal{P}_T(\lambda)$ is the projection of $\lambda \in \mathbb{R}^M$ onto $T = \{\lambda \in \mathbb{R}^M : t \leq \lambda_j \leq U - t, \forall j\}$ 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 \([36, \text{p. 70}]\) that \(\lim_{n \to \infty} \alpha_n = 0\), there exists \(N \in \mathbb{N}\) such that \(\alpha_n \) satisfies (i) for \(n \geq N\). Treating \(\lambda^N \in \text{Int} B\) 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:

\[
p_j = \sum_{i=1}^{N} a_{ij} / M.
\]  

(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 \(\tilde{U} = \lambda_j\) 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 to the log-likelihood scaling by \(\lambda_j\). This is verified experimentally in Section IV.

If we take larger and larger \(U\), then \(B \to \mathbb{R}^p_+\) and \(d_j(\lambda) \to \lambda_j/p_j\). 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:

\[
\lambda^{n,m} = \mathcal{P}_B(\lambda^{n,m-1} + \alpha_n D \nabla f_m(\lambda^{n,m-1}))
\]  

(25)

for \(m = 1, 2, \ldots, M\) where \(\alpha_n > 0\) and \(D = \text{diag}\{d_j\}\) with \(d_j > 0, \forall j\), and \(\mathcal{P}_B(\lambda)\) is the projection of \(\lambda \in \mathbb{R}^p\) onto \(B\). We call these algorithms diagonally-scaled incremental gradient methods since if we choose \(D = I\), 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 (i) \(\sum_{n=0}^{\infty} \alpha_n = \infty\) and (ii) \(\sum_{n=0}^{\infty} \alpha_n^2 < \infty\) as shown in Theorem 2

Here the projection \(\mathcal{P}_F(\lambda)\) is easily calculated componentwise as \(\mathcal{P}_F(\lambda)_j := t\) for \(\lambda_j < t\); \(\mathcal{P}_F(\lambda)_j := \tilde{U} - t\) for \(\lambda_j > \tilde{U} - t\); and \(\mathcal{P}_F(\lambda)_j := \lambda_j\) otherwise. So \((23)\) can be written componentwise as \(\lambda^{n,m} = \lambda^{n,m-1} + \alpha_n \lambda_j^{n,m-1} < 0\), \(\lambda^{n,m} = \tilde{U} - t\) for \(\lambda_j^{n,m} > \tilde{U}\), and \(\lambda^{n,m} = \lambda^{n,m}_j\) otherwise.

The global convergence holds regardless of \(D\) as long as it is diagonal with positive elements.

A practical issue is how to choose \(D\) for fast convergence rate and easy stepsize selection. Fortunately, some hints are given by observing that the ordered subsets separable paraboloidal surrogates (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:

\[
d_j = M \left( \sum_{i=1}^{N} a_{ij} w_i + 2\beta \sum_{k \in N_j} \omega_{jk} \right)^{-1}, \quad \forall j
\]  

(26)

where \(a_{ij} = \sum_{j=1}^{p} a_{ij}\), \(M\) is the number of subsets, and \(w_i = \begin{cases} -\tilde{h}_i(y_i) & \text{for } y_i > 0 \\ 0 & \text{otherwise.} \end{cases}\)

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 non-quadratic 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 \(r_j > 0\) for monotonicity and convergence, we eliminate this requirement here by the modification \((11)\) of the penalized-likelihood.

One of required conditions for the global convergence proofs of diagonally-scaled incremental gradient methods is the boundedness of \(\nabla f_m\) on \(B\). If the gradient \(\nabla R\) of the penalty part is bounded on \(\mathbb{R}^p_+\), then we can take \(B = \mathbb{R}^p_+\) while retaining global convergence since the gradient of the (modified) log-likelihood is bounded on \(\mathbb{R}^p_+\). Such penalties include the Huber penalty:

\[
\psi(x) = \begin{cases} x^2/2, & \text{for } |x| \leq \delta \\ \delta |x| - \delta^2/2, & \text{otherwise,} \end{cases}
\]  

for some \(\delta > 0\).

### C. Regularization into OS Algorithms

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

\[
\gamma_m = |S_m|/N, \quad \forall m
\]  

(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 sub-objective function as in [1]:

\[
\gamma_m = 0 \text{ for } m = 1, 2, \ldots, M, \text{ and } \gamma_{M+1} = 1 \quad (28)
\]

where we have \((M + 1)\) sub-objective 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 never get to the optimum point if an initial point is sufficiently far from it.

One could write a more general form of OS algorithms by allowing different scaling matrices over subiterations:

\[
\lambda^{n,m} = \lambda^{n,m-1} + \alpha_n D_m(\lambda^{n,m-1}) \nabla f_m(\lambda^{n,m-1}), \forall m \quad (29)
\]

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

\[
\lambda^{n+1} = \lambda^n + \alpha_n \sum_{m=1}^{M} D_m(\lambda^{n,m-1}) \nabla f_m(\lambda^{n,m-1})
\]

\[
= \lambda^0 + \sum_{k=0}^{n} \alpha_k \sum_{m=1}^{M} D_m(\lambda^{k,m-1}) \nabla f_m(\lambda^{k,m-1}).
\]

Now suppose that the sequence \(\{\lambda^{n,m}\}\) generated by (29) converges to some \(\lambda^*\). Assuming that \(\lambda_{n} \nabla f_{m}\) is continuous, we have:

\[
\lim_{k \to \infty} D_m(\lambda^{k,m-1}) \nabla f_m(\lambda^{k,m-1}) = D_m(\lambda^*) \nabla f_m(\lambda^*).
\]

\[\text{If } \sum_{m=1}^{M} D_m(\lambda^*) \nabla f_m(\lambda^*) \neq 0, \text{ then } \{\lambda^n\} \text{ diverges since } \sum_{n=0}^{\infty} \alpha_n = \infty. \text{ So it must be the case that:}
\]

\[
\sum_{m=1}^{M} D_m(\lambda^*) \nabla f_m(\lambda^*) = 0. \quad (30)
\]

However, if the \(D_m\)'s are different, then (30) is generally different from the true optimality conditions, e.g., \(\nabla \Phi(\lambda^*) = \sum_{m=1}^{M} \nabla f_m(\lambda^*) = 0\) for unconstrained optimization. So, in general, OS algorithms with subiteration-dependent scaling matrices, including OS-EM and rescaled block-iterative expectation maximization (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 penalized-likelihood 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 \(\nabla f_m(\lambda)\) and \(D(\lambda) \nabla f_m(\lambda)\) are Lipschitz continuous on \(B\). Diagonally-scaled incremental gradient methods including relaxed OS-SPS require that \(\nabla f_m\) is bounded on \(B\) 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. The relaxation parameters obtained thus 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:

\[
\alpha_n = \frac{\alpha_0}{\gamma n + 1}, \forall n \quad (31)
\]

for \(\gamma > 0\) and \(\alpha_0 > 0\). We run simulations using these simple relaxation parameters. Our goal here is 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 3.6mm ray spacing and 120 angles over 360 degrees, and the reconstructed images were 128 \(\times\) 128 with 3.6mm pixel size. The distance from the center of rotation to the detector plane was 288mm. The system matrix \(A\) 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):

$$\text{FWHM} = \sqrt{(0.0868056 \cdot z)^2 + (3\text{mm})^2}$$

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 $(\Phi(\hat{\lambda}) - \Phi(\lambda^*))/(\Phi(\hat{\lambda}) - \Phi(0^*))$ versus iteration number where $\hat{\lambda}$ is the solution estimated by 5,000 iterations of De Pierro’s modified EM, a globally convergent method [38]. One can see that the scaling factors (22) with (24), and (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 the logarithmic scale). We chose $\alpha_n = 1/(\frac{1}{n} + 1)$ 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 $\alpha_n = 1/(\frac{1}{n} + 1)$ 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 $||\lambda^* - \hat{\lambda}||$ 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 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 penalized-likelihood 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

$$U = \max_i \left\{ \frac{\psi_i}{\min_{j: a_{ij} \neq 0, a_{ij} > 0} a_{ij}} \right\}.$$  (32)

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

\[
\begin{align*}
 l_i(\lambda) &= a_{ij_i} \lambda_{j_i} + \sum_{j \neq j_i} a_{ij}\lambda_j + r_i \\
&> a_{ij_i} U + \sum_{j \neq j_i} a_{ij}\lambda_j + r_i = l_i(\hat{\lambda}) \\
&\geq a_{ij_i} U \geq y_i 
\end{align*}
\]

(33) (34)

where in (33) we used the fact that \( \lambda_j \geq \hat{\lambda}_j, \forall j \) and that \( \hat{\lambda}_j > U = \hat{\lambda}_j \), and (34) is due to our construction (32) of \( U \). So \( h_i(l_i(\lambda)) < h_i(l_i(\hat{\lambda})) \) by (3). Second, if such \( j_i \) does not exist for some \( i \), then \( h_i(l_i(\lambda)) = h_i(l_i(\hat{\lambda})) \) since \( l_i(\lambda) = l_i(\hat{\lambda}) \). Third, one can verify that there exists some \( i \) for which such \( j_i \) exists by the assumption of non-zero sensitivity factors. Combining these, we have \( L(\lambda) = \sum_{i=1}^{N} h_i(l_i(\lambda)) < \sum_{i=1}^{N} h_i(l_i(\hat{\lambda})) = L(\hat{\lambda}) \). One can also show that “clipping” all elements of \( \lambda \) 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 \( \phi(\lambda) = L(\lambda) - R(\lambda) < L(\hat{\lambda}) - R(\hat{\lambda}) = \Phi(\hat{\lambda}) \).

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 \( \nu \in D \), e.g., \( \nu = 1 \). Define

\[
\epsilon = \min_{i \in I} \left\{ y_i, \exp\left( \frac{\Phi(\nu) - \sum_{k \neq i} h_k(y_k)}{y_i} \right) \right\} 
\]

(35)

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

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

\[
\Phi(\mu) = \sum_{k=1}^{N} h_k(l_k(\mu)) \leq h_i(l_i(\mu)) + \sum_{k \neq i} h_k(y_k) \leq h_i(\epsilon) + \sum_{k \neq i} h_k(y_k) \leq y_i \log \epsilon + \sum_{k \neq i} h_k(y_k) \leq \Phi(\nu) \leq \sup_{\lambda \in D} \Phi(\lambda).
\]

(36) (37) (38) (39)

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

Similarly, one can verify that \( \Lambda^\ast \overset{\Delta}{=} \{ \lambda^\ast \in \mathbb{R}_+^M : \Phi(\lambda^\ast) \geq \Phi(\lambda), \forall \lambda \in \mathbb{R}_+^M \} \subset \mathcal{E} \). But since \( \Phi(\lambda) = \Phi(\hat{\lambda}) \) for \( \lambda \in \mathcal{E} \), we have \( \Lambda^\ast = \Lambda^\ast \cap B = \Lambda^\ast \) where \( \Lambda^\ast \) 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: \( \Phi(\lambda) \) is strictly concave on \( B \) and \( \nabla f_m(\lambda) \) and \( D(\lambda) \nabla f_m(\lambda) \) are Lipschitz continuous (and thus bounded) on \( B \). They are satisfied by our (modified) Poisson penalized-likelihood.

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

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

\[
\alpha_0 \frac{1}{|p_j|} \frac{\partial}{\partial \lambda_j} f_m(\lambda) < 1, \forall \lambda \in B \text{ and } \forall j, m.
\]

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

\[
\lambda^{n,m} = \lambda^{n,m-1} + \alpha_n \frac{\lambda^{n,m-1}}{|p_j|} \frac{\partial}{\partial \lambda_j} f_m(\lambda^{n,m-1})
\]

\[
= \lambda^{n,m-1} \left( 1 + \alpha_n \frac{1}{|p_j|} \frac{\partial}{\partial \lambda_j} f_m(\lambda^{n,m-1}) \right).
\]

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

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

This implies that \( \lambda^{n,m} \in \text{Int } B \). \( \blacksquare \)

**Lemma 3:** Suppose that \( \{ \lambda^n \} \) 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 \( \lambda^{n,m} \in B, \forall n, m \), then \( \{ \Phi(\lambda^n) \} \) converges in \( \mathbb{R} \) and there exists a limit point \( \lambda^\ast \in B \) of \( \{ \lambda^n \} \) such that

\[
D(\lambda^\ast) \nabla \Phi(\lambda^\ast) = 0.
\]

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

\[
\lambda^{n+1} = \lambda^n + \alpha_n \sum_{m=1}^{M} D(\lambda^{n,m-1}) \nabla f_m(\lambda^{n,m-1})
\]

\[
= \lambda^n + \alpha_n D(\lambda^n) \nabla \Phi(\lambda^n) + \alpha_n \sum_{m=1}^{M} (D(\lambda^{n,m-1}) \nabla f_m(\lambda^{n,m-1}) - D(\lambda^n) \nabla f_m(\lambda^n))
\]

\[
= \lambda^n + \alpha_n D(\lambda^n) \nabla \Phi(\lambda^n) + O(\alpha_n^2)
\]

(40)

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

$$
\left\| \sum_{m=1}^{M} (D(\lambda^{n,m-1}) \nabla f_m(\lambda^{n,m-1}) - D(\lambda^n) \nabla f_m(\lambda^n)) \right\| \\
\leq \sum_{m=1}^{M} \left\| D(\lambda^{n,m-1}) \nabla f_m(\lambda^{n,m-1}) - D(\lambda^n) \nabla f_m(\lambda^n) \right\| \\
\leq L \sum_{m=1}^{M} |\lambda^{n,m-1} - \lambda^n| \\
\leq \alpha_n L M^2 \max_{m,A \in B} \left\| D(\lambda) \nabla f_m(\lambda) \right\|.
$$

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

$$
\Phi(\lambda^{n+1}) = \Phi(\lambda^n) + \nabla \Phi(\lambda^n)(\lambda^{n+1} - \lambda^n) + O(\|\lambda^{n+1} - \lambda^n\|^2).
$$

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

$$
\Phi(\lambda^{n+1}) = \Phi(\lambda^n) + \alpha_n \nabla \Phi(\lambda^n) D(\lambda^n) \nabla \Phi(\lambda^n) + O(\alpha_n^2).
$$

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

$$
\sum_{n=l}^{k} \alpha_n \nabla \Phi(\lambda^n) D(\lambda^n) \nabla \Phi(\lambda^n) < \epsilon, \forall k > l
$$

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

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

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

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

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

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

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

**Proof:** We extend the proof of Proposition 3 of [17]. It is clear that $\frac{\partial}{\partial \lambda^j} \Phi(\lambda^*) = 0$ if $0 < \lambda^*_j < U$. Considering the optimality conditions [39, p. 203], we need to prove that $\lambda^*_j = 0$ implies $\frac{\partial}{\partial \lambda^j} \Phi(\lambda^*) \leq 0$, and $\lambda^*_j = U$ implies $\frac{\partial}{\partial \lambda^j} \Phi(\lambda^*) \geq 0$.

Define $J = J_1 \cup J_2$ where $J_1 = \{j = 1, \ldots, p : \lambda^*_j = 0\}$ and $J_2 = \{j = 1, \ldots, p : \lambda^*_j = U\}$ and $\frac{\partial}{\partial \lambda^j} \Phi(\lambda^*) < 0$. We show that $J = \emptyset$.

Since $\nabla \Phi$ is continuous on $B$, there exists $0 < \delta < U$ such that if $\lambda \in B_\delta$ then $\frac{\partial}{\partial \lambda^j} \Phi(\lambda) > 0$, $\forall j \in J_1$ and $\frac{\partial}{\partial \lambda^j} \Phi(\lambda) < 0$, $\forall j \in J_2$ where $B_\delta = \{\lambda = \{\lambda - \lambda^*\} < \delta\}$.

Suppose that $\lambda^* \in B_\delta$ where $n$ is sufficiently large. Then, using Lemma 4, we have $\lambda^{n,m} \in B_\delta$, $\forall m$ since $n$ is large. For $j \in J_2$, since $d_j(\lambda^{n,m}) = (U - \lambda^{n,m}_j)/p_j$, $\forall m$, one can show:

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

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

$$
U - \lambda^{n+1}_j = (U - \lambda^n_j) \left(1 - \frac{\alpha_n}{p_j} \frac{\partial}{\partial \lambda^j} f_m(\lambda^{n,m-1})\right) + O(\alpha_n^2).
$$

Now we have $U - \lambda^{n+1}_j > U - \lambda^n_j$, i.e., $\lambda^{n+1}_j < \lambda^n_j$ since $\frac{\partial}{\partial \lambda^j} \Phi(\lambda^n) < 0$. Similarly, one can show that $\lambda^{n+1}_j > \lambda^n_j$ for $j \in J_1$.

Let $\{\lambda^n\}$ be a subsequence of $\{\lambda^n\}$ such that $\lim_{n \to \infty} \lambda^n = \lambda^*$. Let $t_k = \max\{q < n_k : \lambda^q \notin B_\delta\}$. If $\lambda^q \in B_\delta$, $\forall q < n_k$ for some $k$, set $t_k = 0$. Then $\{t_k\}$ is a monotone increasing sequence of nonnegative integers such that $\lambda^q \in B_\delta$ for $k+1 \leq q < n_k$ for large $k$. Suppose that $\lim_{k \to \infty} t_k = t < \infty$, i.e., $\lambda^n$ stays in $B_\delta$ for large $n$. Then $\lambda^k_j > \lambda^* - \delta$, $\forall k > l$, $\forall j \in J_1$ and $\lambda^k_j < \lambda^* - \delta$, $\forall k > l$, $\forall j \in J_2$. Then $\lambda^k_j$ is a subsequence of $\{\lambda^{n,k}\}$, which is a subsequence of $\{\lambda^n\}$, such that $\lim_{k \to \infty} \lambda^k_j = \lambda^*$ with $\lambda^*_j = 0$, $\forall j \in J_1$ and $\lambda^*_j = U$, $\forall j \in J_2$ but $\lambda^* \neq \lambda^*$ since $\lambda^k_j \notin B_\delta$ and thus $\lambda^* \notin B_\delta$. Then $\Phi(\lambda^*) = \Phi(\lambda^*)$ by Lemma 3. We have two different maximizers $\lambda^*$ and $\lambda^*$ of $\Phi$ over $\lambda \in B : \lambda_j = 0$, $\forall j \in J_1$ and $\lambda_j = U$, $\forall j \in J_2$. This is a contradiction since $\Phi$ is strictly concave. So it must be the case that $J = \emptyset$. □
Theorem 1: A sequence \( \{\lambda_n\} \) 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 \( \hat{\lambda} = \arg \max_{\lambda \in \mathbb{B}} \Phi(\lambda) \).

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

Corollary 2: \( \lim_{n \to \infty} \lambda_n^{\alpha} = \hat{\lambda}, \forall m. \)

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: \( \nabla F(\lambda) \) is bounded on \( \mathbb{B} \); and \( f_m \) are concave. They are satisfied by our (modified) Poisson penalized-likelihood. Define a norm \( \| \cdot \|_{D^{-1}} \) on \( \mathbb{R}^p \) by \( \| \lambda \|_{D^{-1}} = \langle \lambda, D^{-1} \lambda \rangle^{1/2} \). For all \( \lambda \in \mathbb{R}^p \), suppose that \( \Phi^{*} = \sup_{\lambda \in \mathbb{B}} \Phi(\lambda) \).

Lemma 6: Let \( \{\lambda_n\} \) be a sequence generated by (25). Then for any \( \lambda \in \mathbb{B} \), one can show:

\[
\| \lambda^{n+1} - \lambda^{*} \|_{D^{-1}}^2 \leq \| \lambda^n - \lambda^{*} \|_{D^{-1}}^2 - 2\alpha_n \langle \Phi(\lambda) - \Phi(\lambda^{*}) \rangle + \alpha_n^2 C
\]

for all \( n \) and some \( C > 0 \).

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

\[
x^{n+1} = g_{\alpha_n, \nabla f_m}(x^{n+1}, \alpha_n, \nabla f_m(x^{n+1}))
\]

for \( m = 1, 2, \ldots, M \), where \( x^{n+1} = D^{-1/2} \lambda^{n+1} \) and \( B = \{ x \in \mathbb{R}^p : 0 \leq x_j \leq U_d \} \). Then use Lemma 2.1 of [3].

Lemma 7: Suppose that \( \{\lambda_n\} \) 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 \( \lim_{n \to \infty} \Phi(\lambda^n) = \Phi^{*} \).

Proof: The proof is due to Proposition 1.2 of [4]. Assume for contradiction that there exists \( \delta > 0 \) such that \( \lim_{n \to \infty} \alpha_n = 0 \) and one can assume that \( N \) is so large that \( \alpha_n C < \delta \) for all \( n > N \). Using Lemma 6, one obtains:

\[
\| \lambda^{n+1} - \lambda^{*} \|_{D^{-1}}^2 \leq \| \lambda^n - \lambda^{*} \|_{D^{-1}}^2 - 2\alpha_n \langle \Phi(\lambda) - \Phi(\lambda^{*}) \rangle + \alpha_n^2 C
\]

for all \( n > N \). Summing up, this gives:

\[
0 \leq \| \lambda^n - \lambda^{*} \|_{D^{-1}}^2 \leq \| \lambda^n - \lambda^{*} \|_{D^{-1}}^2 - \delta \sum_{k=N}^{n-1} \alpha_k
\]

for all \( n > N \). This is a contradiction since \( \sum_{k=0}^{\infty} \alpha_k = \infty \).

Corollary 3: \( \lim_{n \to \infty} \lambda_n^{\alpha} = \lambda^{*} \), \( \forall m. \)

Proof: Use Lemma 4 and Theorem 1.

REFERENCES


[29] A. J. R. Gunawardana, “The information geometry of EM variants for optimal point is $\phi^*$ versus iteration number. For the OS method, each subiterate is denoted.


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 3 subsets. The optimal point is $\phi^* = (0.5, 0.5)$ and the initial point is $x^0 = (5, 5)$. (b) Normalized $\phi$ difference ($\Phi(\hat{\phi}) - \Phi(x^{m+1})$) / ($\Phi(\hat{\phi}) - \Phi(x^0)$) versus iteration number. For the OS method, each subiterate is denoted.
Fig. 2. Comparison of normalized $\Phi$ difference ($\Phi(\lambda^{(\ell)}) - \Phi(\lambda^n))/\Phi(\lambda^{(\ell)})$) versus iteration number for non-OS algorithms including SPS and De Pierro’s modified EM denoted by DPEM; and unrelaxed (i.e., constant stepsizes) OS algorithms including unrelaxed OS-SPS and unrelaxed modified BSREM (with 8 and 40 subsets).

Fig. 3. Comparison of normalized $\Phi$ difference ($\Phi(\lambda^{(\ell)}) - \Phi(\lambda^n))/\Phi(\lambda^{(\ell)})$) 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), $\alpha_\ell = 1/(1.5n + 1)$ and $\alpha_n = 1/(n + 1)$ are used, respectively. This figure shows every subiterate.

Fig. 4. Comparison of normalized $\Phi$ difference ($\Phi(\lambda^{(\ell)}) - \Phi(\lambda^n))/\Phi(\lambda^{(\ell)})$) 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), $\alpha_\ell = 1/(1.5n + 1)$ and $\alpha_n = 1/(n + 1)$ are used, respectively. This figure shows every subiterate.
Fig. 5. Comparison of normalized $\Phi$ difference $((\Phi(\lambda) - \Phi(\lambda^n))/\Phi(\lambda) - \Phi(\lambda^n))$ 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 8 subsets. (d) PL reconstruction using 20 iterations of relaxed OS-SPS with 8 subsets.
TABLE I
ALGORITHM OUTLINE FOR THE ALGORITHMS PRESENTED IN THIS PAPER

Compute a bound $U$ on a solution by (32) in Appendix A.

Compute $e$ 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 $d^P_j$ for relaxed OS-SPS. Use (26) for quadratic penalty.

for each iteration $n = 1, \ldots, n_{iter}$

for each subset $m = 1, \ldots, M$

\[
\hat{l}_i = \sum_{j=1}^p a_{ij} \hat{\lambda}_j + r_i \quad \text{for } i \in S_m
\]

\[
h_i = \begin{cases} 
\hat{h}_i(e) + \hat{h}_i(e)(\hat{l}_i - e) & \text{for } i \in \mathcal{I} \text{ and } \hat{l}_i \leq e_i \text{ where } h_i(l) = y_i \log l - l \\
(y_i/\hat{l}_i) - 1 & \text{otherwise}
\end{cases}
\]

\[
\lambda^{old} = \lambda
\]

for $j = 1, \ldots, p$

\[
\tilde{\phi}_j = \sum_{i \in S_m} a_{ij} h_i - \beta \sum_{k \in \mathcal{N}_j} a_{jk} \psi(\lambda_k^{old} - \lambda_k^{old})/M
\]

Update $\hat{\lambda}_j$. (See Table II.)

end

d_{j}^{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}
\]

\[
\hat{\lambda}_j := \hat{\lambda}_j + \alpha_n d_{j}^{B} \tilde{\phi}_j
\]

end

end

TABLE II
COMPARISON OF ALGORITHMS

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Update in Table I</th>
<th>Sufficient conditions for convergence</th>
</tr>
</thead>
</table>
| Modified BSREM-I   | $d_{j}^{B} = \begin{cases} 
\hat{\lambda}_j/p_j & \text{for } \hat{\lambda}_j < U/2 \\
(U - \hat{\lambda}_j)/p_j & \text{for } \hat{\lambda}_j \geq U/2
\end{cases}$ | (i) $\sum_n \alpha_n = \infty$ \quad (ii) $\sum_n \alpha_n^2 < \infty$ \quad (iii) $\alpha_n$ is sufficiently small \quad (vi) $\hat{\lambda}^{\text{Initial}} \in \text{Int } B$ \quad or, instead of (iii) and (iv), \quad (v) All iterates lie in the interior of $B$ |
| Modified BSREM-II  | $\hat{\lambda}_j := \hat{\lambda}_j + \alpha_n d_{j}^{B} \tilde{\phi}_j$ same as above | $\sum_n \alpha_n = \infty$ and $\sum_n \alpha_n^2 < \infty$ |
| Relaxed OS-SPS     | $\hat{\lambda}_j := \hat{\lambda}_j + \alpha_n d_{j}^{B} \tilde{\phi}_j$ | $\sum_n \alpha_n = \infty$ and $\sum_n \alpha_n^2 < \infty$ |

$t$ is a small value, say, $0.001 \max_j \lambda_{j}^{\text{FIBP}}$. 