Penalized Maximum-Likelihood Image Reconstruction using Space-Alternating Generalized EM Algorithms

Jeffrey A. Fessler and Alfred O. Hero
The University of Michigan, Email: fessler@umich.edu

IEEE T-IP 4(10):1417-29 Oct. 1995

Abstract

Most expectation-maximization (EM) type algorithms for penalized maximum-likelihood image reconstruction converge slowly, particularly when one incorporates additive background effects such as scatter, random coincidences, dark current, or cosmic radiation. In addition, regularizing smoothness penalties (or priors) introduce parameter coupling, rendering intractable the M-steps of most EM-type algorithms. This paper presents space-alternating generalized EM (SAGE) algorithms for image reconstruction, which update the parameters *sequentially* using a sequence of small "hidden" data spaces, rather than *simultaneously* using one large complete-data space. The sequential update decouples the M-step, so the maximization can typically be performed analytically. We introduce new hidden-data spaces that are less informative than the conventional complete-data space for Poisson data and that yield significant improvements in convergence rate. This acceleration is due to statistical considerations, not numerical overrelaxation methods, so monotonic increases in the objective function are guaranteed. We provide a general global convergence proof for SAGE methods with nonnegativity constraints.

I. Introduction

IMAGING techniques with Poisson measurement statistics include: positron emission tomography (PET) [1], single photon emission computed tomography (SPECT), gamma astronomy, microscopy methods [2], and photon-limited optical imaging [3]. Statistical methods for image reconstruction or restoration, such as maximum likelihood (ML), penalized maximum-likelihood (PML), or maximum *a posteriori* (MAP), are computationally challenging due to the transcendental form of the Poisson log-likelihood. EM algorithms [4] have proven to be somewhat useful in such problems, except for two important drawbacks. The first problem is slow convergence, particularly when one includes the additive effects of "background" events such as random coincidences [5], scatter [6], dark-current [7], or background cosmic radiation. The second problem is that the M-step of the EM algorithm becomes intractable when one includes smoothness penalties in the objective function, since these functionals further couple the parameters.

In [8,9] we introduced a class of methods called space-alternating generalized EM (SAGE) algorithms that overcome these limitations of EM algorithms. Rather than using the *simultaneous* update of nearly all EM-type algorithms, SAGE algorithms use *sequential* parameter updates in which one iteratively cycles through a sequence of hidden data spaces—one for each pixel. By choosing hidden data spaces whose Fisher information is smaller than the Fisher information of the ordinary EM complete data space, one can accelerate convergence yet maintain the desirable monotonicity properties of EM algorithms. The relationship between Fisher information and convergence rate [4,8,9,10,11,12,13,14] underscores all of the methods we present. In [9] we described two SAGE algorithm for ML image reconstruction and presented anecdotal results showing that one of them converged faster than the EM algorithm. This paper describes a third SAGE algorithm that *supersedes the previous two* in that it converges faster but negligibly increases CPU time. Using a quadratic penalty for illustration, we show empirically over a range of background event fractions that the new SAGE algorithm converges faster than several EM-type algorithms, even when those methods are accelerated using a new complete data space.

Images reconstructed purely by using the ML criterion [1] are unacceptably noisy. Methods for reducing the noise include: stopping rules [15], penalized least squares [16], separable (non-smoothness) priors [17,18], adding smoothing steps [19], and sieves [20]. Recent studies [21] have found that MAP (or equivalently PML) methods outperform sieves. In this paper, we focus on PML image reconstruction, although the new complete-data and hidden-data spaces we introduce are also applicable to unpenalized ML methods. Algorithms for penalized likelihood objective functions for Poisson statistics can be categorized as: 1) *intrinsically monotonic*, 2) *forced monotonic* (typically made monotonic using a line search), and 3) *nonmonotonic* methods. Since one could convert any nonmonotonic method to a forced monotonic method by using a line search, the latter two categories overlap. Nonmonotonic methods can diverge unless one explicitly checks that the objective increases, which often would be expensive in applications with many parameters. The SAGE methods we propose are intrinsically monotonic, so expensive line searches are unnecessary. Although it is not our purpose to argue this point, we believe that convergence properties are relevant to clinical medical imaging, since algorithm divergence could have unfortunate

This work was supported in part by DOE Grant DE-FG02-87ER60561, NSF grant BCS-9024370 and NIH grants CA-54362-02 and CA-60711-01.

consequences.

Intrinsically monotonic methods are those such as the ML-EM algorithm for PET [1] where the statistical formulation of the recursion inherently ensures that the objective function increases every iteration (ignoring finite precision computing). The only intrinsically monotonic methods for penalized maximum-likelihood that we are aware of are: 1) extensions of the EM algorithm including generalized expectation-maximization (GEM) [22], expectation/conditional maximization (ECM) [23,24], and SAGE [9] algorithms, 2) the trivial case with separable (non-smoothness) priors [17,18], 3) De Pierro's algorithms [25,26], 4) and the ICM-EM algorithm [27]. For comparison purposes, we derive an accelerated monotone converging GEM algorithm in Section IV using a new complete-data space. Most intrinsically monotonic algorithms converge globally to the unique maximum for strictly concave objectives.

Perhaps a more accurate name for nonmonotonic methods would be "not guaranteed monotonic" since most such methods do have *local* convergence and the PML estimate is usually a fixed point. An early approach was gradient ascent of the objective starting from an ML estimate [28,29], which was stated to "not guarantee convergence to the global [max]imum." Gradient ascent is complicated by the nonnegativity constraint. Most other nonmonotonic methods are variations of the one-step late (OSL) method of Green [30,31]. In the OSL approach, one circumvents the problem of coupled equations by "plugging in" values from the previous iteration. Unfortunately, such an approach can diverge, unless modified to include a line search [32]. Similar strategies include the BIP algorithm [33], the methods in [34,35], and nested gradient or Jacobi iterations [36,37,21]. Most such strategies include a user-specified step size parameter, and one user has noted that "finding good values for [the step size] and the number of times to iterate requires painful experimentation [38]." Other OSL-like methods are given in [38,39], which have been reported to occasionally diverge [39]. The sequential update of our SAGE methods is close in form (cf Type-III algorithms in Table 1) to the coordinate-wise ascent proposed by Bouman and Sauer [40,41].

One could force any of the above methods to be monotonic by adding a line search. Lange has shown convergence for a line-search modification of OSL [32], and Mumcuoglu *et al.* have adapted the conjugate gradient method [42]. We show in Section VI that an intrinsically monotonic SAGE algorithm converges faster than even a line-search accelerated EM algorithm.

This paper is condensed from [43], in which we compare SAGE to many alternative algorithms and show that the convergence rate of SAGE is comparable to even fast nonmonotonic methods such as [40,41]. Just as one can force a nonmonotonic algorithm to be monotonic by adding a line search, one can also often accelerate monotonic methods by over-relaxation. Thus, for meaningful comparisons, one should first decide whether or not monotonicity is required. In this paper, we focus solely on *monotonic* (intrinsic or forced) algorithms. Additional comparisons can be found in [43].

The organization of this paper is as follows. Section II describes the general structure of the SAGE method. Section III introduces new complete-data spaces and hidden-data spaces for Poisson data, and gives several algorithms for unpenalized maximum-likelihood. Section IV presents PML algorithms. Sections V and VI illustrate the convergence rates. The Appendix gives a global convergence proof.

II. The SAGE Method

Previously we described the SAGE method within a statistical framework [9,8,12]. Here we first describe a generalized version of the method without direct statistical considerations, and then introduce the statistical version as a special case. This new formulation encompasses both the previous SAGE method [9,8,12] and the convexity approach of De Pierro [26,44] as special cases.

A. Problem

Let the observation $\boldsymbol{Y}$ have the probability distribution $f(\boldsymbol{y};\boldsymbol{\theta}_{\rm true})$, where $\boldsymbol{\theta}_{\rm true}$ is a parameter vector residing in a subset Θ of the p-dimensional space $\mathbb{R}^p$. Given a measurement realization $\boldsymbol{Y}=\boldsymbol{y}$, our goal is to compute the penalized maximum-likelihood estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_{\rm true}$, defined by:

$$\hat{\boldsymbol{\theta}} \stackrel{\triangle}{=} \arg\max_{\boldsymbol{\theta}\in\Theta} \Phi(\boldsymbol{\theta}), \quad \text{where} \quad \Phi(\boldsymbol{\theta}) \stackrel{\triangle}{=} \log f(\boldsymbol{y};\boldsymbol{\theta}) - P(\boldsymbol{\theta}). \quad (1)$$

P is an optional penalty function. When analytical solutions for $\hat{\boldsymbol{\theta}}$ are unavailable, one must resort to iterative methods, most of which update all pixels *simultaneously*. SAGE algorithms use *sequential* updates.

B. Algorithm

To describe the SAGE method, we adopt the notation used in [9]. Define an *index set* S to be a nonempty subset of $\{1,\ldots,p\}$, and $\tilde{S}$ its complement. If the cardinality of S is m, then $\boldsymbol{\theta}_S$ denotes the m dimensional vector consisting of the m elements of $\boldsymbol{\theta}$ indexed by the members of S. Similarly $\boldsymbol{\theta}_{\tilde{S}}$ denotes the $p-m$ dimensional vector consisting of the remaining elements of $\boldsymbol{\theta}$. For example, if $p=5$ and $S=\{1,3,4\}$, then $\tilde{S}=\{2,5\}$, $\boldsymbol{\theta}_S=[\theta_1\ \theta_3\ \theta_4]'$, and $\boldsymbol{\theta}_{\tilde{S}}=[\theta_2\ \theta_5]'$, where $'$ denotes vector transpose. Finally, functions like $\Phi(\boldsymbol{\theta})$ expect a p-dimensional vector argument, but it is often convenient to split the argument $\boldsymbol{\theta}$ into two vectors: $\boldsymbol{\theta}_S$ and $\boldsymbol{\theta}_{\tilde{S}}$, as defined above. Therefore, we equate expressions such as: $\Phi(\boldsymbol{\theta}_S,\boldsymbol{\theta}_{\tilde{S}})=\Phi(\boldsymbol{\theta})$.

Let $\boldsymbol{\theta}^0\in\Theta$ be an initial parameter estimate. Given $\boldsymbol{\theta}^i, i=1,2,\ldots$, a SAGE algorithm produces a new estimate $\boldsymbol{\theta}^{i+1}$ by the following two steps:

E-step: Choose an index set S^i
and a functional $\phi^i(\boldsymbol{\theta}_{S^i};\boldsymbol{\theta}^i)$ satisfying:

$$\Phi(\boldsymbol{\theta}_{S^i},\boldsymbol{\theta}^i_{\tilde{S}^i}) - \Phi(\boldsymbol{\theta}^i) \geq \phi^i(\boldsymbol{\theta}_{S^i};\boldsymbol{\theta}^i) - \phi^i(\boldsymbol{\theta}^i_{S^i};\boldsymbol{\theta}^i). \quad (2)$$

$$\text{M-step:} \quad \boldsymbol{\theta}^{i+1}_{S^i} = \arg\max_{\boldsymbol{\theta}_{S^i}} \phi^i(\boldsymbol{\theta}_{S^i};\boldsymbol{\theta}^i) \quad (3)$$

$$\boldsymbol{\theta}^{i+1}_{\tilde{S}^i} = \boldsymbol{\theta}^i_{\tilde{S}^i} \quad (4)$$

The maximization in (3) and the inequality in (2) are over the set $\{\boldsymbol{\theta}_{S^i} : (\boldsymbol{\theta}_{S^i},\boldsymbol{\theta}^i_{\tilde{S}^i})\in\Theta\}$.

This is an "algorithm" in a loose sense, since there is considerable latitude for the algorithm designer when choosing the index sets $\{S^i\}$ and functionals $\{\phi^i\}$ (see Appendix). The basic idea behind the SAGE method is that if maximizing $\Phi(\boldsymbol{\theta}_{S^i}, \boldsymbol{\theta}^i_{\tilde{S}^i})$ over $\boldsymbol{\theta}_{S^i}$ at the ith iteration is difficult, then we instead maximize some user-specified functional $\phi^i(\boldsymbol{\theta}_{S^i}; \boldsymbol{\theta}^i)$, carefully chosen to ensure (using (2)) that increases in ϕ^i yield increases in Φ. Often one can maximize $\phi^i(\cdot; \boldsymbol{\theta}^i)$ *analytically*, obviating expensive line searches. We discuss choices for the index sets S^i in [9]. Here we focus on single-pixel index sets, e.g.: $S^i = \{1 + (i \text{ modulo } p)\}$.

C. Convergence Properties

It follows from (2) and (3) that the sequence of estimates $\{\boldsymbol{\theta}^i\}$ generated by any SAGE algorithm will *monotonically* increase the objective $\Phi(\boldsymbol{\theta}^i)$. If the objective function is bounded above, then this monotonicity ensures that $\{\Phi(\boldsymbol{\theta}^i)\}$ converges, but it does not guarantee convergence of the sequence $\{\boldsymbol{\theta}^i\}$. In [9], we provided regularity conditions under which the sequence $\{\boldsymbol{\theta}^i\}$ also converges monotonically *in norm*, and derived an expression for the asymptotic rate of convergence. The nonnegativity constraint for image reconstruction violates one of the regularity conditions in [9]. Therefore, in the Appendix we prove global convergence under mild conditions suitable for image reconstruction with nonnegativity constraints.

D. Hidden-Data Spaces

A natural approach to choosing functionals ϕ^i that satisfy (2) is to use the underlying statistical structure of the problem. Often one can simplify the form of the log-likelihood by (conceptually) augmenting the observed data with some additional unobservable or "hidden" data. The hidden-data spaces we defined in [9] were all independent of the iteration i. Here we present a less restrictive definition that allows one to use hidden-data spaces that change with iteration.

Definition 1 *Let S^i denote the index set for the ith iteration. A random vector $\boldsymbol{X}$ with probability distribution $f(\boldsymbol{x}; \boldsymbol{\theta}_{S^i}, \boldsymbol{\theta}^i_{\tilde{S}^i})$ is an* admissible hidden-data space *with respect to $\boldsymbol{\theta}_{S^i}$ for $f(\boldsymbol{y}; \boldsymbol{\theta}_{S^i}, \boldsymbol{\theta}^i_{\tilde{S}^i})$ at $\boldsymbol{\theta}^i$ if the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$ satisfies*

$$f(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta}_{S^i}, \boldsymbol{\theta}^i_{\tilde{S}^i}) = f(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta}^i_{\tilde{S}^i}) f(\boldsymbol{x}; \boldsymbol{\theta}_{S^i}, \boldsymbol{\theta}^i_{\tilde{S}^i}), \qquad (5)$$

i.e., the conditional distribution $f(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta}^i_{\tilde{S}^i})$ must be independent of $\boldsymbol{\theta}_{S^i}$.

Any complete-data space associated with a conventional EM algorithm is a special case of this definition [9].

Given an admissible hidden-data space $\boldsymbol{X}$, define the following conditional expectation:

$$Q(\boldsymbol{\theta}_{S^i}; \boldsymbol{\theta}^i) = E\left\{\log f(\boldsymbol{X}; \boldsymbol{\theta}_{S^i}, \boldsymbol{\theta}^i_{\tilde{S}^i}) | \boldsymbol{Y} = \boldsymbol{y}; \boldsymbol{\theta}^i\right\}. \qquad (6)$$

Combine this conditional expectation with the penalty function:

$$\phi(\boldsymbol{\theta}_{S^i}; \boldsymbol{\theta}^i) \stackrel{\triangle}{=} Q(\boldsymbol{\theta}_{S^i}; \boldsymbol{\theta}^i) - P(\boldsymbol{\theta}_{S^i}, \boldsymbol{\theta}^i_{\tilde{S}^i}). \qquad (7)$$

From [9], any ϕ generated using (5)-(7) satisfies (2). Thus, one can design SAGE algorithms by choosing index sets $\{S^i\}$ and admissible hidden-data spaces $\{\boldsymbol{X}^i\}$, and then generating $\{\phi^i\}$ functionals using (5)-(7). The "majorization" method of De Pierro [26,44] is an alternative method for choosing ϕ^i functionals [43].

III. Maximum Likelihood

In this section we first review the linear Poisson model that is often used in image reconstruction problems, and summarize the classical EM algorithm (ML-EM-1) for maximizing the likelihood [1]. We then introduce a new complete-data space that leads to a new, faster converging EM algorithm: ML-EM-3. Even less informative hidden-data spaces lead to new SAGE algorithms that converge faster than both ML-EM-3 and the line-search accelerated EM algorithm (ML-LINU) [45]. We presented some of this material in [9,12]; we include it here since the concepts behind the new complete-data spaces and hidden-data spaces are easier to explain in the maximum-likelihood framework than in the PML case described in the next section.

A. The Problem

Let the emission distribution be discretized into p pixels with nonnegative emission rates $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_p]' \geq 0$. Let N_{nk} denote the number of emissions from the kth pixel that are detected by the nth of N detectors, assumed to have independent Poisson distributions:

$$N_{nk} \sim \text{Poisson}\{a_{nk}\lambda_k\},$$

where the a_{nk} are nonnegative constants that characterize the system [1] with $a_{\cdot k} = \sum_n a_{nk} > 0$. The detectors record emissions from several source locations as well as background events, so we observe

$$Y_n = \sum_k N_{nk} + R_n \sim \text{Poisson}\{\sum_k a_{nk}\lambda_k + r_n\}, \qquad (8)$$

where $\{R_n\}$ are independent Poisson variates: $R_n \sim \text{Poisson}\{r_n\}$. We assume the background rates $\{r_n\}$ are known. This assumption is not essential to the general method, and one could generalize the approach to jointly estimate [11] $\{\lambda_k\}$ and $\{r_n\}$.

Given realizations $\{y_n\}$ of $\{Y_n\}$, the log-likelihood for this problem is given by [1]:

$$L(\boldsymbol{\lambda}) = \log f(\boldsymbol{y}; \boldsymbol{\lambda}) \equiv \sum_n \left(y_n \log \bar{y}_n(\boldsymbol{\lambda}) - \bar{y}_n(\boldsymbol{\lambda})\right), \qquad (9)$$

where

$$\bar{y}_n(\boldsymbol{\lambda}) = \sum_k a_{nk}\lambda_k + r_n. \qquad (10)$$

(We use the symbol "$\equiv$" between expressions that are equivalent up to constant terms that are independent of $\boldsymbol{\lambda}$.) We would like to compute the ML estimate $\hat{\boldsymbol{\lambda}} \geq 0$ from $\boldsymbol{y} = [y_1, \ldots, y_N]'$.

B. ML-EM Algorithms

The classical EM complete-data space [1] for this problem is the set of unobservable random variates

$$\boldsymbol{X}^1 = \{\{N_{nk}\}_{k=1}^p, \{R_n\}\}_{n=1}^N, \qquad (11)$$

which has the following log-likelihood:

$$\log f(\boldsymbol{X}^1;\boldsymbol{\lambda}) \equiv \sum_k \sum_n \left(N_{nk}\log(a_{nk}\lambda_k) - a_{nk}\lambda_k\right).$$

As shown in [1],

$$E\{N_{nk}|\boldsymbol{Y}=\boldsymbol{y};\boldsymbol{\lambda}^i\} = \lambda_k^i a_{nk} y_n/\bar{y}_n(\boldsymbol{\lambda}^i).$$

Thus, the Q function (6) for $\boldsymbol{X}^1$ is (see eqn. (4) of [1]):

$$\begin{aligned} Q_{\mathbf{X}^1}(\boldsymbol{\lambda};\boldsymbol{\lambda}^i) &= E\{\log f(\boldsymbol{X}^1;\boldsymbol{\lambda})|\boldsymbol{Y}=\boldsymbol{y};\boldsymbol{\lambda}^i\} \\ &\equiv \sum_k \left(\lambda_k^i\, e_k(\boldsymbol{\lambda}^i)\log\lambda_k - a_{\cdot k}\lambda_k\right), \qquad (12) \end{aligned}$$

where

$$e_k(\boldsymbol{\lambda}^i) = \sum_n a_{nk} y_n/\bar{y}_n(\boldsymbol{\lambda}^i). \qquad (13)$$

$Q_{\mathbf{X}^1}(\cdot;\boldsymbol{\lambda}^i)$ is a separable, concave function of $\lambda_1,\ldots,\lambda_p$, and one can maximize it analytically. This yields the ML-EM-1 algorithm [1,5], which is a Type-I algorithm in Table 1 with its M-step (41) given by:

$$\lambda_k^{i+1} = \lambda_k^i e_k(\boldsymbol{\lambda}^i)/a_{\cdot k}. \qquad (14)$$

ML-EM-1 converges globally [1,11] but slowly. The slow convergence is partly explained by considering the Fisher information of the complete-data space $\boldsymbol{X}^1$ [11]. One can think of $\boldsymbol{X}^1$ as data from a hypothetical tomograph that knows whether each detected event is a true emission or a background event, and knows in which pixel each event originated. Such a tomograph would clearly be much more *informative* than real tomographs, and this intuition is reflected in the Fisher information matrices. The Fisher information of the parameter vector $\boldsymbol{\lambda}$ for the observed data $\boldsymbol{Y}$ evaluated at the ML estimate $\hat{\boldsymbol{\lambda}}$ is

$$\mathbf{F}_{\mathbf{Y}}(\hat{\boldsymbol{\lambda}}) = E\{-\nabla^2_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda})\}\Big|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}} = \mathbf{A}'\mathrm{diag}\left\{\mathbf{A}\hat{\boldsymbol{\lambda}}+\boldsymbol{r}\right\}^{-1}\mathbf{A},$$

whereas the Fisher information for $\boldsymbol{X}^1$ is diagonal:

$$\mathbf{F}_{\mathbf{X}^1}(\hat{\boldsymbol{\lambda}}) = \mathrm{diag}\left\{a_{\cdot k}/\hat{\lambda}_k\right\}$$

(provided $\hat{\boldsymbol{\lambda}}$ is positive). One can show that $\mathbf{F}_{\mathbf{X}^1} > \mathbf{F}_{\mathbf{Y}}$ (i.e. $\mathbf{F}_{\mathbf{X}^1} - \mathbf{F}_{\mathbf{Y}}$ is a positive definite matrix) using a Fisher version of the data processing inequality [46]. Indeed, $\mathbf{F}_{\mathbf{X}^1}$ is completely independent of the background rates $\{r_n\}$, reflecting the fact that the parameters are *completely isolated from the uncertainty due to the background events* $\{R_n\}$ *in* $\boldsymbol{X}^1$ (see (11)).

To accelerate convergence, we would like a less informative complete-data space than $\boldsymbol{X}^1$, so we depart from the intuitive relationship between $\boldsymbol{X}^1$ and the underlying image formation physics, and instead exploit the statistical structure of (8). The first approach we tried was the following new complete-data space:

$$\boldsymbol{X}^2 = \{\{X_{nk}\}_{k=1}^p\}_{n=1}^N,$$

where the $\{X_{nk}\}$ are unobservable independent Poisson variates that include *all* of the background events:

$$X_{nk} \sim \mathrm{Poisson}\{a_{nk}(\lambda_k + r_n/a_{n\cdot})\}, \qquad (15)$$

where $a_{n\cdot} = \sum_k a_{nk}$. Clearly $Y_n = \sum_k X_{nk}$ has the appropriate distribution (8). The Fisher information for $\boldsymbol{X}^2$ is diagonal and smaller than that of $\boldsymbol{X}^1$:

$$\mathbf{F}_{\mathbf{X}^2}(\hat{\boldsymbol{\lambda}}) = \mathrm{diag}\left\{\sum_n a_{nk}/(\hat{\lambda}_k + r_n/a_{n\cdot})\right\} < \mathbf{F}_{\mathbf{X}^1}.$$

Unfortunately, the function $Q_{\mathbf{X}^2}$ (formed using (6)) has no analytical maximum (unless the ratio $r_n/a_{n\cdot}$ is a constant independent of n), so the M-step appears intractable. Such tradeoffs between convergence rate and computation per-iteration are common [11].

To obtain an tractable M-step, we would like to replace the term $r_n/a_{n\cdot}$ in (15) with a term that is *independent* of n. Therefore, we propose the following new complete data space [12,43]:

$$\boldsymbol{X}^3 = \{\{M_{nk}\}_{k=1}^p, \{B_n\}\}_{n=1}^N,$$

where $\{M_{nk}\}$ and $\{B_n\}$ are unobservable independent Poisson variates:

$$\begin{aligned} M_{nk} &\sim \mathrm{Poisson}\{a_{nk}(\lambda_k + m_k)\} \\ B_n &\sim \mathrm{Poisson}\{r_n - \sum_k a_{nk} m_k\}, \qquad (16) \end{aligned}$$

and where $\{m_k\}$ are design parameters that must satisfy

$$\sum_k a_{nk} m_k \le r_n,\ \forall n, \qquad (17)$$

so that the Poisson rates of $\{B_n\}$ are nonnegative. With these definitions, clearly

$$Y_n = \sum_k M_{nk} + B_n$$

has the appropriate distribution (8).

The Fisher information for $\boldsymbol{X}^3$ is diagonal:

$$\mathbf{F}_{\mathbf{X}^3}(\hat{\boldsymbol{\lambda}}) = \mathrm{diag}\left\{a_{\cdot k}/(\hat{\lambda}_k + m_k)\right\}, \qquad (18)$$

and now depends on r_n though (19) below. This Fisher information is smaller than $\mathbf{F}_{\mathbf{X}^1}(\hat{\boldsymbol{\lambda}})$, which leads to faster convergence. In light of (18), to make $\mathbf{F}_{\mathbf{X}^3}$ small the design parameters $\{m_k\}$ should be “as large as possible,” but still satisfying the constraint (17). We have found it natural to choose a set $\{m_k\}$ whose *smallest element is as large as possible* subject to (17). A simple solution to this min-max problem is:

$$m_k = \min_{n\,:\,a_{n\cdot}\neq 0}\{r_n/a_{n\cdot}\}. \qquad (19)$$

We discuss alternatives to (19) based on other min-max criteria in [43], none of which we have found to perform significantly better than (19) for PET, but that might be advantageous elsewhere.

The design (19) clearly satisfies (17), and at least one of the N constraints in (17) is met with equality. Thus, the M_{nk} terms absorb some of the background events, but usually not all. For tomographic systems, the $a_{n\cdot}$’s vary by orders of magnitude between rays traversing the center of the object and rays grazing the object's edge, so $\sum_k a_{nk} m_k << r_n$ for most n. Many of the background events remain separated in B_n. In contrast, in image restoration problems, if the point-spread function is roughly spatially invariant and the background rates $\{r_n\}$ are uniform, then the ratios $\{r_n / a_{n\cdot}\}$ will be fairly uniform and nearly all of the background events will be absorbed into $\{M_{nk}\}$.

Using a similar derivation as for (12) one can show:

$$Q_{\mathbf{X}^3}(\boldsymbol{\lambda}; \boldsymbol{\lambda}^i) \equiv \sum_k \left((\lambda_k^i + m_k) e_k(\boldsymbol{\lambda}^i) \log(\lambda_k + m_k) - a_{\cdot k}(\lambda_k + m_k) \right), \tag{20}$$

where e_k was defined by (13). Like $Q_{\mathbf{X}^1}$, this function is also separable, and its partial derivatives are:

$$\frac{\partial}{\partial \lambda_k} Q_{\mathbf{X}^3}(\boldsymbol{\lambda}; \boldsymbol{\lambda}^i) = e_k(\boldsymbol{\lambda}^i) \frac{\lambda_k^i + m_k}{\lambda_k + m_k} - a_{\cdot k}.$$

To implement the M-step, one cannot simply maximize $Q_{\mathbf{X}^3}$ by zeroing its partial derivatives, because of the nonnegativity constraint. However, $Q_{\mathbf{X}^3}$ is a concave function with respect to λ_k, so if its derivative vanishes at a negative λ_k, then the point $\lambda_k = 0$ will satisfy the Karush-Kuhn-Tucker conditions for the nonnegativity constraint. This leads to the ML-EM-3 algorithm, which, like ML-EM-1, is also a Type-I algorithm of Table 1, with (14) replaced by:

$$\lambda_k^{i+1} = \left[(\lambda_k^i + m_k) e_k(\boldsymbol{\lambda}^i) / a_{\cdot k} - m_k \right]_+, \tag{21}$$

where $[x]_+ = x$ if $x > 0$ and is zero otherwise. This simple change to the implementation of ML-EM-1 accelerates convergence, both theoretically and empirically, provided that some $m_k > 0$. Random coincidences are pervasive in PET, so $r_n > 0$ for all n and $m_k > 0$ for all k.

Like ML-EM-1, since ML-EM-3 is an EM algorithm it monotonically increases the likelihood every iteration. Unlike ML-EM-1, the iterates generated by ML-EM-3 can move on and off the boundary of the nonnegative orthant each iteration. This may partly explain the faster convergence of ML-EM-3, since when ML-EM-1 converges to the boundary, it can do so at *sublinear* rates [11].

C. ML Line-Search Algorithms

Kaufman [45] noted that ML-EM-1 is the special case where $\alpha = 1$ of the form:

$$\lambda_k^{i+1} = \left[\lambda_k^i + \alpha \left(\frac{\lambda_k^i}{a_{\cdot k}} \right) \frac{\partial}{\partial \lambda_k} L(\boldsymbol{\lambda}^i) \right]_+. \tag{22}$$

The ML-LINB-1 and ML-LINU-1 algorithms [45] use a line-search to choose an $\alpha^i \geq 1$, which accelerates convergence. For ML-LINB-1, the search over α is *bounded* such that $\boldsymbol{\lambda}^{i+1}$ is positive, whereas ML-LINU-1 allows an *unconstrained* "bent line" search, in which α can be chosen large enough that some pixels would become negative, but are set to zero [45]. Similarly, ML-EM-3 is the special case where $\alpha = 1$ of the form:

$$\lambda_k^{i+1} = \left[\lambda_k^i + \alpha \left(\frac{\lambda_k^i + m_k}{a_{\cdot k}} \right) \frac{\partial}{\partial \lambda_k} L(\boldsymbol{\lambda}^i) \right]_+. \tag{23}$$

In the few PET experiments we tried, "accelerating" ML-EM-3 using a line-search to choose $\alpha^i \geq 1$ only slightly increased the convergence rate.

D. ML-SAGE Algorithms

Since ML-EM-3 is a simultaneous update, the background events in (16) must be shared among all the pixels, so the values for m_k are fairly small. We now derive a class of SAGE algorithms that use sequential updates with $S^i = \{k\}$, where $k = 1 + (i \text{ modulo } p)$. Two algorithms in this class were presented in [9]; here we also present a third algorithm. A subtle advantage of sequential updates is that we can associate nearly all of the background events with whichever pixel is being updated, yielding much less informative hidden-data spaces and thus faster convergence.

Define unobservable independent Poisson variates:

$$\begin{aligned} Z_{nk}^i &\sim \text{Poisson}\{a_{nk}(\lambda_k + z_k^i)\} \\ B_{nk}^i &\sim \text{Poisson}\{r_n - a_{nk} z_k^i + \sum_{j \neq k} a_{nj} \lambda_j^i\}, \end{aligned} \tag{24}$$

where $\{z_k^i\}$ are nonnegative design parameters (discussed in more detail below) that must satisfy

$$a_{nk} z_k^i \leq r_n + \sum_{j \neq k} a_{nj} \lambda_j^i, \ \forall n, \tag{25}$$

so that the Poisson rates of B_{nk}^i are nonnegative. This constraint is much less restrictive than (17). Clearly $Y_n = Z_{nk}^i + B_{nk}^i$ has the appropriate distribution (8) for any k. We let the hidden-data space for λ_k *only* be

$$\boldsymbol{X}_k^i = \{Z_{nk}^i, B_{nk}^i\}_{n=1}^N.$$

Using a similar derivation as for $Q_{\mathbf{X}^3}$, one can show:

$$Q_{\mathbf{X}_k^i}(\lambda_k; \boldsymbol{\lambda}^i) \equiv (\lambda_k^i + z_k^i) \ e_k(\boldsymbol{\lambda}^i) \log(\lambda_k + z_k^i) - a_{\cdot k}(\lambda_k + z_k^i). \tag{26}$$

Maximizing $Q_{\mathbf{X}_k^i}(\cdot; \boldsymbol{\lambda}^i)$ analytically (subject to the nonnegativity constraint), yields the ML-SAGE class of algorithms, which are Type-III algorithms of Table 1, with M-steps (44) given by:

$$\lambda_k^{i+1} = \left[(\lambda_k^i + z_k^i) e_k(\boldsymbol{\lambda}^i) / a_{\cdot k} - z_k^i \right]_+. \tag{27}$$

Type-III algorithms update the parameters *sequentially*, and immediately update the predicted measurements $\bar{y}_n$ within the inner loop, whereas Type-I algorithms wait until all parameters have been updated[2].

The recursion (27) does not completely specify an algorithm until we have chosen suitable z_k^i's satisfying the constraint (25). The Fisher information for $\boldsymbol{X}_k^i$ with respect to λ_k is the scalar value

$$F_{\mathbf{X}_k^i}(\hat{\lambda}_k) = a_{\cdot k} / (\hat{\lambda}_k + z_k^i),$$

[2] Incremental updates like (45) will accumulate numerical error, so must be treated with caution if used repeatedly. Fortunately, the SAGE algorithms converge in a small number of iterations. In those rare occasions that we run SAGE for many iterations, we "reset" the estimated projections $\{\bar{y}_n\}$ using (10) roughly every 20 iterations.

which (cf (18)) again suggests that we would like the z_k^i's to be as large as possible subject to (25).

The obvious choice is $z_k^i = z_k^{(1)} = 0$, which trivially satisfies (25), and we then refer to the recursion (27) as ML-SAGE-1 [9,8,12]. This algorithm is generally ineffective except for well conditioned problems [8], which is unsurprising since when $z_k^i = 0$ the Fisher information for $\boldsymbol{X}_k^i$ is just the kth diagonal entry of $\mathbf{F}_{\mathbf{X}^1}$.

A second choice is based on the following idea: *since we are updating one pixel at a time, we can associate nearly all of the background events with each pixel as it is updated*. This may be unintuitive in terms of the imaging physics, but is completely admissible and sensible from a statistical perspective. The choice

$$z_k^i = z_k^{(2)} = \min_{n: a_{nk} \neq 0} \{r_n / a_{nk}\}, \tag{28}$$

clearly satisfies (25), and when substituted into the recursion (27) we refer to the resulting algorithm as ML-SAGE-2 [9]. We precompute the $z_k^{(2)}$ terms in (28) prior to iterating, so the computation difference between ML-SAGE-1 and -2 is negligible. However, this small change significantly accelerates convergence [9].

When the rates $\{r_n\}$ are small, then the $\{z_k^{(2)}\}$ are also small, and ML-SAGE-2 is little better than ML-SAGE-1. Therefore, we now introduce a new choice for z_k^i that is effective even when the $\{r_n\}$ are small or even zero. When updating a single pixel, we can consider the contributions from all of the other pixels as "pseudo-background" events. This opportunity is indicated by the form of (24), which the reader should contrast with (16). The following choice also satisfies (25):

$$\begin{aligned} z_k^i = z_k^{i,(3)} &= \min_{n: a_{nk} \neq 0} \{(r_n + \sum_{j \neq k} a_{nj} \lambda_j^i) / a_{nk}\} \\ &= \min_{n: a_{nk} \neq 0} \{\bar{y}_n(\boldsymbol{\lambda}^i) / a_{nk}\} - \lambda_k^i. \end{aligned} \tag{29}$$

Clearly $z_k^{i,(3)} > z_k^{(2)}$, which yields faster convergence. Note that $z_k^{i,(3)}$ *changes with iteration*, which is nevertheless admissible as defined by Definition 1. We refer to the recursion (27) with the choice (29) as the ML-SAGE-3 algorithm.

Remarkably, with an efficient implementation[3] the "extra work" suggested by the minimizations (29) adds negligibly to the execution time. Since the ratios $\bar{y}_n(\boldsymbol{\lambda}^i)/a_{nk}$ in (29) are already needed for computing e_k (see (13)), no extra floating point divides are required. Only the comparisons for the minimization are needed, and those add negligible CPU time (at least on our DEC 3000).

The definitions (28) and (29) involve only a single a_{nk} in each denominator, rather than the sum $a_{n\cdot}$ contained in the definition (19) of m_k. Thus, $z_k^{(2)}$ and $z_k^{i,(3)}$ are orders of magnitude larger than m_k, and $F_{\mathbf{X}_k^i}$ is much smaller than the kth diagonal entry of $\mathbf{F}_{\mathbf{X}^3}$, leading to faster convergence.

IV. Penalized Maximum Likelihood

We described the ML algorithms above primarily to introduce the new hidden data spaces. In this section we turn to penalized likelihood objectives. We first present SAGE algorithms based on the hidden-data spaces $\{\boldsymbol{X}_k^i\}$. For fair comparison with alternative methods, we also derive a new version of the GEM algorithm [22] using the new complete-data space $\boldsymbol{X}^3$. We derived modified versions of the parallelizable algorithm of De Pierro [26] and the one-step late algorithm of Green [30] in [43]. As we show in Section V, these modified algorithms based on $\boldsymbol{X}^3$ all converge somewhat faster than their original versions based on $\boldsymbol{X}^1$, but still none converge as fast as SAGE on a serial computer. Nevertheless, they should be useful for some parallel computers, and they allow us to perform the most conservative comparison between SAGE and its alternatives.

We have implemented the SAGE method with non-quadratic penalties [43]. However, to simplify notation, in this paper we focus on a simple quadratic smoothness penalty:

$$P(\boldsymbol{\lambda}) = \beta \frac{1}{2} \sum_k \sum_{j \in \mathcal{N}_k} \frac{1}{2} w_{kj} (\lambda_k - \lambda_j)^2 \tag{30}$$

where $\mathcal{N}_k$ is a neighborhood of the kth pixel and $w_{kj} = w_{jk}$. For the results in Section V, we let $\mathcal{N}_k$ be the 8 pixels adjacent to the kth pixel, and set $w_{kj} = 1$ for horizontal and vertical neighbors and $w_{kj} = 1/\sqrt{2}$ for diagonal neighbors. Combining (30) with the log-likelihood (9) yields the penalized likelihood objective function (1):

$$\Phi(\boldsymbol{\lambda}) = \sum_n (y_n \log \bar{y}_n(\boldsymbol{\lambda}) - \bar{y}_n(\boldsymbol{\lambda})) - \beta P(\boldsymbol{\lambda}).$$

For the penalty given by (30), Φ is strictly concave under mild conditions on $\mathbf{A} = \{\{a_{nk}\}\}$.

A. Penalized SAGE Algorithm

For generality we derive the SAGE algorithm for PML with generic z_k^i arguments: any choice satisfying (25) can be used. Following (7), define

$$\begin{aligned} \phi_k(\lambda_k; \boldsymbol{\lambda}^i) &= Q_{\mathbf{X}_k^i}(\lambda_k; \boldsymbol{\lambda}^i) - P(\lambda_k, \boldsymbol{\lambda}_{-k}^i) \\ &\equiv (\lambda_k^i + z_k^i) e_k(\boldsymbol{\lambda}^i) \log(\lambda_k + z_k^i) - a_{\cdot k}(\lambda_k + z_k^i) \\ &\quad - \beta \sum_{j \in \mathcal{N}_k} w_{kj} \frac{1}{2} (\lambda_k - \lambda_j^i)^2, \end{aligned} \tag{31}$$

where $Q_{\mathbf{X}_k^i}$ was defined in (26), and $\boldsymbol{\lambda}_{-k}^i$ is the vector of length $(p-1)$ obtained by removing the kth element from $\boldsymbol{\lambda}^i$. The M-step (3) requires maximizing $\phi_k(\cdot; \boldsymbol{\lambda}^i)$, which we can do analytically by zeroing its derivative since $\phi_k(\lambda_k; \boldsymbol{\lambda}^i)$ is a strictly concave function of λ_k. The derivative of $\phi_k(\cdot; \boldsymbol{\lambda}^i)$ with respect to λ_k is:

$$\frac{\partial}{\partial \lambda_k} \phi_k(\lambda_k; \boldsymbol{\lambda}^i) = -a_{\cdot k} + e_k(\boldsymbol{\lambda}^i) \frac{\lambda_k^i + z_k^i}{\lambda_k + z_k^i} - \beta \sum_{j \in \mathcal{N}_k} w_{kj} (\lambda_k - \lambda_j^i).$$

[3] The pessimistic results given in [43] were for a very inefficient implementation of (29). In our more recent optimized implementations, the execution times per iteration of ML-SAGE-1,2,3 were indistinguishable.

Note that updating only one parameter obviates coupled equations (cf (34)). Zeroing this derivative yields a quadratic formula:

$$(\lambda_k + z_k^i)^2 + 2B_k(\lambda_k + z_k^i) - e_k(\boldsymbol{\lambda}^i)(\lambda_k^i + z_k^i)/(\beta w_{k\cdot}) = 0,$$

where $w_{k\cdot} = \sum_{j\in\mathcal{N}_k} w_{kj}$ and

$$B_k = [a_{\cdot k} - \beta \sum_{j\in\mathcal{N}_k} w_{kj}(\lambda_j^i + z_k^i)]/(2\beta w_{k\cdot}). \tag{32}$$

Just as in the derivation of (21), the constrained maximum of $\phi_k(\cdot;\boldsymbol{\lambda}^i)$ corresponds to either the positive root of the quadratic, or the value $\lambda_k = 0$, since ϕ_k is strictly concave. This leads to the PML-SAGE class of algorithms, which are of Type-III in Table 1 with M-steps (44) given by:

$$\lambda_k^{i+1} = \left[-B_k + \sqrt{B_k^2 + e_k(\boldsymbol{\lambda}^i)(\lambda_k^i + z_k^i)/(\beta w_{k\cdot})} - z_k^i\right]_+ . \tag{33}$$

In words, we first compute the e_k correction term from the current projection estimate, then update the kth pixel using a quadratic formula that involves both the data and the neighboring pixels, and then immediately update the projection estimate before proceeding to the next pixel. In practice, the actual implementation has two important differences: 1) the pixels are updated in four different raster scan orders rather than using the same order each iteration (cf frequency analysis in [47]), and 2) the quadratic formula is computed using numerically stable formulae [43] [48] (p. 156), rather than the conventional form (33).

We refer to the recursion (33) with the choice (29) for z_k^i as PML-SAGE-3, and define PML-SAGE-1 and -2 analogously. (PML-SAGE-1 is essentially identical to the ICM-EM algorithm of Abdalla and Kay [27].) As described in Section V, PML-SAGE-3 converges fastest. Global convergence of PML-SAGE is established in the Appendix.

B. Modified GEM Algorithm

The GEM algorithm for image reconstruction [22] is an intuitive approach to extending the EM algorithm to the PML case. Rather than using $\boldsymbol{X}^1$ as in [22], here we develop a GEM algorithm using the new complete-data space $\boldsymbol{X}^3$. Following (7), let

$$\phi^3(\boldsymbol{\lambda};\boldsymbol{\lambda}^i) = Q_{\mathbf{X}^3}(\boldsymbol{\lambda};\boldsymbol{\lambda}^i) - P(\boldsymbol{\lambda}),$$

where $Q_{\mathbf{X}^3}$ was defined in (20). The GEM algorithm is similar to the special case of the SAGE algorithm of Section II where $S^i = \{1,\ldots,p\}$ and $\phi^i = \phi^3$ for all i. Thus, the M-step (3) requires us to maximize $\phi^3(\cdot;\boldsymbol{\lambda}^i)$. Unfortunately, its partial derivatives are coupled:

$$\frac{\partial}{\partial\lambda_k}\phi^3(\boldsymbol{\lambda};\boldsymbol{\lambda}^i) =$$

$$-a_{\cdot k} + e_k(\boldsymbol{\lambda}^i)\frac{\lambda_k^i + m_k}{\lambda_k + m_k} - \beta \sum_{j\in\mathcal{N}_k} w_{kj}(\lambda_k - \lambda_j),\ k = 1,\ldots,p. \tag{34}$$

This coupling prohibits analytical maximization. The GEM method [22] abandons maximization in favor of simply increasing $\phi^3(\cdot;\boldsymbol{\lambda}^i)$ using coordinate-ascent. It is easier to increase ϕ^3 than $\Phi(\cdot)$ using coordinate-ascent since we can solve (34) with respect to λ_k (while holding the other parameters fixed) using essentially the same quadratic formula as (33). Extending the derivation in [22] leads to the PML-GEM-3 algorithm, which is a Type-II algorithm of Table 1, with the M-step (42) given by:

$$B_k = [a_{\cdot k} - \beta \sum_{j\in\mathcal{N}_k} w_{kj}(\lambda_j^\star + m_k)]/(2\beta w_{k\cdot}) \tag{35}$$

$$\lambda_k^{i+1} = \left[-B_k + \sqrt{B_k^2 + e_k(\boldsymbol{\lambda}^i)(\lambda_k^i + m_k)/(\beta w_{k\cdot})} - m_k\right]_+ . \tag{36}$$

Here, $\lambda_k^\star$ denotes the *most recent* estimate of λ_k, e.g. $\lambda_j^\star = \lambda_j^{i+1}$ if $j < k$, otherwise $\lambda_j^\star = \lambda_j^i$, i.e., the updates are done "in place". We refer to the GEM algorithm based on $\boldsymbol{X}^1$ (where $m_k = 0\ \forall k$) as PML-GEM-1.

One can easily verify that λ_k^{i+1} given by (36) satisfies the one-dimensional Karush-Kuhn-Tucker conditions with respect to the nonnegativity constraint. Thus PML-GEM-3 monotonically increases the objective Φ. Global convergence of GEM follows from Theorem 3 of [23], provided the objective is strictly concave.

Note that PML-SAGE-2,3 and PML-GEM-3 are somewhat similar, except that PML-SAGE-2,3 use the less informative hidden data space $\boldsymbol{X}_k^i$, and update the projections immediately after each parameter update. Although subtle, these two differences lead to PML-SAGE-2,3 converging significantly faster.

C. Modified One-Step-Late (OSL) Algorithm

An alternative to the above algorithms is preconditioned steepest-ascent. Using a derivation similar to that for (22), Lange [32] has shown that Green's OSL algorithm [30] can be expressed in the form:

$$\lambda_k^{i+1} = \left[\lambda_k^i + \alpha\left(\frac{\lambda_k^i}{a_{\cdot k} + \frac{\partial}{\partial\lambda_k}P(\boldsymbol{\lambda}^i)}\right)\frac{\partial}{\partial\lambda_k}\Phi(\boldsymbol{\lambda}^i)\right]_+ , \tag{37}$$

which is a Type-I algorithm in Table 1. For $\alpha = 1$, this method is not necessarily monotonic, but by choosing α using a line-search, one can ensure monotonicity and global convergence [32], and also accelerate OSL (often significantly, see [43]). We refer to (37) as PML-LINB-1 or PML-LINU-1, depending on whether the line-search for α is bounded or unbounded (cf Section III.C).

V. Simulation Methods

We have evaluated the convergence rates of the algorithms using a 2-D slice of the digital Hoffman brain phantom shown in Fig. 1, with intensity 4 in the gray matter, 1 in the white matter, and 0 in the background, discretized on a 80 by 110 grid with 2mm square pixels. The phantom was forward projected using precomputed factors a_{nk} corresponding to an idealized PET system having 100 angles evenly spaced over 180°, and

70 radial samples with 3mm spacing. Each a_{nk} was precomputed as the area of intersection between the square pixel and a strip of width 6mm. (Since the strip width is wider than the radial spacing, the strips overlap.) The detector response is thus a 6mm rectangular function. Only image pixels within a support ellipse of radii 39 by 54 pixels were reconstructed.

The projections were multiplied by nonuniform attenuation factors corresponding to an ellipse with radii 90 and 100 mm with attenuation coefficient 0.01/mm, surrounded by an elliptical 5mm thick skull with attenuation coefficient 0.015/mm. Nonuniform detector efficiencies were applied using pseudo-random log-normal variates with standard deviation 0.2. The sinogram was globally scaled to a mean sum of 900000 true events. All of the above effects were also incorporated into the a_{nk} factors. Pseudo-random independent Poisson variates were drawn according to (8), and a uniform field of Poisson distributed background events with known mean were added. Three data sets were studied, with 0%, 5%, and 35% background events, representing the range of random coincidences in PET scans. The initial estimate $\boldsymbol{\lambda}^0$ was reconstructed using FBP with a 3rd order Butterworth filter with cutoff 0.6 of Nyquist (10mm resolution). FBP image values below 0.1 were set to 0.1 so that $\boldsymbol{\lambda}^0 > \mathbf{0}$.

VI. Results

We found that the LINU algorithms converged faster than the LINB algorithms only in the 0% background case, so the LINU results are shown only then. We focus on the PML algorithms here; results for the unregularized ML case show the same trends [43].

Figures 3-5 display the objective $\Phi(\boldsymbol{\lambda}^i) - \Phi(\boldsymbol{\lambda}^0)$ and illustrate the following points.

- In all cases, the GEM algorithm and OSL algorithms had indistinguishable convergence rates.
- PML-GEM-3 converged faster than the conventional PML-GEM-1, and the increase in speed grows with the background fraction.
- Even with only 5% random coincidences, PML-SAGE-3 clearly increased faster and reached its asymptote sooner than PML-GEM-3. The advantage for 35% background is even greater.
- For 0% background events, $z_k^{(2)} = 0$, so PML-SAGE-2 is identical to PML-SAGE-1 (and the ICM-EM algorithm of [27]), and converged at the same rate as PML-GEM-1 (which is identical to PML-GEM-3 with 0% background). However, for 0% background PML-SAGE-3 converged faster than all of the other algorithms.
- The results given above in terms of the convergence in the objective function $\Phi(\boldsymbol{\lambda}^i)$ also apply to convergence in L_2 norm [43].

Since PML-SAGE is a monotonic algorithm applied to a strictly concave objective, it is robust to the initial estimate. Figure 2 displays PML-SAGE-2 estimates initialized with a uniform image and a FBP image. The difference images rapidly decrease to values that are invisible on a conventional 8-bit display, so we have amplified the differences by a factor of 4 for display. The effects of the initial estimate are negligible by 15-20 iterations for 5% background, or by 10 iterations with 35% background [43].

As detailed in [43], the SAGE algorithms require about 25% more floating point operations per iteration than ML-EM-1, although this could be eliminated by doubling the memory requirements. Nevertheless, even when $\Phi(\boldsymbol{\lambda}^i)$ is graphed against CPU time [43], the SAGE algorithms still have the fastest convergence among monotonic algorithms.

VII. Discussion

This paper presents algorithms for image reconstruction that converge rapidly, monotonically, globally, and naturally enforce nonnegativity constraints. There are two main principles that lead to the improved convergence rates. The first is to update the pixel estimates sequentially rather than simultaneously. This idea has been used successfully by other authors as well [27,40,41]. The second is to use less informative hidden data spaces formed by "mixing together" some of the emission events with the background events. Either idea *by itself* only slightly improves convergence rates (cf PML-SAGE-1, PML-GEM-3 relative to PML-GEM-1 in Fig. 4), but *in tandem* (e.g. PML-SAGE-2 or PML-SAGE-3) the two principles significantly accelerate convergence.

The monotonicity of SAGE stems in part from the fact that all of the (relevant) measured data is used even though only one pixel is updated at a time. In contrast, another method proposed for accelerating EM is the ordered-subsets EM (OS-EM) algorithm [49], in which all pixels are updated simultaneously but *the measurements are used sequentially*. OS-EM is not guaranteed to be monotonic, and its convergence properties are poorly understood. OS-EM achieves a limited form of "regularization" through stopping rules, whereas SAGE can be used with any convex penalty function, including edge-preserving penalties.

We have attempted a fair comparison between SAGE methods and the alternatives. We presented slightly improved versions of several alternatives (GEM, OSL, De Pierro, etc.) in [43], and experimented with several choices for the design parameters m_k. Nevertheless, it is possible that a better choice for $\{m_k\}$, or even a better complete-data space will be eventually found. Such an extension could be very useful since algorithms such as De Pierro's method are more suitable for fine-grain parallel computers than the SAGE algorithms herein. However, the generic SAGE method (Section II) offers more flexibility than we have used here, and we are currently studying alternatives that may be more suitable to parallel computing [43].

We have compared several algorithms, and the reader may wonder what is the impact of these results on "practitioners" of penalized likelihood image reconstruction? Based on our experience, PML-SAGE-3 is the fastest intrinsically monotonic algorithm for Poisson measurements that we are aware of, and it converges faster even than PML-LINU, which is the fastest forced monotonic method we are aware of. Therefore we recommend using PML-SAGE-3 when using conventional serial computers. However, given the considerable recent progress in accelerating algorithms for image reconstruction, it is doubtful that SAGE will be the final word. It is noteworthy that the statistical principles behind the SAGE methods yield convergence rates that rival conventional numerical tools such as line-searches and Newton's methods, yet ensuring algorithm monotonicity. Further development using statistical perspectives will likely lead to additional improvements.

VIII. Appendix: Convergence

The proof in [9] of local monotonic convergence in norm to a fixed point is inapplicable to problems with nonnegativity constraints, except when the fixed point lies in the interior of the nonnegative orthant. In this appendix, we prove convergence of a general form of SAGE that allows the limit to lie on the boundary of the nonnegative orthant. The proof structure is based on [1] and [26].

We begin by stating some general sufficient conditions for convergence. These conditions make no specific references to the Poisson likelihood or penalty used in this paper, so this proof will apply to a broad class of nonnegatively constrained estimation problems. Following the general proof, we verify that the specific SAGE algorithms presented in this paper meet the required conditions under the linear Poisson model.

Define the following sets:

$$\begin{aligned}\Re_S^+ &= \{\boldsymbol{\theta}_S : \theta_k \geq 0,\ k \in S\},\\ \Theta^+ &= \{\boldsymbol{\theta} \in \Re^p : \theta_k \geq 0,\ k = 1, \ldots, p\},\\ \mathcal{S}(\boldsymbol{\theta}^0) &= \{\boldsymbol{\theta}\ :\ \Phi(\boldsymbol{\theta}) \geq \Phi(\boldsymbol{\theta}^0)\}.\end{aligned}$$

Also define:

$$\nabla_k^{10} \phi^i(\boldsymbol{\theta}_{S^i}^\star; \bar{\boldsymbol{\theta}}) \stackrel{\triangle}{=} \left.\frac{\partial}{\partial \theta_k} \phi^i(\boldsymbol{\theta}_{S^i}; \bar{\boldsymbol{\theta}})\right|_{\boldsymbol{\theta}_{S^i} = \boldsymbol{\theta}_{S^i}^\star}$$

and

$$\mathbf{J}^i(\boldsymbol{\theta}_{S^i}; \bar{\boldsymbol{\theta}}) \stackrel{\triangle}{=} -\frac{1}{2} \nabla^{20} \phi^i(\boldsymbol{\theta}_{S^i}; \bar{\boldsymbol{\theta}}),$$

where for $k, j \in S^i$

$$\left[\nabla^{20} \phi^i(\boldsymbol{\theta}_{S^i}; \bar{\boldsymbol{\theta}})\right]_{kj} = \frac{\partial^2}{\partial \theta_k \partial \theta_j} \phi^i(\boldsymbol{\theta}_{S^i}; \bar{\boldsymbol{\theta}}).$$

To eliminate the interior restriction used in [9], we impose the following two regularity conditions on Φ.

Assumption 1 *$\Phi(\boldsymbol{\theta})$ is strictly concave (and continuous and differentiable) on Θ^+.*

Assumption 2 *For any $\boldsymbol{\theta}^0 \in \Theta^+$, the set $\mathcal{S}(\boldsymbol{\theta}^0)$ is bounded.*

As noted in [1], the assumption of strict concavity is adequate to "make up for" relaxing the restriction to the interior of Θ^+. We do not consider strict concavity to be an overly restrictive assumption; if Φ is not strictly concave, then typically either it does not have a unique maximum, in which case it is a questionable choice of objective, or it has local maxima, and no known deterministic algorithms are guaranteed to find the global maxima, including SAGE. Like any monotonic algorithm, for a non-strictly concave objective SAGE will only find a global maximum if initialized suitably close to one.

We assume the iterates are produced by an algorithm having the general form given in Section II, i.e., each iteration is associated with an index set S^i and a functional $\phi^i(\boldsymbol{\theta}_{S^i}; \boldsymbol{\theta}^i)$, and the iterates satisfy $\boldsymbol{\theta}_{\tilde{S}^i}^{i+1} = \boldsymbol{\theta}_{\tilde{S}^i}^i$. We assume that the functionals ϕ^i satisfy the following conditions.

Condition 1 *The functionals ϕ^i satisfy (2), i.e.:*

$$\Phi(\boldsymbol{\theta}_{S^i}, \boldsymbol{\theta}_{\tilde{S}^i}^i) - \Phi(\boldsymbol{\theta}^i) \geq \phi^i(\boldsymbol{\theta}_{S^i}; \boldsymbol{\theta}^i) - \phi^i(\boldsymbol{\theta}_{S^i}^i; \boldsymbol{\theta}^i),$$

for $\boldsymbol{\theta}_{S^i} \in \Re_{S^i}^+$ and $\boldsymbol{\theta}^i \in \Theta^+$.

Condition 2 *Each functional $\phi^i(\cdot; \boldsymbol{\theta})$ is strictly concave and twice differentiable on $\Re_{S^i}^+$ for any $\boldsymbol{\theta} \in \Theta^+$, and each $\phi^i(\cdot; \cdot)$ is continuous on $\Re_{S^i}^+ \times \Theta^+$.*

Condition 3 *The following derivatives match $\forall i$:*

$$\frac{\partial}{\partial \theta_k} \Phi(\boldsymbol{\theta}) = \nabla_k^{10} \phi^i(\boldsymbol{\theta}_{S^i}; \boldsymbol{\theta})$$

for any $\boldsymbol{\theta} \in \Theta^+$ and $k \in S^i$.

Condition 4 *For $\boldsymbol{\theta}^i \in \Theta^+$, the iterates satisfy the Karush-Kuhn-Tucker conditions $\forall k \in S^i$:*

$$\nabla_k^{10} \phi^i(\boldsymbol{\theta}_{S^i}^{i+1}; \boldsymbol{\theta}^i)\ \begin{cases} = 0, & \theta_k^{i+1} > 0 \\ \leq 0, & \theta_k^{i+1} = 0 \end{cases}.$$

Condition 5 *For any bounded set $\mathcal{S}$, there exists a $C_{\mathcal{S}} > 0$ such that for every i, for all $\bar{\boldsymbol{\theta}} \in \mathcal{S}$, and for all $(\boldsymbol{\theta}_{S^i}, \bar{\boldsymbol{\theta}}_{\tilde{S}}) \in \mathcal{S}$:*

$$\lambda_{\min}\left\{\mathbf{J}^i(\boldsymbol{\theta}_{S^i}; \bar{\boldsymbol{\theta}})\right\} \geq C_{\mathcal{S}},$$

where $\lambda_{\min}\{\mathbf{J}\}$ denotes the minimum eigenvalue of $\mathbf{J}$.

Condition 6 *For each $k \in \{1, \ldots, p\}$, there is an index set $S^{(k)}$ containing k and functional $\phi^{(k)}$ that is used regularly to update the kth element of the parameter $\boldsymbol{\theta}$. Define $\mathcal{I}_\| = \{\rangle\ :\ \mathcal{S}^\rangle = \mathcal{S}^{(\|)}$ and $\phi^\rangle = \phi^{(\|)}\}$. Then for each k there exists an integer $i_{\max}$ (which may depend on k) such that*

$$\forall n \geq 0\ \exists i \in [n, n + i_{\max}]\ \text{s.t. } i \in \mathcal{I}_\|.$$

(This condition is clearly satisfied if the index sets and functionals are chosen periodically.)

Using the above conditions, we now prove a series of Lemmas that establish global convergence.

Lemma 1 *The iterates $\{\boldsymbol{\theta}^i\}$ yield monotonic increases in $\Phi(\boldsymbol{\theta}^i)$, and are thus contained in the set $\mathcal{S}(\boldsymbol{\theta}^0)$. Furthermore, $\mathcal{S}(\boldsymbol{\theta}^0)$ is compact and convex.*

Proof: Monotonicity follows from Conditions 1 and 4. Since Φ is strictly concave (Assumption 1), $\mathcal{S}(\boldsymbol{\theta}^0)$ is strictly convex. Since Φ is continuous (Assumption 1), $\mathcal{S}(\boldsymbol{\theta}^0)$ is closed (p. 91 of [50]). Thus $\mathcal{S}(\boldsymbol{\theta}^0)$ is compact since it is closed and bounded (Assumption 2), by the Heine-Borel theorem (p. 58 of [50]). □

Lemma 2 *There exists a $C > 0$ such that for any i*

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

Proof: From Condition 1 and since $\boldsymbol{\theta}^{i+1}_{\tilde{S}^i} = \boldsymbol{\theta}^{i}_{\tilde{S}^i}$, it suffices to show $\forall i$:

$$\|\boldsymbol{\theta}^{i+1}_{S^i} - \boldsymbol{\theta}^{i}_{S^i}\|^2 \leq C^{-1}(\phi^i(\boldsymbol{\theta}^{i+1}_{S^i}; \boldsymbol{\theta}^i) - \phi^i(\boldsymbol{\theta}^{i}_{S^i}; \boldsymbol{\theta}^i)).$$

Expand $\phi^i(\cdot; \boldsymbol{\theta}^i)$ about $\boldsymbol{\theta}^{i+1}_{S^i}$ using Taylor's expansion with remainder (see p. 599 of [51]):

$$\phi^i(\boldsymbol{\theta}_{S^i}; \boldsymbol{\theta}^i) = \phi^i(\boldsymbol{\theta}^{i+1}_{S^i}; \boldsymbol{\theta}^i) +$$

$$\nabla^{10}\phi^i(\boldsymbol{\theta}^{i+1}_{S^i}; \boldsymbol{\theta}^i)(\boldsymbol{\theta}_{S^i} - \boldsymbol{\theta}^{i+1}_{S^i}) + (\boldsymbol{\theta}_{S^i} - \boldsymbol{\theta}^{i+1}_{S^i})'$$

$$\int_0^1 (1-t)\mathbf{J}^i((1-t)\boldsymbol{\theta}^{i+1}_{S^i} + t\boldsymbol{\theta}_{S^i}; \boldsymbol{\theta}^i)\, dt\, (\boldsymbol{\theta}_{S^i} - \boldsymbol{\theta}^{i+1}_{S^i}). \quad (38)$$

From Condition 4, it follows that

$$\nabla^{10}\phi^i(\boldsymbol{\theta}^{i+1}_{S^i}; \boldsymbol{\theta}^i)(\boldsymbol{\theta}^{i+1}_{S^i} - \boldsymbol{\theta}^{i}_{S^i}) \geq 0,$$

so setting in $\boldsymbol{\theta}_{S^i} = \boldsymbol{\theta}^i_{S^i}$ in (38) and applying Condition 5 yields

$$C\|\boldsymbol{\theta}^{i+1}_{S^i} - \boldsymbol{\theta}^{i}_{S^i}\|^2 \leq \phi^i(\boldsymbol{\theta}^{i+1}_{S^i}; \boldsymbol{\theta}^i) - \phi^i(\boldsymbol{\theta}^{i}_{S^i}; \boldsymbol{\theta}^i),$$

where $C = C_{\mathcal{S}(\boldsymbol{\theta}^0)}$. We have used the fact that $\boldsymbol{x}'\mathbf{A}\boldsymbol{x} \geq \|\boldsymbol{x}\|^2\lambda_{\min}\{\mathbf{A}\}$ for any positive definite matrix $\mathbf{A}$. □

Lemma 3 $\|\boldsymbol{\theta}^{i+1} - \boldsymbol{\theta}^i\| \to 0$ *as* $i \to \infty$.

Proof: Since $\{\Phi(\boldsymbol{\theta}^i)\}$ is monotone increasing (Lemma 1) and bounded above (by continuity of Φ and compactness (Lemma 1) of $\mathcal{S}(\boldsymbol{\theta}^0)$, see p. 78 of [50]), it follows that $\Phi(\boldsymbol{\theta}^{i+1}) - \Phi(\boldsymbol{\theta}^i) \to 0$. Apply Lemma 2. □

Lemma 4 *The sequence* $\{\boldsymbol{\theta}^i\}$ *has a limit point*[4] $\boldsymbol{\theta}^\star$. *For any such limit point, if* $\theta^\star_k > 0$, *then* $\frac{\partial}{\partial\theta_k}\Phi(\boldsymbol{\theta}^\star) = 0$.

Proof: By Lemma 1 and [50] (p. 56) there is a subsequence i_m and limit point $\boldsymbol{\theta}^\star \in \mathcal{S}(\boldsymbol{\theta}^0)$ such that $\|\boldsymbol{\theta}^{i_m} - \boldsymbol{\theta}^\star\|^2 \to 0$ as $m \to \infty$. Now pick any index k, and define k_m to be the smallest $i \geq i_m$ such that $i \in \mathcal{I}_{\|}$. By Condition 6, $k_m \leq i_m + i_{\max}$. By the triangle inequality:

$$\|\boldsymbol{\theta}^{k_m} - \boldsymbol{\theta}^\star\|^2 \leq \|\boldsymbol{\theta}^{k_m} - \boldsymbol{\theta}^{i_m}\|^2 + \|\boldsymbol{\theta}^{i_m} - \boldsymbol{\theta}^\star\|^2;$$

the second term of which goes to 0 as $m \to \infty$. For the first term, applying the triangle inequality repeatedly:

$$\|\boldsymbol{\theta}^{k_m} - \boldsymbol{\theta}^{i_m}\|^2 \leq \sum_{i=k_m}^{i_m-1} \|\boldsymbol{\theta}^{i+1} - \boldsymbol{\theta}^i\|^2,$$

which is a sum of at most $i_{\max}$ terms by Condition 6, each of which goes to 0 as $m \to \infty$ by Lemma 3. Thus $\|\boldsymbol{\theta}^{k_m} - \boldsymbol{\theta}^\star\| \to 0$ as $m \to \infty$. Again using the triangle inequality:

$$\|\boldsymbol{\theta}^{k_m+1} - \boldsymbol{\theta}^\star\|^2 \leq \|\boldsymbol{\theta}^{k_m+1} - \boldsymbol{\theta}^{k_m}\|^2 + \|\boldsymbol{\theta}^{k_m} - \boldsymbol{\theta}^\star\|^2.$$

Thus $\|\boldsymbol{\theta}^{k_m+1} - \boldsymbol{\theta}^\star\| \to 0$ as $m \to \infty$.

Since $k_m \in S^{(k)}$, i.e. on iterations $\{k_m\}$ one updates θ_k, by Condition 4: $\theta^{k_m+1}_k \cdot \nabla^{10}_k\phi^{(k)}(\boldsymbol{\theta}^{k_m}_{S^{(k)}}; \boldsymbol{\theta}^{k_m}) = 0$. Taking the limit as $m \to \infty$ and using continuity (Condition 2) shows: $\theta^\star_k \cdot \nabla^{10}_k\phi^{(k)}(\boldsymbol{\theta}^\star_{S^{(k)}}; \boldsymbol{\theta}^\star) = 0$. The Lemma then follows from Condition 3. □

Lemma 5 *The sequence* $\{\boldsymbol{\theta}^i\}$ *converges to a limit* $\boldsymbol{\theta}^\infty$.

Proof: As in Lemma 3 of [1], the number of limit points is finite (at most 2^p), due to Assumption 1, the nonnegativity constraint, and Lemma 4. However, since a bounded (Assumption 2) sequence $\{\boldsymbol{\theta}^i\}$ for which $\|\boldsymbol{\theta}^{i+1} - \boldsymbol{\theta}^i\| \to 0$ (Lemma 3) has a connected and compact set of limit points (see p. 173 of [52]), there must be only one limit point. □

Lemma 6 *The limit* $\boldsymbol{\theta}^\infty$ *satisfies the Karush-Kuhn-Tucker conditions for* Φ.

Proof: For an element $\theta^\infty_k > 0$, we have $\partial/\partial\theta_k\Phi(\boldsymbol{\theta}^\infty) = 0$ by Lemma 4. Now suppose for some k we have $\theta_k = 0$ but $\partial/\partial\theta_k\Phi(\boldsymbol{\theta}^\infty) > 0$. Then by continuity (Assumption 1) and Lemma 3, $\partial/\partial\theta_k\Phi(\boldsymbol{\theta}^i) > 0$ for all i sufficiently large. Thus by Conditions 3 and 6,

$$\nabla^{10}_k\phi^{(k)}(\boldsymbol{\theta}^i_{S^{(k)}}; \boldsymbol{\theta}^i) > 0$$

for all $i \in \mathcal{I}_{\|}$ sufficiently large. But since $\phi^{(k)}(\cdot; \boldsymbol{\theta}^i)$ is strictly concave (Condition 2), if $\nabla^{10}_k\phi^{(k)}(\boldsymbol{\theta}^i_{S^{(k)}}; \boldsymbol{\theta}^i) > 0$, then $\theta^{i+1}_k > \theta^i_k$. This contradicts $\theta^i_k \to 0$, so if $\theta^\infty_k = 0$ we must have $\partial/\partial\theta_k\Phi(\boldsymbol{\theta}^\infty) \leq 0$, establishing the Karush-Kuhn-Tucker conditions. □

Since a strictly concave objective has only one point that satisfies the Karush-Kuhn-Tucker conditions, namely the constrained maximum, the limit $\boldsymbol{\theta}^\infty$ must be that point. Lemma 6 thus establishes global convergence under a generic set of assumptions and conditions. All that remains is to verify that the conditions are satisfied for the SAGE algorithms presented in this paper.

Remark:

In all of SAGE algorithms in this paper, the ϕ^i functionals are *additively separable* in their first argument, which means that the curvature matrices $\mathbf{J}^i(\boldsymbol{\theta}_{S^i}; \boldsymbol{\theta}^i)$ are diagonal. In this case, Condition 5 reduces to verifying that the diagonal elements of $\mathbf{J}^i$ have a positive lower bound. This is clearly the case for convex penalties such as the quadratic penalty (30). In other words, for separable ϕ^i functionals, a sufficient condition for Condition 5 is:

Condition 5$'$: *For any bounded set* $\mathcal{S}$, *there exists a* $C_{\mathcal{S}} > 0$ *such that for all* $\boldsymbol{\theta} \in \mathcal{S}$

$$-\frac{1}{2}\frac{\partial}{\partial\theta_k^2}P(\boldsymbol{\theta}) \geq C_{\mathcal{S}}.$$

Theorem 1 *A sequence* $\{\boldsymbol{\theta}^i\}$ *generated by any of the PML-SAGE algorithms for penalized maximum-likelihood image reconstruction converges globally to the unique maximum of a strictly concave objective function* Φ *having a penalty function satisfying Condition 5$'$, provided* $z^i_k > 0\ \forall k$.

Proof:

- Assumption 2 follows from the behavior of the Poisson log-likelihood as $\lambda_k \to \infty$ [1].
- Condition 1 follows from Theorem 1 of [9].

[4] The reader should note the distinction between limits and limit points (or cluster points) (p. 55 of [50]).

- Condition 2 is easily verified for the hidden-data spaces and penalty functions used in this paper.
- Condition 3 follows by the construction of ϕ_k using (5)-(7).
- Condition 4 is built into the definition (3), and is satisfied by (33).
- Condition 5 follows from Condition 5' since the SAGE algorithms have separable ϕ^i functionals.
- Condition 6 is inherently satisfied by the cyclical sequential update used in PML-SAGE.

□

If one hopes for global convergence, then Condition $5'$ is a reasonable restriction; it is clearly satisfied for the quadratic penalty (30), and for most strictly convex penalties.

There is an important technical difference between our proof and the assumptions in [1]. In [1] it was assumed that the sequence was initialized in the interior of Θ^+, and remained in the interior of Θ^+ for every iteration. With our new complete-data spaces and hidden-data spaces, the iterates can come and go from boundary of Θ^+ since the terms z_k^i are nonzero. However, when z_k^i is positive, one can verify that the corresponding functions ϕ_k are well-defined and differentiable on an open interval containing zero.

Condition 2 as stated is only met if $z_k^i > 0$ for all k, which will be true if $r_n > 0$ for all n. If one were to include the effects of say, cosmic radiation, then in practice it is always the case that $r_n > 0$. However, if some r_n, and hence some z_k^i are zero, it is simple to modify the proof to establish global convergence to the maximum. There is one important technical detail however; one cannot use $z_k^i > 0$ in one iteration and then switch to $z_k^i = 0$ in a later iteration, since then λ_k^i could get stuck on the boundary of Θ^+. Provided that one consistently uses either *only* the original complete-data space or *only* the new complete-data spaces, then global convergence is assured.

As stated above, the proof does not always apply to the unpenalized maximum-likelihood algorithms ML-EM-1, ML-EM-3, and ML-SAGE-1,2,3 because the curvature assumption Condition 5 is not necessarily satisfied without a strictly convex penalty. However, one can replace Condition 5 with an alternative condition that each $\phi^i(\boldsymbol{\theta}_{S^i}; \boldsymbol{\theta}^i)$ must be a monotonically decreasing function of θ_k. This approach was used in [1,11]. With this condition, a small modification of the above proof establishes global convergence of the unpenalized algorithms, *provided that Assumption 1 is still satisfied*. This strict concavity will not be satisfied if the system matrix $\mathbf{A}$ does not have full column rank. We consider this to be a minor point since in the underdetermined case regularization is particularly essential, and the above proof shows that PML-SAGE converges globally for strictly concave penalized maximum-likelihood objectives. We conjecture that the methods of [53] could be extended to establish convergence of ML-EM-3, ML-SAGE-1,2,3, etc. without the strict concavity assumption, but such a proof would probably be of limited academic interest since one rarely iterates a ML algorithm to convergence in the unregularized, underdetermined case.

If one is willing to be content with a local convergence result, then it is possible to relax the assumption of strict concavity for the ϕ^i functionals, using a region of convergence idea similar to that in [9,13].

IX. Acknowledgement

The authors gratefully acknowledge Neal Clinthorne and W. Leslie Rogers for helpful discussions, Stephanie Sandor for bringing the `mmap()` function to their attention, and Ed Hoffman for providing the digital brain data.

References

[1] K Lange and R Carson, "EM reconstruction algorithms for emission and transmission tomography", *J. Comp. Assisted Tomo.*, vol. 8, no. 2, pp. 306–316, Apr. 1984.

[2] S Joshi and M I Miller, "Maximum a posteriori estimation with good's roughness for three-dimensional optical-sectioning microscopy", *J. Opt. Soc. Amer. Ser. A*, vol. 10, no. 5, pp. 1078–1085, May 1993.

[3] D L Snyder, A M Hammoud, and R L White, "Image recovery from data acquired with a charge-couple-device camera", *J. Opt. Soc. Amer. Ser. A*, vol. 10, no. 5, pp. 1014–1023, May 1993.

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

[5] D G Politte and D L Snyder, "Corrections for accidental coincidences and attenuation in maximum-likelihood image reconstruction for positron-emission tomography", *IEEE Trans. Med. Im.*, vol. 10, no. 1, pp. 82–89, Mar. 1991.

[6] M E Daube-Witherspoon, R E Carson, Y Yan, and T K Yap, "Scatter correction in maximum likelihood reconstruction of PET data", in *Conf. Rec. of the IEEE Nuc. Sci. Symp. Med. Im. Conf.*, 1992.

[7] R Molina and B D Ripley, "Deconvolution in optical astronomy. A Bayesian approach", in *Stochastic Models, Statistical Methods, and Algorithms in Im. Analysis*, P Barone, A Frigessi, and M Piccioni, Eds., vol. 74 of *Lecture Notes in Statistics*, pp. 233–239. Springer, New York, 1992.

[8] J A Fessler and A O Hero, "Complete-data spaces and generalized EM algorithms", in *Proc. IEEE Conf. Acoust. Speech Sig. Proc.*, 1993, vol. 4, pp. 1–4.

[9] J A Fessler and A O Hero, "Space-alternating generalized expectation-maximization algorithm", *IEEE Trans. Sig. Proc.*, vol. 42, no. 10, pp. 2664–2677, Oct. 1994.

[10] J A Fessler, “Hidden data spaces for maximum-likelihood PET reconstruction”, in *Conf. Rec. of the IEEE Nuc. Sci. Symp. Med. Im. Conf.*, 1992, vol. 2, pp. 898–900.

[11] J A Fessler, N H Clinthorne, and W Leslie Rogers, “On complete data spaces for PET reconstruction algorithms”, *IEEE Trans. Nuc. Sci.*, vol. 40, no. 4, pp. 1055–1061, Aug. 1993.

[12] J A Fessler and A O Hero, “New complete-data spaces and faster algorithms for penalized-likelihood emission tomography”, in *Conf. Rec. of the IEEE Nuc. Sci. Symp. Med. Im. Conf.*, 1993, pp. 1897–1901.

[13] A O Hero and J A Fessler, “Asymptotic convergence properties of EM-type algorithms”, Tech. Rep. 282, Comm. and Signal Proc. Lab., Dept. of EECS, Univ. of Michigan, Ann Arbor, MI, 48109, Apr. 1993.

[14] A O Hero and J A Fessler, “Convergence in norm for alternating expectation-maximization (EM)-type algorithms”, *Statistica Sinica*, vol. 5, no. 1, pp. 0–0, Jan. 1995, To appear.

[15] E Veklerov and J Llacer, “Stopping rule for the MLE algorithm based on statistical hypothesis testing”, *IEEE Trans. Med. Im.*, vol. 6, no. 4, pp. 313–319, Dec. 1987.

[16] J A Fessler, “Penalized weighted least-squares image reconstruction for positron emission tomography”, *IEEE Trans. Med. Im.*, vol. 13, no. 2, pp. 290–300, June 1994.

[17] K Lange, M Bahn, and R Little, “A theoretical study of some maximum likelihood algorithms for emission and transmission tomography”, *IEEE Trans. Med. Im.*, vol. 6, no. 2, pp. 106–114, June 1987.

[18] E M Levitan and G T Herman, “A maximum *a posteriori* probability expectation maximization algorithm for image reconstruction in emission tomography”, *IEEE Trans. Med. Im.*, vol. 6, no. 3, pp. 185–192, Sept. 1987.

[19] B W Silverman, M C Jones, J D Wilson, and D W Nychka, “A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography”, *J. Royal Stat. Soc. Ser. B*, vol. 52, no. 2, pp. 271–324, 1990.

[20] D L Snyder, M I Miller, L J Thomas, and D G Politte, “Noise and edge artifacts in maximum-likelihood reconstructions for emission tomography”, *IEEE Trans. Med. Im.*, vol. 6, no. 3, pp. 228–238, Sept. 1987.

[21] C S Butler and M I Miller, “Maximum A Posteriori estimation for SPECT using regularization techniques on massively parallel computers”, *IEEE Trans. Med. Im.*, vol. 12, no. 1, pp. 84–89, Mar. 1993.

[22] T Hebert and R Leahy, “A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors”, *IEEE Trans. Med. Im.*, vol. 8, no. 2, pp. 194–202, June 1989.

[23] X L Meng and D B Rubin, “Maximum likelihood estimation via the ECM algorithm: A general framework”, *Biometrika*, vol. 80, no. 2, pp. 267–278, 1993.

[24] C H Liu and D B Rubin, “The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence”, *Biometrika*, vol. 81, no. 4, 1994, To appear.

[25] A R De Pierro, “A generalization of the EM algorithm for maximum likelihood estimates from incomplete data”, Tech. Rep. MIPG119, Med. Im. Proc. Group, Dept. of Radiol., Univ. of Pennsylvania, 1987.

[26] A R De Pierro, “A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography”, 1994, To appear in IEEE Trans. Med. Im.

[27] M Abdalla and J W Kay, “Edge-preserving image restoration”, in *Stochastic Models, Statistical Methods, and Algorithms in Im. Analysis*, P Barone, A Frigessi, and M Piccioni, Eds., vol. 74 of *Lecture Notes in Statistics*, pp. 1–13. Springer, New York, 1992.

[28] S Geman and D E McClure, “Bayesian image analysis: an application to single photon emission tomography”, in *Proc. of Stat. Comp. Sect. of Amer. Stat. Assoc.*, 1985, pp. 12–18.

[29] S Geman and D E McClure, “Statistical methods for tomographic image reconstruction”, *Proc. 46 Sect. ISI, Bull. ISI*, vol. 52, pp. 5–21, 1987.

[30] P J Green, “Bayesian reconstructions from emission tomography data using a modified EM algorithm”, *IEEE Trans. Med. Im.*, vol. 9, no. 1, pp. 84–93, Mar. 1990.

[31] P J Green, “Statistical methods for spatial image analysis (discussion)”, *Proc. 46 Sect. ISI, Bull. ISI*, vol. 52, pp. 43–44, 1987.

[32] K Lange, “Convergence of EM image reconstruction algorithms with Gibbs smoothing”, *IEEE Trans. Med. Im.*, vol. 9, no. 4, pp. 439–446, Dec. 1990, Corrections, June 1991.

[33] Z Liang, R Jaszczak, and K Greer, “On Bayesian image reconstruction from projections: uniform and nonuniform *a priori* source information”, *IEEE Trans. Med. Im.*, vol. 8, no. 3, pp. 227–235, Sept. 1989.

[34] G T Herman and D Odhner, “Performance evaluation of an iterative image reconstruction algorithm for positron-emission tomography”, *IEEE Trans. Med. Im.*, vol. 10, no. 3, pp. 336–346, Sept. 1991.

[35] G T Herman, D Odhner, K D Toennies, and S A Zenios, “A parallelized algorithm for image reconstruction from noisy projections”, in *Large-Scale Numerical Optimization*, T F Coleman and Y Li, Eds., pp. 3–21. SIAM, Philadelphia, 1990.

[36] A W McCarthy and M I Miller, "Maximum likelihood SPECT in clinical computation times using mesh-connected parallel processors", *IEEE Trans. Med. Im.*, vol. 10, no. 3, pp. 426–436, Sept. 1991.

[37] M I Miller and B Roysam, "Bayesian image reconstruction for emission tomography incoporating Good's roughness prior on massively parallel processors", *Proc. Natl. Acad. Sci.*, vol. 88, pp. 3223–3227, Apr. 1991.

[38] A D Lantermann, "A new way to regularize maximum-likelihood estimates for emission tomography with Good's roughness penalty", Tech. Rep. ESSRL-92-05, Elec. Sys. and Sig. Res. Lab., Washington Univ., Jan. 1992.

[39] D L Snyder, A D Lanterman, and M I Miller, "Regularizing images in emission tomography via an extension of Good's roughness measure", Tech. Rep. ESSRL-92-17, Washington Univ., Jan. 1992.

[40] C Bouman and K Sauer, "Fast numerical methods for emission and transmission tomographic reconstruction", in *Proc. 27th Conf. Info. Sci. Sys., Johns Hopkins*, 1993, pp. 611–616.

[41] C Bouman and K Sauer, "A unified approach to statistical tomography using coordinate descent optimization", 1993, Submitted to Trans. Im. Proc.

[42] E U Mumcuoglu, R Leahy, S R Cherry, and Z Zhou, "Fast gradient-based methods for Bayesian reconstruction of transmission and emission PET images", Tech. Rep. 241, Signal and Image Processing Inst., USC, Sept. 1993.

[43] J A Fessler and A O Hero, "Space-alternating generalized EM algorithms for penalized maximum-likelihood image reconstruction", Tech. Rep. 286, Comm. and Sign. Proc. Lab., Dept. of EECS, Univ. of Michigan, Ann Arbor, MI, 48109-2122, Feb. 1994.

[44] A R De Pierro, "On the relation between the ISRA and the EM algorithm for positron emission tomography", *IEEE Trans. Med. Im.*, vol. 12, no. 2, pp. 328–333, June 1993.

[45] L Kaufman, "Implementing and accelerating the EM algorithm for positron emission tomography", *IEEE Trans. Med. Im.*, vol. 6, no. 1, pp. 37–51, Mar. 1987.

[46] A O Hero and J A Fessler, "A recursive algorithm for computing Cramer-Rao-type bounds on estimator covariance", *IEEE Trans. Info. Theory*, vol. 40, no. 4, pp. 1205–1210, July 1994.

[47] K Sauer and C Bouman, "A local update strategy for iterative reconstruction from projections", *IEEE Trans. Sig. Proc.*, vol. 41, no. 2, pp. 534–548, Feb. 1993.

[48] W H Press, B P Flannery, S A Teukolsky, and W T Vetterling, *Numerical Recipes in C*, Cambridge Univ. Press, 1988.

[49] H M Hudson and R S Larkin, "Acceleration image reconstruction using ordered subsets of projection data", *IEEE Trans. Med. Im.*, vol. 0, no. 0, pp. 0–0, 0 1994, To appear.

[50] M Rosenlicht, *Introduction to Analysis*, Dover, New York, 1985.

[51] R E Williamson, R H Crowell, and H F Trotter, *Calculus of Vector Functions*, Prentice Hall, New Jersey, 1972.

[52] A M Ostrowski, *Solution of equations in Euclidian and Banach spaces*, Academic Press, 1973.

[53] T M Cover, "An algorithm for maximizing expected log investment return", *IEEE Trans. Info. Theory*, vol. 30, no. 2, pp. 369–373, Mar. 1984.

Type-I Algorithm (e.g. ML-EM, PML-OSL)
and Type-II Algorithm (e.g. PML-GEM)

```
Initialize $\boldsymbol{\lambda}^0$
    for $i = 0, 1, \ldots$ {
        $\bar{y}_n = \sum_k a_{nk} \lambda_k^i + r_n, \; n = 1, \ldots, N$
        $s_n = y_n / \bar{y}_n, \; n = 1, \ldots, N$        (39)
        $e_k = \sum_n a_{nk} s_n, \; k = 1, \ldots, p$        (40)
        for $k = 1, \ldots, p$ {
            $\lambda_k^{i+1} = g_k(e_k; \boldsymbol{\lambda}^i)$,        (Type-I)  (see (14), (21), (22), (23), (37))        (41)
            $\lambda_k^{i+1} = g_k(e_k; \boldsymbol{\lambda}^\star; \boldsymbol{\lambda}^i)$,        or (Type-II)  (see (36))        (42)
        }
    }
```

Type-III Algorithm (e.g. ML-SAGE, PML-SAGE)

```
Initialize $\boldsymbol{\lambda}^0$,        $\bar{y}_n = \sum_k a_{nk} \lambda_k^0 + r_n, \; n = 1, \ldots, N.$
    for $i = 0, 1, \ldots$ {
        $k = 1 + (i \text{ modulo } p)$
        $e_k = \sum_n a_{nk} y_n / \bar{y}_n$        (43)
        $\lambda_k^{i+1} = g_k(e_k; \boldsymbol{\lambda}^i)$  (see (27), (33))        (44)
        $\lambda_j^{i+1} = \lambda_j^i, \; j \neq k,$
        $\bar{y}_n := \bar{y}_n + (\lambda_k^{i+1} - \lambda_k^i) a_{nk}, \; \forall n : a_{nk} \neq 0$        (45)
    }
```

Table 1: Three generic pseudo-code algorithm types for penalized maximum-likelihood image reconstruction. All of the algorithms presented in the text are of one of these three types. Within each type, the algorithms differ in form of the functions $g()$ used in the M-step.

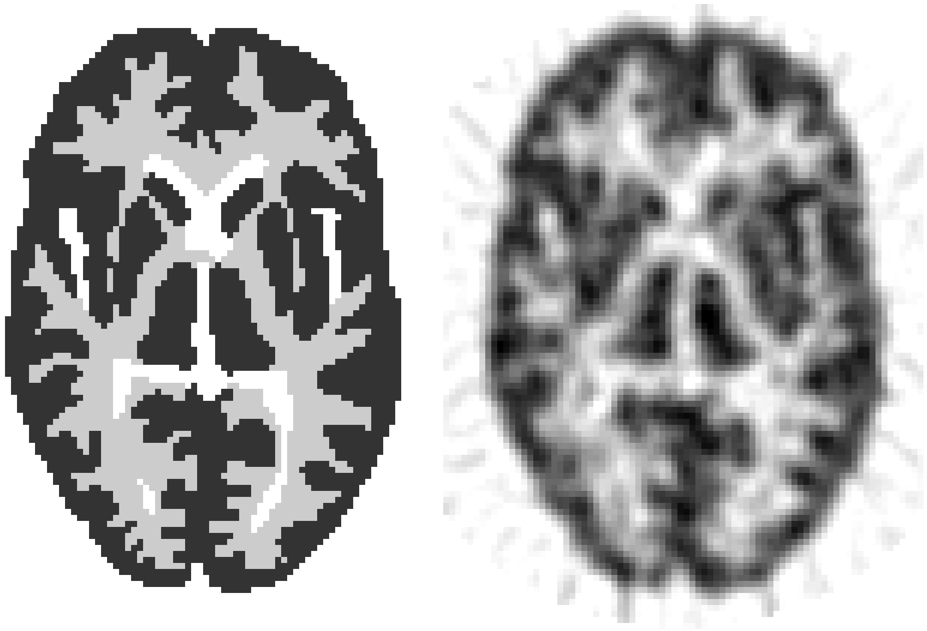

Figure 1: Digital brain phantom (left), and filtered backprojection reconstructed image (right).

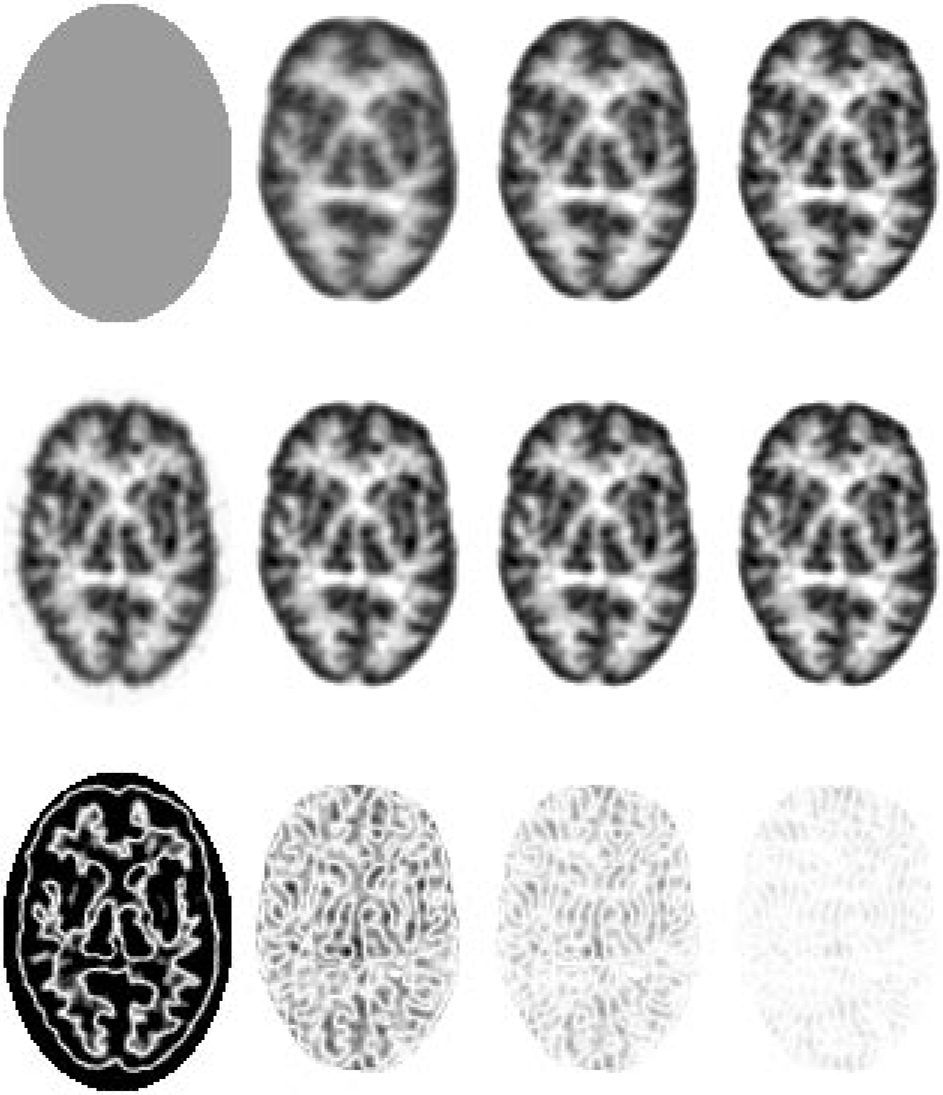

Figure 2: PML-SAGE-2 estimates from data with 5% random coincidences at iterations $i = 0, 5, 10, 20$ (left to right). Top row: initialized with uniform image. Middle row: initialized with thresholded filtered-backprojection image. Bottom row: absolute value of difference between top and middle rows amplified by a factor of 4.

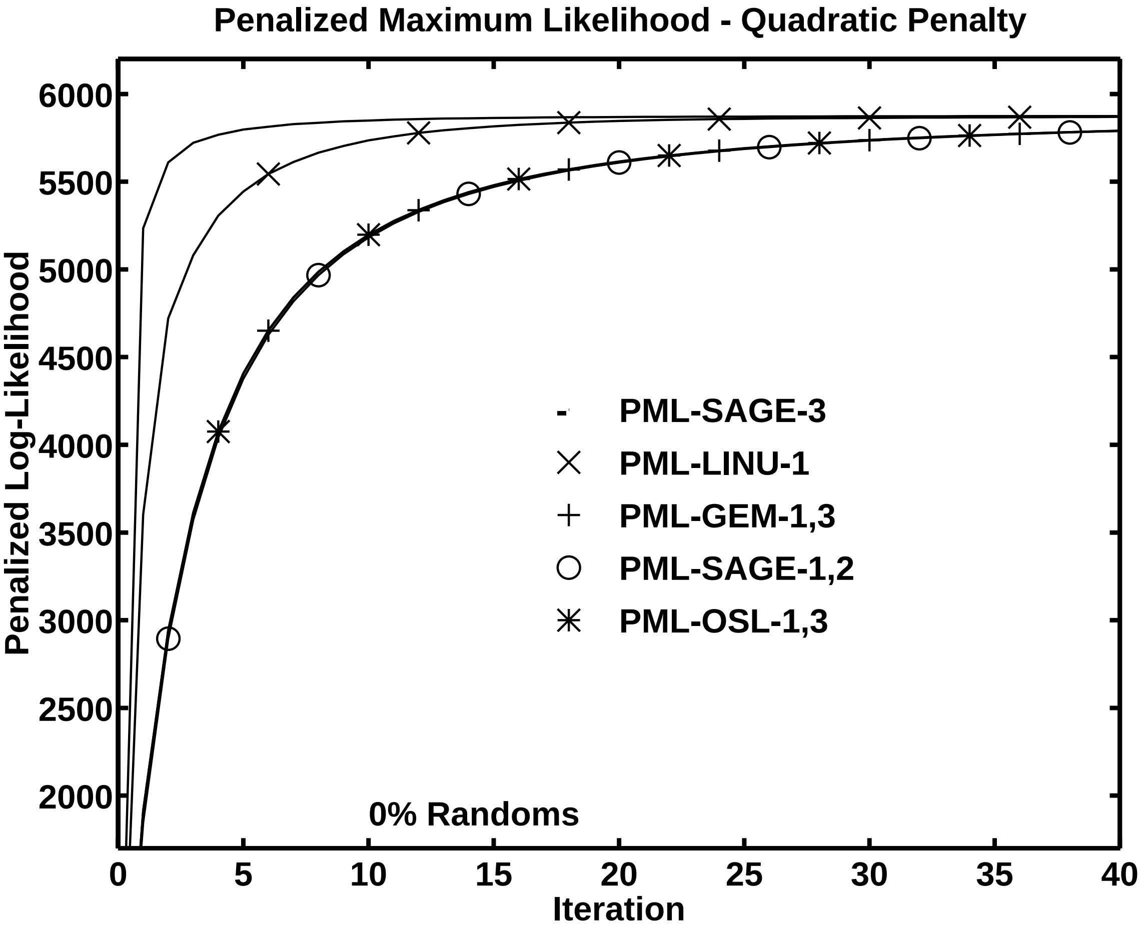

Figure 3: Penalized likelihood $\Phi(\boldsymbol{\lambda}^i) - \Phi(\boldsymbol{\lambda}^0)$ vs. iteration from data with 0% random coincidences.

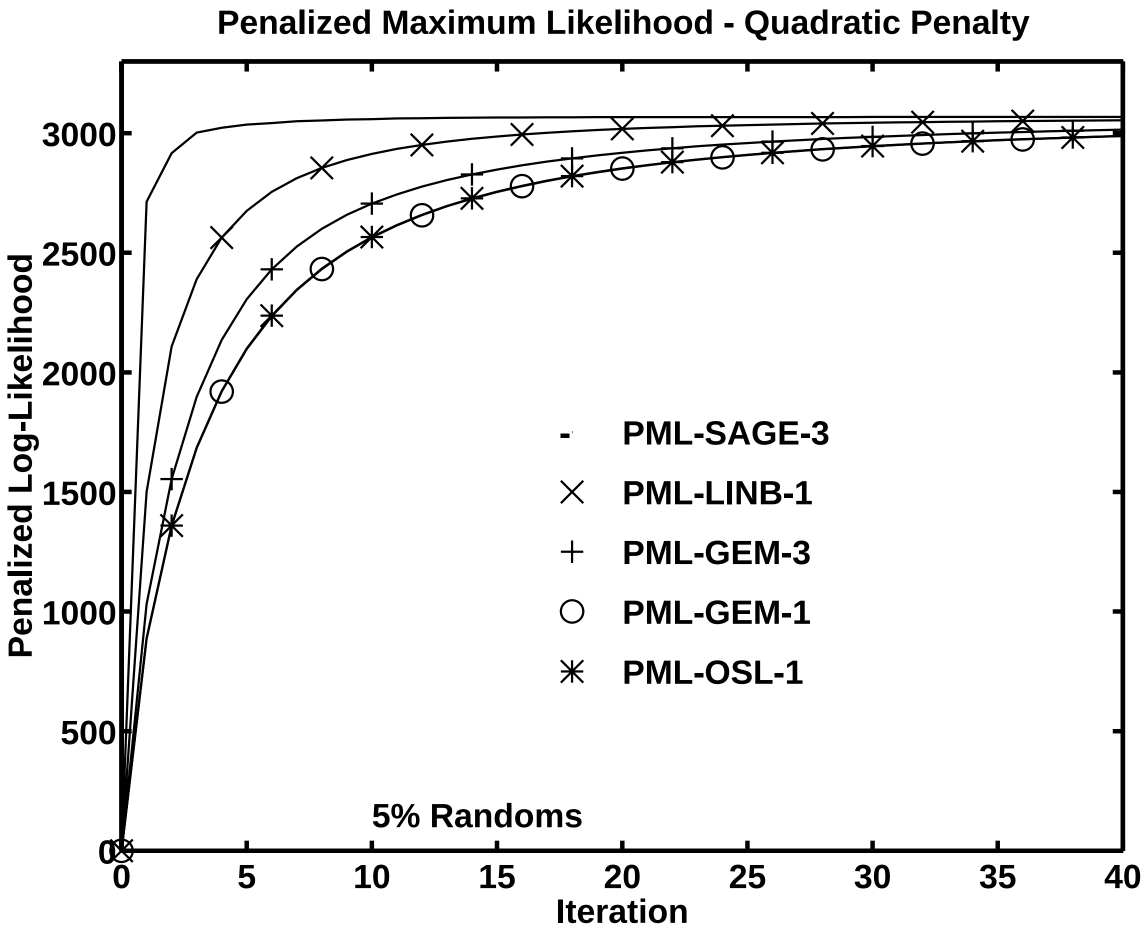

Figure 4: Penalized likelihood $\Phi(\boldsymbol{\lambda}^i) - \Phi(\boldsymbol{\lambda}^0)$ vs. iteration from data with 5% random coincidences. Not shown is PML-SAGE-2, which converges slightly slower than PML-SAGE-3. Also not shown is PML-SAGE-1, which is indistinguishable from PML-OSL-1.

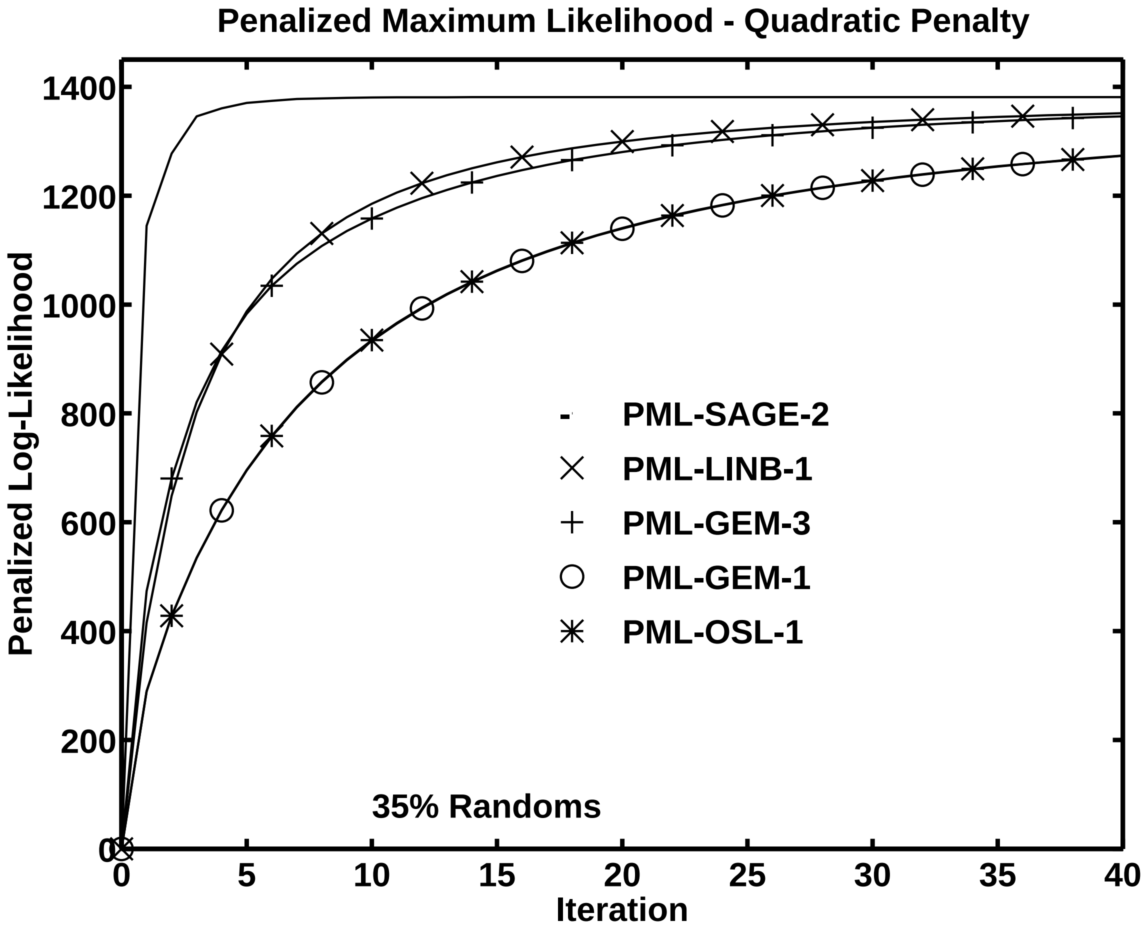

Figure 5: As in Fig. 3, but with 35% random coincidences.