On complete-data spaces for PET reconstruction algorithms

Jeffrey A. Fessler, Neal H. Clinthorne, and W. Leslie Rogers
The University of Michigan

IEEE T-NS 40(4):1055-61, Aug. 1993

Abstract

As investigators consider more comprehensive measurement models for emission tomography, there will be more choices for the complete-data spaces of the associated expectation-maximization (EM) algorithms for maximum-likelihood (ML) estimation. In this paper, we show that EM algorithms based on smaller complete-data spaces will typically converge faster. We discuss two practical applications of these concepts: (i) the ML-IA and ML-IB image reconstruction algorithms of Politte and Snyder [1] which are based on measurement models that account for attenuation and accidental coincidences in positron-emission tomography (PET), and (ii) the problem of simultaneous estimation of emission and transmission parameters. Although the PET applications may often violate the necessary regularity conditions, our analysis predicts heuristically that the ML-IB algorithm, which has a smaller complete-data space, should converge faster than ML-IA. This is corroborated by the empirical findings in [1].

I. Introduction

The ML criterion for tomographic image reconstruction has received considerable attention since Shepp and Vardi [2] introduced an EM algorithm for computing ML estimates. Although the medical imaging community often refers to "the" ML-EM algorithm, there are in fact a multitude of feasible EM algorithms, each based on a different complete-data space. A useful complete-data space supplements the observed measurements in a way that facilitates parameter estimation [3]. Although only one complete-data space has been suggested for PET under the simple measurement model used in the early papers [2,4], there will be more choices as investigators consider more comprehensive measurement models, such as those accounting for photon attenuation [5], accidental coincidences [1], deadtime, and scatter [6,7]. This paper illustrates the importance of parsimony in choosing complete-data spaces, and some of the tradeoffs that result.

Accurate quantification of radiotracer activity using PET must include corrections for the effects of attenuation and accidental coincidences. Recently, Politte and Snyder proposed two ML-EM algorithms for PET image reconstruction that directly incorporate the effects of known attenuation and accidental coincidences into the statistical measurement model [1]. The algorithms are based on two different complete-data spaces, one of which is a subset of the other. They observed in experiments that the algorithm based on the smaller complete-data space converged faster. In this paper we corroborate their observations by proving that smaller complete-data spaces yield EM algorithms with faster asymptotic convergence rates.

The measurement models used in [1] assumed exact knowledge of the survival probabilities, the probability that both photons of a positron-produced pair escape unattenuated. In practice, one must obtain these factors experimentally, typically by a transmission scan that precedes the radiotracer injection. As mentioned in [1], a more accurate approach would account for the statistical uncertainties in both the emission data and the transmission measurements. An iterative method for simultaneously estimating the emission intensities and the survival probabilities has been recently proposed by Clinthorne *et al.* [8]. In this paper, we present two algorithms for joint emission/transmission estimation based on generalizations of the two complete-data spaces in [1]. We demonstrate that although the smaller complete-data space may provide a faster algorithm in theory, in practice the larger complete-data space leads to an EM algorithm with an easier maximization step. Such tradeoffs have been observed in other EM applications [3], but they may be particularly important to future investigations of PET reconstruction methods because of the large dimensions of the parameter spaces.

In Section II, we briefly review the EM algorithm, and prove that smaller complete-data spaces result in faster convergence. This is applied to PET in Section III. In Section IV we analyze the joint emission/transmission estimation method for PET.

II. Theory

A. EM algorithm

We observe $\boldsymbol{y}$, a realization of a random vector $\boldsymbol{Y}$ having known density $g(\boldsymbol{y};\boldsymbol{\theta})$, with the goal of computing the ML estimate of $\boldsymbol{\theta}$. In many problems, including emission tomography, the measurements are "incomplete" in the sense that several components of $\boldsymbol{\theta}$ may contribute to each component of $\boldsymbol{Y}$. In such problems one can often postulate a "complete data" random vector $\boldsymbol{X}$ that

This work was supported in part by a National Institute of Health Training Fellowship, a DOE Alexander Hollaender Postdoctoral Fellowship, and DOE Grant DE-FG02-87ER65061.

is more naturally related to the parameter vector $\boldsymbol{\theta}$, and is related to the observed measurements by a many-to-one mapping $\boldsymbol{Y} = \mathrm{h}(\boldsymbol{X})$. The density $f(\boldsymbol{x};\boldsymbol{\theta})$ of the complete data $\boldsymbol{X}$ must be consistent with the incomplete data $\boldsymbol{Y}$ in that:

$$g(\boldsymbol{y};\boldsymbol{\theta}) = \int_{\{\boldsymbol{x}:\boldsymbol{y}=h(\boldsymbol{x})\}} f(\boldsymbol{x};\boldsymbol{\theta})\,d\boldsymbol{x}.$$

Let

$$Q(\bar{\boldsymbol{\theta}};\boldsymbol{\theta}) \stackrel{\triangle}{=} E\left\{\log f(\boldsymbol{X};\bar{\boldsymbol{\theta}})\,|\boldsymbol{Y}=\boldsymbol{y};\boldsymbol{\theta}\right\} \tag{1}$$

$$= \int \log f(\boldsymbol{x};\bar{\boldsymbol{\theta}})\, f(\boldsymbol{x}|\boldsymbol{Y}=\boldsymbol{y};\boldsymbol{\theta})\,d\boldsymbol{x}$$

$$= H(\bar{\boldsymbol{\theta}};\boldsymbol{\theta}) + L(\bar{\boldsymbol{\theta}}), \tag{2}$$

where

$$\begin{aligned} H(\bar{\boldsymbol{\theta}};\boldsymbol{\theta}) &\stackrel{\triangle}{=} E\left\{\log f(\boldsymbol{X}|\boldsymbol{Y}=\boldsymbol{y};\bar{\boldsymbol{\theta}})\,|\boldsymbol{Y}=\boldsymbol{y};\boldsymbol{\theta}\right\}, \\ L(\boldsymbol{\theta}) &\stackrel{\triangle}{=} \log g(\boldsymbol{y};\boldsymbol{\theta}). \end{aligned}$$

The EM algorithm for ML estimation [3], calls for iterating over the following steps:
E-step:

$$\text{Compute } Q(\boldsymbol{\theta};\boldsymbol{\theta}^i),$$

M-step:

$$\boldsymbol{\theta}^{i+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta};\boldsymbol{\theta}^i),$$

where $\boldsymbol{\theta}^i$ denotes the parameter estimate after the ith iteration. Note that by Jensen's inequality [3]:

$$H(\bar{\boldsymbol{\theta}};\boldsymbol{\theta}) \le H(\boldsymbol{\theta};\boldsymbol{\theta}) \;\; \forall \bar{\boldsymbol{\theta}},$$

so the EM algorithm produces a likelihood sequence $L(\boldsymbol{\theta}^i)$ that is monotonically increasing. The basic idea is to compute Q, the conditional expectation of the complete data given the most recent parameter estimate, and then to maximize the parameter's likelihood as if one had observed the complete data [3]. The EM algorithm is most useful when the complete-data space is chosen such that $Q(\boldsymbol{\theta};\boldsymbol{\theta}^i)$ can be maximized analytically for the M-step, although other approaches are possible [9].

B. EM convergence rate

Several investigators have observed empirically that larger complete-data spaces correspond to slower EM convergence [3, pp. 25,34]. In this section we formally establish a version of this result. For our purposes, asymptotic convergence rate is defined by the following well known result [10, p. 301].

Linear Convergence Theorem: *If (i) $G : \mathcal{D} \in \Re^n \to \Re^n$ has a fixed point $\boldsymbol{\theta}^\star \in \mathcal{D}_+ = \mathrm{int}(\mathcal{D})$, (ii) G is Fréchet differentiable at $\boldsymbol{\theta}^\star$, and (iii) $\rho(\nabla^1 G(\boldsymbol{\theta}^\star)) < 1$, where $\rho()$ denotes spectral radius[1], then the root-convergence factor [10, p. 288] R_1 at $\boldsymbol{\theta}^\star$ for the iterative process $\boldsymbol{\theta}^{i+1} = G(\boldsymbol{\theta}^i)$ is given by $R_1(G,\boldsymbol{\theta}^\star) = \rho(\nabla^1 G(\boldsymbol{\theta}^\star))$.*

This leads to the following definition of convergence rate for EM algorithms corresponding to strictly concave likelihood functions [3].

Theorem 1: *Let $\boldsymbol{\theta}^{i+1} = G(\boldsymbol{\theta}^i)$ define the iterations for an EM algorithm such that (i) G and $\boldsymbol{\theta}^\star$ satisfy conditions (i) and (ii) of the Linear Convergence Theorem, (ii) G is defined by solving the system of equations $\nabla^{10} Q(\bar{\boldsymbol{\theta}};\boldsymbol{\theta})|_{\bar{\boldsymbol{\theta}}=G(\boldsymbol{\theta}))} = \mathbf{0}$, and (iii) $\boldsymbol{L} \stackrel{\triangle}{=} -\nabla^2 L(\boldsymbol{\theta}^\star)$ is positive definite, then the root-convergence factor at $\boldsymbol{\theta}^\star$ for the EM iteration G is*

$$R_1 = \rho\left(\boldsymbol{I} - \boldsymbol{Q}^{-1}\boldsymbol{L}\right) < 1, \tag{3}$$

where $\boldsymbol{I}$ is the $n \times n$ identity matrix and $\boldsymbol{Q} \stackrel{\triangle}{=} -\nabla^{20} Q(\boldsymbol{\theta}^\star;\boldsymbol{\theta}^\star)$.

Proof: $\boldsymbol{H} \stackrel{\triangle}{=} -\nabla^{20} H(\boldsymbol{\theta}^\star;\boldsymbol{\theta}^\star)$ is nonnegative definite since it is a (conditional) Fisher information matrix [11, p. 126]. From (2), $\boldsymbol{Q} = \boldsymbol{H} + \boldsymbol{L}$, so by (iii), $\boldsymbol{Q}$ is positive definite and therefore invertible. From (ii) we see that $\nabla^{10} Q(G(\boldsymbol{\theta});\boldsymbol{\theta}) = \mathbf{0}$, which can be differentiated again to yield $\nabla^{20} Q(G(\boldsymbol{\theta});\boldsymbol{\theta})\nabla^1 G(\boldsymbol{\theta}) + \nabla^{11} Q(G(\boldsymbol{\theta});\boldsymbol{\theta}) = \mathbf{0}$. Therefore,

$$\nabla^1 G(\boldsymbol{\theta}^\star) = -(\nabla^{20} Q(\boldsymbol{\theta}^\star;\boldsymbol{\theta}^\star))^{-1}\nabla^{11} Q(\boldsymbol{\theta}^\star;\boldsymbol{\theta}^\star),$$

[1] The (n,m) element of the $\dim(\boldsymbol{\theta})$ by $\dim(\boldsymbol{\theta})$ matrix $\nabla^1 G$ is $\partial/\partial\theta_m G_n(\boldsymbol{\theta})$. In general, the $\dim(\boldsymbol{\theta})$ by $\dim(\boldsymbol{\theta})$ matrix $\nabla^{ij} F(\bar{\boldsymbol{\theta}};\boldsymbol{\theta})$ denotes the matrix of partials $\partial^{i+j} F(\bar{\boldsymbol{\theta}};\boldsymbol{\theta})/\partial\bar{\boldsymbol{\theta}}^i\partial\boldsymbol{\theta}^j$.

IEEE Trans. on Nuclear Science, 40(4):1055-61, Aug. 1993

since $G(\boldsymbol{\theta}^\star) = \boldsymbol{\theta}^\star$ by (i). But it is well known [3] that for $\boldsymbol{\theta} \in \mathcal{D}_+$

$$\nabla^{11} Q(\boldsymbol{\theta}; \boldsymbol{\theta}) = \nabla^{11} H(\boldsymbol{\theta}; \boldsymbol{\theta}) = -\nabla^{20} H(\boldsymbol{\theta}; \boldsymbol{\theta})$$
$$= \nabla^2 L(\boldsymbol{\theta}) - \nabla^{20} Q(\boldsymbol{\theta}; \boldsymbol{\theta}).$$

The theorem then follows from combining the two equations above and the Linear Convergence Theorem, provided $\rho\left(\boldsymbol{I} - \boldsymbol{Q}^{-1}\boldsymbol{L}\right) < 1$. To prove this, we parallel an argument of Green [12]. If α is an eigenvalue of $\boldsymbol{I} - \boldsymbol{Q}^{-1}\boldsymbol{L}$, then $|\boldsymbol{I} - \boldsymbol{Q}^{-1}\boldsymbol{L} - \alpha \boldsymbol{I}| = 0$, hence $|(1-\alpha)\boldsymbol{Q} - \boldsymbol{L}| = 0$. Since $\boldsymbol{Q} = \boldsymbol{H} + \boldsymbol{L}$, $|(1-\alpha)\boldsymbol{H} - \alpha\boldsymbol{L}| = 0$. Thus, by assumption (iii), we must have $\alpha \in [0,1)$, hence $\rho\left(\boldsymbol{I} - \boldsymbol{Q}^{-1}\boldsymbol{L}\right) < 1$. □

Since $\boldsymbol{Q} = \boldsymbol{H} + \boldsymbol{L}$, where $\boldsymbol{H}$ is a conditional Fisher information matrix, one sees from (3) that if a larger complete-data space has more Fisher information, then the corresponding root-convergence factor will be larger, and the asymptotic convergence rate will be slower. This is the idea behind the next lemma and theorem, the main results of this section.

Lemma 1: *If (i) $\boldsymbol{Q}_{\rm B} = \boldsymbol{H} + \boldsymbol{L}$, where $\boldsymbol{H}$ is symmetric nonnegative definite and $\boldsymbol{L}$ is symmetric positive definite, and (ii) $\boldsymbol{Q}_{\rm A} = \boldsymbol{Q}_{\rm B} + \boldsymbol{N}$ where $\boldsymbol{N}$ is symmetric nonnegative definite, then $\rho_{\rm B} \leq \rho_{\rm A}$, where $\rho_{\rm A} = \rho\left(\boldsymbol{I} - \boldsymbol{Q}_{\rm A}^{-1}\boldsymbol{L}\right)$ and $\rho_{\rm B} = \rho\left(\boldsymbol{I} - \boldsymbol{Q}_{\rm B}^{-1}\boldsymbol{L}\right)$. Furthermore, if $\boldsymbol{N}$ is symmetric positive definite, then $\rho_{\rm B} < \rho_{\rm A}$.*

Proof: Again we borrow from Green [12]. By the arguments in Theorem 1, $0 \leq \rho_{\rm A} < 1$ and $0 \leq \rho_{\rm B} < 1$. Since $\rho_{\rm B} = \rho\left(\boldsymbol{I} - \boldsymbol{Q}_{\rm B}^{-1}\boldsymbol{L}\right)$, $\exists \boldsymbol{u} \neq 0$ s.t. $(\boldsymbol{I} - \boldsymbol{Q}_{\rm B}^{-1}\boldsymbol{L})\boldsymbol{u} = \rho_{\rm B}\boldsymbol{u}$, so $(1-\rho_{\rm B})\boldsymbol{Q}_{\rm B}\boldsymbol{u} - \boldsymbol{L}\boldsymbol{u} = 0$. By (ii), $(1-\rho_{\rm B})\boldsymbol{Q}_{\rm A}\boldsymbol{u} - \boldsymbol{L}\boldsymbol{u} = (1-\rho_{\rm B})\boldsymbol{N}\boldsymbol{u}$, so $(1-\rho_{\rm B})\boldsymbol{Q}_{\rm A}^{\frac{1}{2}}\boldsymbol{u} - \boldsymbol{Q}_{\rm A}^{-\frac{1}{2}}\boldsymbol{L}\boldsymbol{u} = (1-\rho_{\rm B})\boldsymbol{Q}_{\rm A}^{-\frac{1}{2}}\boldsymbol{N}\boldsymbol{u}$. Defining $\boldsymbol{v} = \boldsymbol{Q}_{\rm A}^{\frac{1}{2}}\boldsymbol{u}$, it follows that $(1-\rho_{\rm B})\boldsymbol{v}'\boldsymbol{v} - \boldsymbol{v}'\boldsymbol{Q}_{\rm A}^{-\frac{1}{2}}\boldsymbol{L}\boldsymbol{Q}_{\rm A}^{-\frac{1}{2}}\boldsymbol{v} = (1-\rho_{\rm B})\boldsymbol{v}'\boldsymbol{Q}_{\rm A}^{-\frac{1}{2}}\boldsymbol{N}\boldsymbol{Q}_{\rm A}^{-\frac{1}{2}}\boldsymbol{v} \geq 0$, if $\boldsymbol{N}$ is nonnegative definite. Hence, $\boldsymbol{v}'[\boldsymbol{I} - \boldsymbol{Q}_{\rm A}^{-\frac{1}{2}}\boldsymbol{L}\boldsymbol{Q}_{\rm A}^{-\frac{1}{2}}]\boldsymbol{v} \geq \rho_{\rm B}\boldsymbol{v}'\boldsymbol{v}$, so $\rho(\boldsymbol{I} - \boldsymbol{Q}_{\rm A}^{-\frac{1}{2}}\boldsymbol{L}\boldsymbol{Q}_{\rm A}^{-\frac{1}{2}}) \geq \rho_{\rm B}$. But $\rho(\boldsymbol{I} - \boldsymbol{Q}_{\rm A}^{-\frac{1}{2}}\boldsymbol{L}\boldsymbol{Q}_{\rm A}^{-\frac{1}{2}}) = \rho(\boldsymbol{I} - \boldsymbol{Q}_{\rm A}^{-1}\boldsymbol{L}) = \rho_{\rm A}$, so $\rho_{\rm A} \geq \rho_{\rm B}$. The case when $\boldsymbol{N}$ is positive definite is similar. □

Theorem 2: *If (i) $G_{\rm A}$ and $G_{\rm B}$ are two EM algorithms that satisfy the conditions of Theorem 1 and that correspond to complete-data spaces $\boldsymbol{X}_{\rm A}$ and $\boldsymbol{X}_{\rm B}$ respectively, (ii) $\boldsymbol{X}_{\rm B}$ is a subset of $\boldsymbol{X}_{\rm A}$, i.e., $\boldsymbol{X}_{\rm A} = [\boldsymbol{X}_{\rm B}', \boldsymbol{X}_\circ']'$, (iii) $f_{\rm A}(\boldsymbol{x}_{\rm A}|\boldsymbol{y};\boldsymbol{\theta}) = f_{\rm A}([\boldsymbol{x}_{\rm B}, \boldsymbol{x}_\circ]|\boldsymbol{y};\boldsymbol{\theta}) = f_{\rm B}(\boldsymbol{x}_{\rm B}|\boldsymbol{y};\boldsymbol{\theta}) f_\circ(\boldsymbol{x}_\circ|\boldsymbol{y};\boldsymbol{\theta})$, and (iv) $f_\circ(\boldsymbol{x}_\circ|\boldsymbol{y};\boldsymbol{\theta}) = f_\circ(\boldsymbol{x}_\circ;\boldsymbol{\theta})$, i.e., $\boldsymbol{X}_\circ$ is extraneous complete-data, then algorithm* B *converges faster than algorithm* A *asymptotically at a common fixed point $\boldsymbol{\theta}^\star$.*

Proof: By integrating (iii) and using (iv), one sees that $f_{\rm A}(\boldsymbol{x}_{\rm A};\boldsymbol{\theta}) = f_{\rm A}([\boldsymbol{x}_{\rm B}, \boldsymbol{x}_\circ];\boldsymbol{\theta}) = f_{\rm B}(\boldsymbol{x}_{\rm B};\boldsymbol{\theta}) f_\circ(\boldsymbol{x}_\circ;\boldsymbol{\theta})$. Let $Q_{\rm A}$ and $Q_{\rm B}$ denote the Q function (1) for $\boldsymbol{X}_{\rm A}$ and $\boldsymbol{X}_{\rm B}$ respectively, then

$$Q_{\rm A}(\bar{\boldsymbol{\theta}};\boldsymbol{\theta}) = \int \log f_{\rm A}(\boldsymbol{x}_{\rm A};\bar{\boldsymbol{\theta}})\; f_{\rm A}(\boldsymbol{x}_{\rm A}|\boldsymbol{y};\boldsymbol{\theta})\; d\boldsymbol{x}_{\rm A}$$
$$= \int \log f_{\rm A}([\boldsymbol{x}_{\rm B}, \boldsymbol{x}_\circ];\bar{\boldsymbol{\theta}})\; f_{\rm A}([\boldsymbol{x}_{\rm B}, \boldsymbol{x}_\circ]|\boldsymbol{y};\boldsymbol{\theta})\; d\boldsymbol{x}_{\rm B}\; d\boldsymbol{x}_\circ$$
$$= \int \log(f_{\rm B}(\boldsymbol{x}_{\rm B};\bar{\boldsymbol{\theta}}) f_\circ(\boldsymbol{x}_\circ;\bar{\boldsymbol{\theta}}))\; f_{\rm B}(\boldsymbol{x}_{\rm B}|\boldsymbol{y};\boldsymbol{\theta}) f_\circ(\boldsymbol{x}_\circ|\boldsymbol{y};\boldsymbol{\theta})\; d\boldsymbol{x}_{\rm B}\; d\boldsymbol{x}_\circ$$
$$= \int \log f_{\rm B}(\boldsymbol{x}_{\rm B};\bar{\boldsymbol{\theta}})\; f_{\rm B}(\boldsymbol{x}_{\rm B}|\boldsymbol{y};\boldsymbol{\theta})\; d\boldsymbol{x}_{\rm B}$$
$$+ \int \log f_\circ(\boldsymbol{x}_\circ;\bar{\boldsymbol{\theta}})\; f_\circ(\boldsymbol{x}_\circ;\boldsymbol{\theta})\; d\boldsymbol{x}_\circ$$
$$= Q_{\rm B}(\bar{\boldsymbol{\theta}};\boldsymbol{\theta}) + E\{\log f_\circ(\boldsymbol{X}_\circ;\bar{\boldsymbol{\theta}});\boldsymbol{\theta}\}.$$

Therefore,

$$-\nabla^{20} Q_{\rm A}(\boldsymbol{\theta};\boldsymbol{\theta}) = -\nabla^{20} Q_{\rm B}(\boldsymbol{\theta};\boldsymbol{\theta}) + \boldsymbol{J}_\circ(\boldsymbol{\theta}),$$

where $\boldsymbol{J}_\circ(\boldsymbol{\theta})$ is the Fisher information matrix for the extraneous complete data $\boldsymbol{X}_\circ$ at $\boldsymbol{\theta}$. The conclusion then follows from Lemma 1 and Theorem 1. □

In summary, Theorem 2 shows that a complete-data space with extraneous variables will lead to an EM algorithm with slower convergence[2]. Our derivation of this theorem relies on several assumptions, including convergence to an interior point, and strict concavity of the likelihood. The applicability of these assumptions in the PET context is discussed in the next section.

[2]We have since shown in [13] that Theorem 2 is true under considerably less restrictive conditions than (ii)-(iv), but this version is sufficient for the purposes of this paper.

III. PET RECONSTRUCTION

A. Two complete-data spaces

In this section, we briefly review the two PET reconstruction algorithms investigated in [1], and then discuss their convergence rates. For simplicity, we assume that the object is discretized into B voxels, and denote the radioactivity in the bth voxel by λ_b. PET measurements are acquired using D detector pairs, with the number of counts recorded by the dth pair denoted Y_d. The iterative algorithms attempt to estimate $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_B]'$ (corresponding to the parameter $\boldsymbol{\theta}$ in the preceding section) from a realization $\boldsymbol{y}$ of $\boldsymbol{Y} = [Y_1, \ldots, Y_D]'$. The parameter domain is $\mathcal{D} = \{\boldsymbol{\lambda} : \lambda_b \geq 0,\ b = 1, \ldots, B\}$.

Let p_{db} denote the point-spread function of the system, normalized so that $\sum_{d=1}^{D} p_{db} = 1$. Let q_b denote the detection probability for an unattenuated event originating in voxel b, and α_d denote the survival probability, assumed known here, for a photon pair emitted towards the dth detector pair. Then $\alpha_d w_{db}$ is the probability that an event in voxel b is detected by the dth detector pair, where $w_{db} = p_{db} q_b$. Let N_{db} denote the number of events from voxel b contributing to detector pair d; the N_{db}'s have independent Poisson distributions with means $\alpha_d w_{db} \lambda_b$. Let R_d denote the number of accidental coincidences counted by detector pair d; the R_d's have independent Poisson distributions with mean r_d, assumed known. The total counts in the dth detector pair is then

$$Y_d = \sum_{b=1}^{B} N_{db} + R_d,\ d = 1, \ldots, D. \tag{4}$$

The two EM algorithms described in [1] were called ML-IA and ML-IB. The complete-data space for ML-IB is

$$\boldsymbol{X}_{\mathrm{IB}} = \{\{N_{db}\},\ \{R_d\}\},\ \begin{array}{l} d = 1, \ldots, D \\ b = 1, \ldots, B \end{array}.$$

Under the distributions given above,

$$\log f_{\mathrm{IB}}(\boldsymbol{X}_{\mathrm{IB}}; \bar{\boldsymbol{\lambda}}) = \sum_{d=1}^{D} (-r_d + R_d \log(r_d))$$
$$+ \sum_{d=1}^{D} \sum_{b=1}^{B} \left(-\alpha_d w_{db} \bar{\lambda}_b + N_{db} \log(\alpha_d w_{db} \bar{\lambda}_b)\right). \tag{5}$$

It follows from (4) that [1]

$$E\{N_{db} | \boldsymbol{y}; \boldsymbol{\lambda}\} = y_d \frac{\alpha_d w_{db} \lambda_b}{\hat{y}_d(\boldsymbol{\lambda})} \tag{6}$$
$$E\{R_d | \boldsymbol{y}; \boldsymbol{\lambda}\} = y_d \frac{r_d}{\hat{y}_d(\boldsymbol{\lambda})}, \tag{7}$$

where

$$\hat{y}_d(\boldsymbol{\lambda}) = \sum_{b=1}^{B} \alpha_d w_{db} \lambda_b + r_d. \tag{8}$$

Combining equations (5)-(7) and (1) yields

$$Q_{\mathrm{IB}}(\bar{\boldsymbol{\lambda}}; \boldsymbol{\lambda}) = \sum_{d=1}^{D} \left(-r_d + y_d \frac{r_d}{\hat{y}_d(\boldsymbol{\lambda})} \log(r_d)\right)$$
$$+ \sum_{d=1}^{D} \sum_{b=1}^{B} \left(-\alpha_d w_{db} \bar{\lambda}_b + y_d \frac{\alpha_d w_{db} \lambda_b}{\hat{y}_d(\boldsymbol{\lambda})} \log(\alpha_d w_{db} \bar{\lambda}_b)\right). \tag{9}$$

Maximizing $Q_{\mathrm{IB}}(\bar{\boldsymbol{\lambda}}; \boldsymbol{\lambda})$ over $\bar{\boldsymbol{\lambda}}$ yields the ML-IB iteration $\boldsymbol{\lambda}^{i+1} = G_{\mathrm{IB}}(\boldsymbol{\lambda}^i)$, where

$$G_{\mathrm{IB}}(\boldsymbol{\lambda}) \stackrel{\triangle}{=} \boldsymbol{\lambda} \odot [\boldsymbol{W}'(\boldsymbol{\alpha} \odot \boldsymbol{y} \oslash \hat{\boldsymbol{y}}(\boldsymbol{\lambda}))] \oslash [\boldsymbol{W}'\boldsymbol{\alpha}],$$

$\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_D]'$, $\hat{\boldsymbol{y}}(\boldsymbol{\lambda}) = [\hat{y}_1(\boldsymbol{\lambda}), \ldots, \hat{y}_D(\boldsymbol{\lambda})]'$, $\boldsymbol{W} = \{w_{db}\}$, and $\boldsymbol{r} = [r_1, \ldots, r_D]'$. The symbols $\odot$ and $\oslash$ denote component-wise multiplication and division respectively.

The complete-data space $\boldsymbol{X}_{\mathrm{IA}}$ for ML-IA is

$$\boldsymbol{X}_{\mathrm{IA}} = \{\{N_{db}\},\ \{R_d\},\ \{N_{db}^{\circ}\}\},\ \begin{array}{l} d = 1, \ldots, D \\ b = 1, \ldots, B \end{array},$$

which includes not only the components of $\boldsymbol{X}_{\text{IB}}$, but also the attenuated counts N°_{db}. These have independent Poisson distributions with mean $(1-\alpha_d)w_{db}\lambda_b$. Thus,

$$\log f_{\text{IA}}(\boldsymbol{X}_{\text{IA}}; \bar{\boldsymbol{\lambda}}) = \log f_{\text{IB}}(\boldsymbol{X}_{\text{IB}}; \bar{\boldsymbol{\lambda}}) + \log f_{\circ}(\boldsymbol{N}^{\circ}; \bar{\boldsymbol{\lambda}}), \tag{10}$$

where

$$\log f_{\circ}(\boldsymbol{N}^{\circ}; \bar{\boldsymbol{\lambda}}) = \sum_{d=1}^{D}\sum_{b=1}^{B} \left(-(1-\alpha_d)w_{db}\bar{\lambda}_b + N^{\circ}_{db}\log((1-\alpha_d)w_{db}\bar{\lambda}_b)\right). \tag{11}$$

Since N°_{db} makes *no contribution* to the measurements $\boldsymbol{Y}$,

$$E\{N^{\circ}_{db}|\boldsymbol{y};\boldsymbol{\lambda}\} = (1-\alpha_d)w_{db}\lambda_b. \tag{12}$$

Combining equations (10)-(12) and (1) yields

$$\begin{aligned} Q_{\text{IA}}(\bar{\boldsymbol{\lambda}};\boldsymbol{\lambda}) &= Q_{\text{IB}}(\bar{\boldsymbol{\lambda}};\boldsymbol{\lambda}) + \sum_{d=1}^{D}\sum_{b=1}^{B}\left(-(1-\alpha_d)w_{db}\bar{\lambda}_b\right. \\ &+ \left.(1-\alpha_d)w_{db}\lambda_b\log((1-\alpha_d)w_{db}\bar{\lambda}_b)\right). \end{aligned}$$

Maximizing $Q_{\text{IA}}(\bar{\boldsymbol{\lambda}};\boldsymbol{\lambda})$ over $\bar{\boldsymbol{\lambda}}$ yields the ML-IA iteration $\boldsymbol{\lambda}^{i+1} = G_{\text{IA}}(\boldsymbol{\lambda}^i)$, where

$$\begin{aligned} G_{\text{IA}}(\boldsymbol{\lambda}) &\triangleq \boldsymbol{\lambda} - \boldsymbol{\lambda}\odot(\boldsymbol{W}'\boldsymbol{\alpha})\oslash(\boldsymbol{W}'\mathbf{1}) \\ &+ \boldsymbol{\lambda}\odot[\boldsymbol{W}'(\boldsymbol{\alpha}\odot\boldsymbol{y}\oslash\hat{\boldsymbol{y}}(\boldsymbol{\lambda}))]\oslash[\boldsymbol{W}'\mathbf{1}]. \end{aligned}$$

The vector $\mathbf{1}$ is the $D\times 1$ vector of 1's.

Note that the fixed point(s) of G_{IA} and G_{IB} are identical, and if there were no attenuation, i.e. $\boldsymbol{\alpha} = \mathbf{1}$, then the two algorithms would be identical.

B. Convergence rates

Since the complete-data space $\boldsymbol{X}_{\text{IA}}$ contains the attenuated events N°_{db} that do not contribute to the measurements, conditions (ii)-(iv) of Theorem 2 are satisfied. Before invoking Theorem 2 to conclude that ML-IB is faster, we must verify condition (i) of Theorem 2, including strict concavity, convergence of the algorithms, and convergence to an interior point. Under the assumption of strict concavity, discussed further below, one can apply the same arguments as in the appendix of [4] to show that ML-IA and ML-IB converge globally to the same estimate. (Strict concavity is a sufficient condition for convergence, but it may not be necessary.) From (4) and (8),

$$L(\boldsymbol{\lambda}) = \log g(\boldsymbol{y};\boldsymbol{\lambda}) = \sum_{d=1}^{D}\left(-\hat{y}_d(\boldsymbol{\lambda}) + y_d\log(\hat{y}_d(\boldsymbol{\lambda}))\right), \tag{13}$$

so the Hessian of the likelihood is

$$\nabla^2 L(\boldsymbol{\lambda}) = -\boldsymbol{W}'\,\text{diag}\left\{\boldsymbol{y}\oslash\hat{\boldsymbol{y}}^2(\boldsymbol{\lambda})\odot\boldsymbol{\alpha}^2\right\}\boldsymbol{W}. \tag{14}$$

For strict concavity it is sufficient to have $\alpha_d y_d > 0, \forall d$, provided $\boldsymbol{W}$ has full column rank. In the presence of accidental coincidences, all PET detectors record nonzero coincidences with very high probability. For an appropriate sampling scheme and a well-designed PET system, $\boldsymbol{W}$ should have full column rank. If $\boldsymbol{W}$ is not full rank, such as when "too many" pixels are used, then the ML estimate is not unique, and the likelihood criterion is inappropriate. In such cases a penalized likelihood estimate is preferable, as we discuss in Section V. We conclude then, that if the $\boldsymbol{W}$ is full rank, and if the ML estimate is strictly positive, then the ML-IB algorithm converges faster than ML-IA.

Unfortunately, in most cases the ML estimate in PET will have components that are zero [14], i.e., not in the interior of $\mathcal{D}$. Strictly speaking, the above analysis is inconclusive for such examples. How likely is it that ML-IB, with its smaller complete-data space, would converge slower than ML-IA simply because some of the components converge to zero? In the next section we explore this question by considering a one-dimensional analogue.

C. Scalar example

One can obtain some insight into the convergence behavior of these two algorithms by considering the following scalar version of the problem. Suppose the measurement model is:

$$y \sim \mathrm{Poisson}(a\lambda + r)$$

where the attenuation $a \in (0,1)$ and the accidental coincidence rate $r \geq 0$ are known. In this case, the ML estimate for emission rate λ over $\mathcal{D} = \{\lambda : \lambda \geq 0\}$ is given by:

$$\hat{\lambda} = \max\left\{0, \frac{y-r}{a}\right\},$$

a truncated subtraction. The ML-IA and ML-IB algorithms are given respectively by the maps

$$G_{\mathrm{IA}}(\lambda) = (1-a)\lambda + \lambda a \frac{y}{a\lambda + r}$$

and

$$G_{\mathrm{IB}}(\lambda) = \lambda \frac{y}{a\lambda + r}.$$

Note that in the absence of attenuation ($a = 1$), the two algorithms are identical. One can also verify that both algorithms are globally convergent if $\lambda^0 > 0$. Differentiating:

$$\begin{aligned} \frac{d}{d\lambda} G_{\mathrm{IA}}(\lambda) &= (1-a) + \frac{ayr}{(a\lambda + r)^2} \\ \frac{d}{d\lambda} G_{\mathrm{IB}}(\lambda) &= \frac{yr}{(a\lambda + r)^2}, \end{aligned}$$

so in particular

$$\begin{aligned} \rho_{\mathrm{B}} = \frac{d}{d\lambda} G_{\mathrm{IB}}(\lambda)|_{\hat{\lambda}} &= \min\left\{\frac{r}{y}, \frac{y}{r}\right\} \\ \rho_{\mathrm{A}} = \frac{d}{d\lambda} G_{\mathrm{IA}}(\lambda)|_{\hat{\lambda}} &= 1 - a + a \frac{d}{d\lambda} G_{\mathrm{IB}}(\lambda)|_{\hat{\lambda}} \\ = 1 + a(\rho_{\mathrm{B}} - 1) &\geq \rho_{\mathrm{B}}, \end{aligned} \tag{15}$$

showing that the root-convergence factor for ML-IB is smaller than that of ML-IA. Does ML-IB converge faster? There are three cases to consider.

Case 1: If $y > r$, then both estimates converge to $\hat{\lambda} > 0$, at asymptotic rates governed by the Linear Convergence Theorem, so by (15), ML-IB converges faster.

Case 2: If $y \leq r$, then both estimates converge to $\hat{\lambda} = 0$, on the boundary of $\mathcal{D}$, so at first it seems that the Linear Convergence Theorem does not apply. However, if $r > 0$ then we can actually make the object domain slightly larger, say: $\mathcal{D}_- = \{\lambda : \lambda \geq -\frac{1}{2}r/a\}$, since G_{IA} and G_{IB} are both differentiable on $\mathcal{D}_-$. Directly applying[3] the Linear Convergence Theorem to G_{IA} and G_{IB} using (15) shows that if $y < r$, then ML-IB converges faster than ML-IA even though the ML estimate is 0!

Case 3: If $y = r$, then $\rho_{\mathrm{A}} = \rho_{\mathrm{B}} = 1$, so the asymptotic convergence rate is not well defined by the Linear Convergence Theorem. However, since y is an integer number of counts, and r is a real number, the outcome $y = r$ seems rather unlikely in practice. For a non-asymptotic comparison, one can verify that if $\lambda > 0$, then

$$|G_{\mathrm{IB}}(\lambda) - \hat{\lambda}| \leq |G_{\mathrm{IA}}(\lambda) - \hat{\lambda}|, \tag{16}$$

so the ML-IB algorithm takes larger steps towards the ML estimate than the ML-IA algorithm. Therefore, even though the convergence is sub-linear when $y = r$, the ML-IB algorithm will converge faster in a sub-linear sense.

In summary, we have shown that under this scalar model, ML-IB usually has faster asymptotic convergence rate than ML-IA, and always takes larger steps (16). It is difficult to predict what the analogous boundary situations would be in higher dimensions. The fact that there exists a situation where $\rho(\nabla G) = 1$ even in the scalar case suggests that a comprehensive rigorous comparison of ML-IB and ML-IA will be difficult to obtain.

IV. Joint estimation of α and λ

As in [1], the above discussion assumed that the survival probabilities α_d were known exactly. In practice, one estimates these survival probabilities from transmission measurements acquired prior to the emission scan. Let M_d denote the transmission measurement, with realization m_d, for the dth detector pair. It is reasonable to assume that the M_d's have independent Poisson distributions with mean $t_d\alpha_d$, where t_d is proportional to the transmission scan time[4] and the efficiency of the dth detector pair.

[3] We cannot apply Theorem 1 to this larger domain since Q_{IA} and Q_{IB} are not differentiable at 0.

[4] Here, we ignore the statistical uncertainty in the accidental coincidences measurement, the statistical uncertainty in the blank scan, and the contribution of accidental coincidences to the transmission measurements. For a more complete treatment, see [8].

The t_d factors are determined by a "blank scan," i.e. a transmission scan without the patient in the scanner. The conventional approach is simply to estimate α_d by m_d/t_d, and to "precorrect" the emission measurements. The approach suggested in [8] is to jointly estimate $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$ from $\boldsymbol{y}$ and $\boldsymbol{m} = [m_1, \ldots, m_D]'$ by maximizing the joint log-likelihood:

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= \log g(\boldsymbol{y}, \boldsymbol{m}; \boldsymbol{\lambda}, \boldsymbol{\alpha}) = \log g(\boldsymbol{y}; \boldsymbol{\lambda}, \boldsymbol{\alpha}) + \log g(\boldsymbol{m}; \boldsymbol{\alpha}) \\
&= \sum_{d=1}^{D} \left(-\hat{y}_d(\boldsymbol{\theta}) + y_d \log(\hat{y}_d(\boldsymbol{\theta})) - t_d\alpha_d + m_d \log(t_d\alpha_d)\right),
\end{aligned}
$$

where $\boldsymbol{\theta} = \{\boldsymbol{\lambda}, \boldsymbol{\alpha}\}$ and $\hat{y}_d(\boldsymbol{\theta}) = \sum_{b=1}^{B} \alpha_d w_{db}\lambda_b + r_d$. It follows that

$$
-\nabla^2 L(\boldsymbol{\theta}) =
$$
$$
\begin{bmatrix} \boldsymbol{W}' & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \left(\boldsymbol{K}(\boldsymbol{\theta}) + \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \operatorname{diag}\{\boldsymbol{m} \oslash \boldsymbol{\alpha}^2\} \end{bmatrix} \right) \begin{bmatrix} \boldsymbol{W} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix},
$$

where

$$
\boldsymbol{K}(\boldsymbol{\theta}) =
$$
$$
\begin{bmatrix} \operatorname{diag}\{\boldsymbol{y} \oslash \hat{\boldsymbol{y}}^2(\boldsymbol{\theta}) \odot \boldsymbol{\alpha}^2\} & \operatorname{diag}\{\boldsymbol{1} - \boldsymbol{y} \odot \boldsymbol{r} \oslash \hat{\boldsymbol{y}}^2(\boldsymbol{\theta})\} \\ \operatorname{diag}\{\boldsymbol{1} - \boldsymbol{y} \odot \boldsymbol{r} \oslash \hat{\boldsymbol{y}}^2(\boldsymbol{\theta})\} & \operatorname{diag}\{\boldsymbol{y} \oslash \hat{\boldsymbol{y}}^2(\boldsymbol{\theta}) \odot (\boldsymbol{W}\boldsymbol{\lambda})^2\} \end{bmatrix}.
$$

Thus, $L(\boldsymbol{\theta})$ is not necessarily strictly concave, and convergence of the two EM algorithms discussed below remains an open problem.

As in Section III, we again have the option of including or excluding the attenuated emissions N°_{db} from the complete-data space, leading to two algorithms we denote ML-JA and ML-JB.

For ML-JB, the complete data consists of

$$
\boldsymbol{X}_{\mathrm{JB}} = \{\{N_{db}\},\ \{R_d\},\ \{M_d\}\}, \quad \begin{array}{l} d = 1, \ldots, D \\ b = 1, \ldots, B \end{array},
$$

so

$$
\log f_{\mathrm{JB}}(\boldsymbol{X}_{\mathrm{JB}}; \bar{\boldsymbol{\theta}}) =
$$
$$
\sum_{d=1}^{D}\sum_{b=1}^{B} \left(-\bar{\alpha}_d w_{db}\bar{\lambda}_b + N_{db}\log(\bar{\alpha}_d w_{db}\bar{\lambda}_b)\right)
$$
$$
+ \sum_{d=1}^{D} \left(-r_d + R_d \log(r_d)\right) + \sum_{d=1}^{D} \left(-t_d\bar{\alpha}_d + M_d \log(t_d\bar{\alpha}_d)\right).
$$

It is easily verified that

$$
\begin{aligned}
E\{N_{db} | \boldsymbol{y}, \boldsymbol{m}; \boldsymbol{\theta}\} &= y_d \frac{\alpha_d w_{db}\lambda_b}{\hat{y}_d(\boldsymbol{\theta})}, \\
E\{R_d | \boldsymbol{y}, \boldsymbol{m}; \boldsymbol{\theta}\} &= y_d \frac{r_d}{\hat{y}_d(\boldsymbol{\theta})}, \\
E\{M_d | \boldsymbol{y}, \boldsymbol{m}; \boldsymbol{\theta}\} &= m_d;
\end{aligned}
$$

therefore,

$$
Q_{\mathrm{JB}}(\bar{\boldsymbol{\theta}}; \boldsymbol{\theta}) =
$$
$$
\sum_{d=1}^{D}\sum_{b=1}^{B} \left(-\bar{\alpha}_d w_{db}\bar{\lambda}_b + y_d \frac{\alpha_d w_{db}\lambda_b}{\hat{y}_d(\boldsymbol{\theta})} \log(\bar{\alpha}_d w_{db}\bar{\lambda}_b)\right)
$$
$$
+ \sum_{d=1}^{D} \left(-r_d + y_d \frac{r_d}{\hat{y}_d(\boldsymbol{\theta})} \log(r_d)\right) + \sum_{d=1}^{D} \left(-t_d\bar{\alpha}_d + m_d \log(t_d\bar{\alpha}_d)\right).
$$

Setting the derivatives of $Q_{\mathrm{JB}}(\cdot, \boldsymbol{\theta})$ to zero yields the following equations for the M-step:

$$
\begin{aligned}
0 &= \sum_{d=1}^{D} \left(-\bar{\alpha}_d w_{db} + y_d \frac{\alpha_d w_{db}\lambda_b}{\hat{y}_d(\boldsymbol{\theta})} / \bar{\lambda}_b\right), \\
0 &= \sum_{b=1}^{B} \left(-w_{db}\bar{\lambda}_b + y_d \frac{\alpha_d w_{db}\lambda_b}{\hat{y}_d(\boldsymbol{\theta})} / \bar{\alpha}_d\right) - t_d + m_d/\bar{\alpha}_d,
\end{aligned}
$$

yielding the pseudo-iteration:

$$\begin{aligned}\boldsymbol{\lambda}^{i+1} &= \boldsymbol{\lambda}^{i} \odot [\boldsymbol{W}'(\boldsymbol{\alpha}^{i} \odot \boldsymbol{y} \oslash \hat{\boldsymbol{y}}(\boldsymbol{\theta}^{i}))] \oslash [\boldsymbol{W}'\boldsymbol{\alpha}^{i+1}], \\ \boldsymbol{\alpha}^{i+1} &= [\boldsymbol{y} \oslash \hat{\boldsymbol{y}}(\boldsymbol{\theta}^{i}) \odot \boldsymbol{\alpha}^{i} \odot (\boldsymbol{W}\boldsymbol{\lambda}^{i}) + \boldsymbol{m}] \oslash [\boldsymbol{W}\boldsymbol{\lambda}^{i+1} + \boldsymbol{t}],\end{aligned}$$

where $\boldsymbol{t} = [t_1, \ldots, t_D]'$. This set of equations is coupled in $\boldsymbol{\lambda}^{i+1}$ and $\boldsymbol{\alpha}^{i+1}$, and no analytical solution seems likely. However, they can form the basis for a GEM algorithm [8, 13].

Fortunately, the equations become uncoupled when using the less parsimonious complete-data space for ML-JA. Let

$$\boldsymbol{X}_{\mathrm{JA}} = \{\{N_{db}\},\ \{R_d\},\ \{M_d\},\ \{N^{\circ}_{db}\}\},\ \begin{array}{l} d = 1, \ldots, D \\ b = 1, \ldots, B \end{array},$$

then

$$\log f_{\mathrm{JA}}(\boldsymbol{X}_{\mathrm{JA}}; \bar{\boldsymbol{\theta}}) = \log f_{\mathrm{JB}}(\boldsymbol{X}_{\mathrm{JB}}; \bar{\boldsymbol{\theta}}) + \sum_{d=1}^{D}\sum_{b=1}^{B} \left(-(1-\bar{\alpha}_d) w_{db} \bar{\lambda}_b + N^{\circ}_{db} \log((1-\bar{\alpha}_d) w_{db} \bar{\lambda}_b)\right).$$

Since N°_{db} is independent of the measurements,

$$E\{N^{\circ}_{db} | \boldsymbol{y}, \boldsymbol{m}; \boldsymbol{\theta}\} = (1-\alpha_d) w_{db} \lambda_b,$$

thus

$$Q_{\mathrm{JA}}(\bar{\boldsymbol{\theta}}; \boldsymbol{\theta}) = Q_{\mathrm{JB}}(\bar{\boldsymbol{\theta}}; \boldsymbol{\theta}) + \sum_{d=1}^{D}\sum_{b=1}^{B} \Big(-(1-\bar{\alpha}_d) w_{db} \bar{\lambda}_b + (1-\alpha_d) w_{db} \lambda_b \log((1-\bar{\alpha}_d) w_{db} \bar{\lambda}_b)\Big).$$

Setting the derivatives of $Q_{\mathrm{JA}}(\cdot, \boldsymbol{\theta})$ to zero yields the following equations for the M-step:

$$\begin{aligned}0 &= \sum_{d=1}^{D} \left(-\bar{\alpha}_d w_{db} + y_d \frac{\alpha_d w_{db} \lambda_b}{\hat{y}_d(\boldsymbol{\theta})} / \bar{\lambda}_b)\right) \\ &\quad + \sum_{d=1}^{D} \left(-(1-\bar{\alpha}_d) w_{db} + (1-\alpha_d) w_{db} \lambda_b / \bar{\lambda}_b)\right), \\ 0 &= \sum_{b=1}^{B} \left(-w_{db} \bar{\lambda}_b + y_d \frac{\alpha_d w_{db} \lambda_b}{\hat{y}_d(\boldsymbol{\theta})} / \bar{\alpha}_d)\right) \\ &\quad + (-t_d + m_d / \bar{\alpha}_d)) \\ &\quad + \sum_{b=1}^{B} \left(w_{db} \bar{\lambda}_b + (1-\alpha_d) w_{db} \lambda_b / (1-\bar{\alpha}_d)\right).\end{aligned}$$

This set of equations is uncoupled in $\boldsymbol{\lambda}^{i+1}$ and $\boldsymbol{\alpha}^{i+1}$, and yields the ML-JA iteration:

$$\begin{aligned}\boldsymbol{\lambda}^{i+1} &= \boldsymbol{\lambda}^{i} - \boldsymbol{\lambda}^{i} \odot (\boldsymbol{W}'\boldsymbol{\alpha}^{i}) \oslash (\boldsymbol{W}'\boldsymbol{1}) \\ &+ \boldsymbol{\lambda}^{i} \odot [\boldsymbol{W}'(\boldsymbol{\alpha}^{i} \odot \boldsymbol{y} \oslash \hat{\boldsymbol{y}}(\boldsymbol{\theta}^{i}))] \oslash [\boldsymbol{W}'\boldsymbol{1}],\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{t} &= [(\boldsymbol{W}\boldsymbol{\lambda}^{i}) \odot \boldsymbol{\alpha}^{i} \odot \boldsymbol{y} \oslash \hat{\boldsymbol{y}}(\boldsymbol{\theta}^{i}) + \boldsymbol{m}] \oslash \boldsymbol{\alpha}^{i+1} \\ &- [(\boldsymbol{W}\boldsymbol{\lambda}^{i}) \odot (\boldsymbol{1} - \boldsymbol{\alpha}^{i})] \oslash (\boldsymbol{1} - \boldsymbol{\alpha}^{i+1}),\end{aligned}$$

where the latter is an easily solved quadratic in $\boldsymbol{\alpha}^{i+1}$. The resulting joint emission/transmission estimation algorithm is only slightly more computationally expensive per iteration than the ML-IA or ML-IB algorithms of [1], yet it accounts for the statistical uncertainty in both the emission measurements and the transmission measurements, unlike ML-IA and ML-IB.

V. Discussion

We have shown that smaller complete-data spaces yield EM algorithms with faster asymptotic convergence. This theoretical result, combined with the empirical results in [1] suggests strongly that the ML-IB algorithm should be used in practice over the ML-IA algorithm. The heuristic explanation for this is that the complete-data space for ML-IA includes the attenuated events that make no contribution to the measurements. Since EM algorithms are notorious for slow convergence, this comparison has practical importance. Even a small decrease in the root-convergence factor can significantly reduce the required number of iterations.

We have also shown that the story gets more complicated if one wants to jointly estimate both the emission and the transmission parameters. In this case, although theoretically ML-JB would converge faster than ML-JA, the M-step of ML-JB seems intractable. We are currently investigating a space-alternating approach that may circumvent this problem [13]. Meanwhile, it appears that the best strategy is the following: *use the smallest complete-data space that results in a tractable maximization step*.

Several investigators have shown that more appealing images are produced by regularizing the ML estimate by including a penalty term or Bayesian "prior" [12, 15–18]. In principle, our Theorems 1 and 2 directly generalize to the case where concave penalties such as those discussed in [18] are added to the likelihood, again supporting the conclusion that smaller complete-data spaces correspond to faster convergence. There is one important caveat however: except in the trivial case of independent priors, the maximization steps of penalized EM algorithms become intractable due to the coupling introduced by the penalties. Consequently, the algorithms for the penalized case are usually of the generalized EM (GEM) type [3, 16]. GEM algorithms only provide an increase in $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^i)$ at each iteration, rather than truly maximizing Q. Therefore, GEM algorithms do not usually satisfy condition (ii) of our Theorem 1. They are also usually not globally convergent unless line-searches are employed [18]. These factors inhibit making formal statements about asymptotic convergence rates for penalized likelihood algorithms. We have implemented penalized-likelihood algorithms based on Hebert's GEM strategy [16] for both the ML-IA and ML-IB complete-data spaces. We have also implemented both ML-IA and ML-IB with sieve constraints [1, 8, 19]. We found empirically that the penalized ML-IB algorithm converged substantially more rapidly, in terms of both likelihood increase and apparent image contrast. These empirical results are further motivation for using smaller complete-data spaces where possible.

Acknowledgement

The authors gratefully acknowledge discussions with P. Chiao and A. Hero.

References

[1] D. G. Politte and D. L. Snyder. Corrections for accidental coincidences and attenuation in maximum-likelihood image reconstruction for positron-emission tomography. *IEEE Tr. Med. Im.*, 10(1):82–9, March 1991.

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

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc. Ser. B*, 39(1):1–38, 1977.

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

[5] M. I. Miller, D. L. Snyder, and T. R. Miller. Maximum-likelihood reconstruction for single-photon emission computed-tomography. *IEEE Tr. Nuc. Sci.*, 32(1):769–78, February 1985.

[6] N. H. Clinthorne, X. H. Wang, and J. A. Fessler. Multi-energy maximum-likelihood reconstruction algorithms for SPECT and PET. *J. Nuc. Med. (Abs. Book)*, 33(5):831, May 1992.

[7] M. E. Daube-Witherspoon, R. E. Carson, Y. Yan, and T. K. Yap. Scatter correction in maximum likelihood reconstruction of PET data. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 2, pages 945–947, 1992.

[8] N. H. Clinthorne, J. A. Fessler, G. D. Hutchins, and W. L. Rogers. Joint maximum likelihood estimation of emission and attenuation densities in PET. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 3, pages 1927–1932, 1991.

[9] M. Segal and E. Weinstein. The cascade EM algorithm. *Proc. IEEE*, 76(10):1388–1390, October 1988.

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

[11] E. L. Lehmann. *Theory of point estimation*. Wiley, New York, 1983.

[12] P. J. Green. On use of the EM algorithm for penalized likelihood estimation. *J. Royal Stat. Soc. Ser. B*, 52(3):443–452, 1990.

[13] J. A. Fessler and A. O. Hero. Complete-data spaces and generalized EM algorithms. In *Proc. IEEE Conf. Acoust. Speech Sig. Proc.*, volume 4, pages 1–4, 1993.

[14] C. L. Byrne. Iterative image reconstruction algorithms based on cross-entropy minimization. *IEEE Tr. Im. Proc.*, 2(1):96–103, January 1993. Erratum and addendum: 4(2):226-7, Feb. 1995.

[15] 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.*, pages 12–8, 1985.

[16] T. Hebert and R. Leahy. A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors. *IEEE Tr. Med. Im.*, 8(2):194–202, June 1989.

[17] C. T. Chen, V. E. Johnson, W. H. Wong, X. Hu, and C. E. Metz. Bayesian image reconstruction in positron emission tomography. *IEEE Tr. Nuc. Sci.*, 37(2):636–641, April 1990.

[18] K. Lange. Convergence of EM image reconstruction algorithms with Gibbs smoothing. *IEEE Tr. Med. Im.*, 9(4):439–46, December 1990. Corrections, T-MI, 10:2(288), June 1991.

[19] 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 Tr. Med. Im.*, 6(3):228–38, September 1987.