Hidden-Data Spaces for Maximum-Likelihood PET Reconstruction

Jeffrey A. Fessler*
Division of Nuclear Medicine, University of Michigan

ABSTRACT

As investigators consider more comprehensive measurement models for emission tomography, there will be more choices for the *hidden* or *complete* data spaces of the associated expectation-maximization (EM) algorithms. In this paper, we show that EM algorithms based on smaller complete data spaces will typically converge faster. As a example, we compare the two maximum-likelihood (ML) 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).

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 *hidden*, or *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, there will be more choices as investigators consider more comprehensive measurement models, such as those accounting for photon attenuation [4], accidental coincidences [1], deadtime, and scatter [5]. This paper illustrates the importance of parsimony in choosing complete-data spaces.

Recently, Politte and Snyder proposed two ML-EM algorithms for PET that directly incorporate the affects of 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. Due to space constraints, all proofs are omitted, and a simplified version of the PET application is presented.

*This work was supported in part by by a DOE Alexander Hollaender Postdoctoral Fellowship, and by DOE Grant DE-FG02-87ER65061.

0-7803-0883-2/93/$3.00 ©1993 IEEE

II. THEORY

We observe $\mathbf{y}$, a realization of a random vector $\mathbf{Y}$ having density $g(\mathbf{y};\boldsymbol{\theta})$ with the goal of computing the ML estimate of $\boldsymbol{\theta}$. When the measurements are "incomplete," one can often postulate a "complete data" random vector $\mathbf{X}$ having density $f(\mathbf{x};\boldsymbol{\theta})$ that is more naturally related to $\boldsymbol{\theta}$, and is related to the observed measurements by a many-to-one mapping $\mathbf{Y}= \mathrm{h}(\mathbf{X})$. Let

$$Q(\bar{\boldsymbol{\theta}};\boldsymbol{\theta}) \triangleq E\left\{\log f(\mathbf{X};\bar{\boldsymbol{\theta}}) \,|\mathbf{Y}=\mathbf{y};\boldsymbol{\theta}\right\} \quad (1)$$

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

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

where

$$H(\bar{\boldsymbol{\theta}};\boldsymbol{\theta}) \triangleq E\left\{\log f(\mathbf{X}|\mathbf{Y}=\mathbf{y};\bar{\boldsymbol{\theta}}) \,|\mathbf{Y}=\mathbf{y};\boldsymbol{\theta}\right\},$$

$$L(\boldsymbol{\theta}) \triangleq \log g(\mathbf{y};\boldsymbol{\theta}).$$

The EM algorithm [3] repeats these 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]:

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

so an 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.

Several investigators have observed empirically that larger complete-data spaces correspond to slower EM convergence [3, pp. 25,34]. In this section we formalize a version of this result. Here, asymptotic convergence rate is defined by the following theorem [6, p. 301].

Linear Convergence Theorem: *If (i)* $G : \mathcal{D} \in \Re^n \to \Re^n$ *has a fixed point* $\boldsymbol{\theta}^* \in \mathcal{D}_+ = \mathrm{int}(\mathcal{D})$, *(ii)* G *is Fréchet differentiable at* $\boldsymbol{\theta}^*$, *and (iii)* $\rho(\nabla^1 G(\boldsymbol{\theta}^*)) < 1$, *where* $\rho()$ *denotes*

spectral radius, then the root-convergence factor [6, p. 288] R_1 at θ^* for the iterative process $\theta^{i+1} = G(\theta^i)$ is given by $R_1(G, \theta^*) = \rho(\nabla^1 G(\theta^*))$.

This theorem leads to the following [7]:

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

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

where $\mathbf{I}$ is the $n \times n$ identity matrix and $\mathbf{Q} \triangleq -\nabla^{20} Q(\theta^*; \theta^*)$.

Since $\mathbf{Q} = \mathbf{H} + \mathbf{L}$, where $\mathbf{H} \triangleq -\nabla^{20} H(\theta^*; \theta^*)$ is a conditional Fisher information matrix, one sees from (3) that if a larger complete-data space has greater 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) $\mathbf{Q}_\mathrm{B} = \mathbf{H} + \mathbf{L}$, where $\mathbf{H}$ is symmetric nonnegative definite and $\mathbf{L}$ is symmetric positive definite, and (ii) $\mathbf{Q}_\mathrm{A} = \mathbf{Q}_\mathrm{B} + \mathbf{N}$ where $\mathbf{N}$ is symmetric nonnegative definite, then $\rho_\mathrm{B} \leq \rho_\mathrm{A}$, where $\rho_\mathrm{A} = \rho\left(\mathbf{I} - \mathbf{Q}_\mathrm{A}^{-1}\mathbf{L}\right)$ and $\rho_\mathrm{B} = \rho\left(\mathbf{I} - \mathbf{Q}_\mathrm{B}^{-1}\mathbf{L}\right)$. Furthermore, if $\mathbf{N}$ is symmetric positive definite, then $\rho_\mathrm{B} < \rho_\mathrm{A}$.*

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

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

In [1], Politte and Snyder developed two ML algorithms for PET image reconstruction accounting for attenuation and accidental coincidences. One algorithm, called ML-IB, was based on a complete-data space consisting of the coincidence contributions of each voxel to each detector, as well as the accidental coincidences. The complete-data space for the other algorithm, called ML-IA, consisted of the ML-IB components as well as the photon emissions that are attenuated. These latter events are unobservable, so they are extraneous in the sense of Theorem 2. Unfortunately, Theorem 2 is not strictly applicable to PET because condition (i) of the Linear Convergence Theorem requires convergence to an interior point, and ML estimates in PET usually include zero components.

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

$$\begin{aligned} \rho_\mathrm{B} = \frac{d}{d\lambda} G_\mathrm{IB}(\lambda)|_{\hat{\lambda}} &= \min\left\{\frac{r}{y}, \frac{y}{r}\right\} \\ \rho_\mathrm{A} = \frac{d}{d\lambda} G_\mathrm{IA}(\lambda)|_{\hat{\lambda}} &= 1 - a + a \frac{d}{d\lambda} G_\mathrm{IB}(\lambda)|_{\hat{\lambda}} \\ = 1 + a(\rho_\mathrm{B} - 1) &\geq \rho_\mathrm{B}, \end{aligned} \tag{4}$$

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 (4), 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[1] the Linear Convergence Theorem to G_IA and G_IB using (4) shows

[1] We cannot apply Theorem 1 to this larger domain since Q_IA and Q_IB are not differentiable at 0.

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 $\rho_A = \rho_B = 1$, 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

$$|G_{IB}(\lambda) - \hat{\lambda}| \leq |G_{IA}(\lambda) - \hat{\lambda}|, \quad (5)$$

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 some 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 (5). 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. Discussion

We have shown that smaller complete-data choices yield EM algorithms with faster asymptotic convergence, provided the ML estimate lies in the interior of the parameter space. 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. Although this theoretical result is not always directly applicable to the ML-IB and ML-IA algorithms for PET, due to the boundary conditions, the basic concept should provide useful insight for choosing complete-data spaces. The fact that the theoretical result agrees with the empirical findings in [1] suggests strongly to us 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. We would be most interested in any counter-examples in which an EM algorithm based on a complete-data space with larger Fisher information converges to a boundary estimate faster than an EM algorithm having smaller Fisher information.

In principle, our Theorems 1 and 2 directly generalize to the case where convex penalties such as those discussed in [8] 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. As a consequence, the algorithms for the penalized case are usually of the generalized EM (GEM) type [3,9]. GEM algorithms only provide an increase in $Q(\theta, \theta^i)$ at each iteration, rather than maximizing Q. Therefore, GEM algorithms do not usually satisfy condition (ii) of our Theorem 1. They are also usually not globally convergent unless possibly computationally expensive line-searches are employed [8]. 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 [9] 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]. In both cases, we found empirically that the regularized ML-IB algorithm converged substantially more rapidly, in terms of both likelihood increase and apparent image contrast, than the regularized ML-IA algorithm. This empirical result is further motivation for using smaller complete-data spaces where possible.

V. Acknowledgement

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

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 Transactions on Medical Imaging*, 10(1):82–89, March 1991.

[2] L. A. Shepp and Y. Vardi. Maximum likelihood reconstruction for emission tomography. *IEEE Transactions on Medical Imaging*, 1(2):113–122, October 1982.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39(1):1–38, 1977.

[4] M. I. Miller, D. L. Snyder, and T. R. Miller. Maximum-likelihood reconstruction for single-photon emission computed-tomography. *IEEE Transactions on Nuclear Science*, 32(1):769–778, February 1985.

[5] N. H. Clinthorne, X. H. Wang, and J. A. Fessler. Multi-energy maximum-likelihood reconstruction algorithms for SPECT and PET. *Journal of Nuclear Medicine (Abstract Book)*, 33(5):831, 1992.

[6] J. M. Ortega and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York, 1970.

[7] P. J. Green. On use of the EM algorithm for penalized likelihood estimation. *Journal of the Royal Statistical Society Series B*, 52(3):443–452, 1990.

[8] K. Lange. Convergence of EM image reconstruction algorithms with Gibbs smoothing. *IEEE Transactions on Medical Imaging*, 9(4):439–446, December 1990. Corrections, June 1991 TMI.

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