COMPLETE-DATA SPACES AND GENERALIZED EM ALGORITHMS

J. A. Fessler and A. O. Hero
University of Michigan

ABSTRACT

Expectation-maximization (EM) algorithms have been applied extensively for computing maximum-likehood and penalized-likelihood parameter estimates in signal processing applications. Intrinsic to each EM algorithm is a complete-data space (CDS)—a hypothetical set of random variables that is related to the parameters more naturally than the measurements are. This paper describes two generalizations of the EM paradigm: (i) allowing the relationship between the CDS and the measured data to be nondeterministic, and (ii) using a sequence of alternating complete-data spaces. These generalizations are motivated in part by the influence of the CDS on the convergence rate, a relationship that we formalize through a data-processing inequality for Fisher Information. We apply these concepts to the problem of estimating superimposed signals in Gaussian noise, and demonstrate that the new space-alternating generalized EM algorithm converges significantly faster than the ordinary EM algorithm.

INTRODUCTION

Signal processing applications of EM algorithms for computing maximum-likehood (ML) parameter estimates have included tomography, image restoration, and estimation of superimposed signals [1–5]. Intrinsic to an EM algorithm is the notion of a complete-data space (CDS), which is a hypothetical set of random variables that, had they been measured, would have facilitated parameter estimation [1]. The conventional EM algorithm requires that the complete-data space be larger than the measurement space in the sense that every point in the CDS determines a point in the original measurement space via a deterministic functional relation. EM algorithms are notorious for slow convergence, and the choice of CDS is affects the convergence rate. In this paper we describe two generalizations of the EM algorithm, and we establish a formal relation between the EM convergence rate and the conditional Fisher information of the CDS given the observations.

This work was supported in part by a DOE Alexander Hollaender Postdoctoral Fellowship, by DOE grant DE-FG02-87ER65061, and by the National Science Foundation under grant BCS-9024370.

The two generalizations are (i) allowing the relationship between the CDS and the measurements to be nondeterministic, and (ii) a "space alternating" generalized EM algorithm in which multiple complete-data spaces are used iteratively. These generalized EM algorithms allow more flexibility in algorithm implementation and can converge faster than the conventional EM algorithm. The convergence rates of these algorithms decrease as the difference between the Fisher information matrices associated with the CDS and the original data space increases. Therefore, given two possible choices of CDS, the one having smaller Fisher information is a better choice in terms of EM convergence rate. While a larger CDS may simplify the implementation of an EM algorithm, using a new *data-processing inequality* we show that it also has larger associated Fisher information and therefore slows the convergence of the algorithm.

This work has been motivated by applications in emission tomography [3–5] and in superimposed signals estimation [2]. In this summary we focus on the latter application, and show that not only can the asymptotic convergence rate be improved by using a space-alternating method, but also that using a smaller CDS leads to estimates that are closer to the ML estimate at every iteration. This non-asymptotic result further highlights the practical importance of consideration of the size of the CDS in terms of computational requirements.

GENERALIZED FORM EM ALGORITHM

Given a measurement $\mathbf{y}$, a realization of a random vector $\mathbf{Y}$ with density $g(\mathbf{y};\boldsymbol{\theta})$, our goal is to compute the ML estimate of $\boldsymbol{\theta}$. In many problems, direct maximization of g over $\boldsymbol{\theta}$ is impractical.[2] Our first generalization of the EM algorithm requires the following definition.

Definition 1. *A random vector $\mathbf{X}$ with density $f(\mathbf{x};\boldsymbol{\theta})$ is* an admissible *CDS for $g(\mathbf{y};\boldsymbol{\theta})$ if the joint density of $\mathbf{X}$ and $\mathbf{Y}$ satisfies*

$$f(\mathbf{y},\mathbf{x};\boldsymbol{\theta}) = f(\mathbf{y}|\mathbf{x})f(\mathbf{x};\boldsymbol{\theta}), \qquad (1)$$

[2] In the missing-data statistical problems that motivated [1], direct maximization was difficult because of the incompleteness of the actual measurements. The terms "complete" and "incomplete" are less natural for most signal processing applications, but we adhere to this standard terminology.

0-7803-0946-4/93 $3.00

where $f(\mathbf{y}|\mathbf{x})$ is independent of $\boldsymbol{\theta}$.

The conditional density $f(\mathbf{y}|\mathbf{x})$ may include Dirac delta functions (as addressed in [6]). Thus (1) reduces to the conventional CDS definition when $\mathbf{Y}$ is a *deterministic* function of $\mathbf{X}$. The generalization (1) offers more flexibility in the choice of CDS, and is more natural for some signal processing applications with additive noise.

Having identified an admissible CDS $\mathbf{X}$, define the following conditional expectation and apply Bayes' rule:

$$Q(\boldsymbol{\theta};\bar{\boldsymbol{\theta}}) \triangleq E\left\{\log f(\mathbf{X};\boldsymbol{\theta}) \,|\mathbf{Y}=\mathbf{y};\bar{\boldsymbol{\theta}}\right\} \tag{2}$$

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

$$= H(\boldsymbol{\theta};\bar{\boldsymbol{\theta}}) + L(\boldsymbol{\theta}) - W(\bar{\boldsymbol{\theta}}), \tag{3}$$

where

$$\begin{aligned} H(\boldsymbol{\theta};\bar{\boldsymbol{\theta}}) &\triangleq E\left\{\log f(\mathbf{X}|\mathbf{Y}=\mathbf{y};\boldsymbol{\theta}) \,|\mathbf{Y}=\mathbf{y};\bar{\boldsymbol{\theta}}\right\}, \\ L(\boldsymbol{\theta}) &\triangleq \log g(\mathbf{y};\boldsymbol{\theta}), \\ W(\bar{\boldsymbol{\theta}}) &\triangleq \int \log f(\mathbf{y}|\mathbf{x})\, f(\mathbf{x}|\mathbf{Y}=\mathbf{y};\bar{\boldsymbol{\theta}})\, d\mathbf{x}. \end{aligned}$$

The generalized CDS (1) influences both H and W. However, since $W(\bar{\boldsymbol{\theta}})$ is independent of $\boldsymbol{\theta}$, the form of the "M-step" of the conventional EM algorithm [1] is unaffected, so we adopt the same two-step iteration as in [1]:

E-step:

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

M-step:

$$\boldsymbol{\theta}^{i+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta};\boldsymbol{\theta}^i), \tag{4}$$

where $\boldsymbol{\theta}^i$ denotes the parameter estimate after the ith iteration. Note that $Q(\boldsymbol{\theta}^{i+1};\boldsymbol{\theta}^i) \geq Q(\boldsymbol{\theta}^i;\boldsymbol{\theta}^i)$ implies that

$$L(\boldsymbol{\theta}^{i+1}) - L(\boldsymbol{\theta}^i) \geq H(\boldsymbol{\theta}^i;\boldsymbol{\theta}^i) - H(\boldsymbol{\theta}^{i+1};\boldsymbol{\theta}^i) = D(\boldsymbol{\theta}^{i+1}\|\boldsymbol{\theta}^i),$$

where

$$D(\boldsymbol{\theta}\|\bar{\boldsymbol{\theta}}) \triangleq \int \log \frac{f(\mathbf{x}|\mathbf{y};\bar{\boldsymbol{\theta}})}{f(\mathbf{x}|\mathbf{y};\boldsymbol{\theta})} f(\mathbf{x}|\mathbf{y};\bar{\boldsymbol{\theta}}) \geq 0,$$

denotes the nonnegative Kullback-Liebler distance [7]. Therefore (4) produces a monotonically increasing likelihood sequence.

EM CONVERGENCE RATE

In this section we formalize the relationship between Fisher Information and the convergence rate of the EM algorithm. Full proofs can be found in [6]. For our purposes, the asymptotic convergence rate is defined by the R_1 root-convergence factor [8]. The arguments in [1] for the conventional EM algorithm apply directly to (4), and show the following.

Theorem 1: *If $\mathbf{X}$ is a CDS inducing an EM algorithm whose sequence of estimates converges to $\boldsymbol{\theta}^*$, then the root-convergence factor is given by:*

$$\rho_{\mathbf{X}} = \rho\left(\mathbf{I} - (\mathbf{H}_{\mathbf{X}|\mathbf{y}} + \mathbf{L})^{-1}\mathbf{L}\right) < 1, \tag{5}$$

where $\mathbf{I}$ is the identity matrix, $\rho()$ denotes spectral radius (largest absolute eigenvalue), $\mathbf{L} \triangleq -\nabla^2 L(\boldsymbol{\theta}^*)$, and $\mathbf{H}_{\mathbf{X}|\mathbf{y}} \triangleq -\nabla^{20} H(\boldsymbol{\theta}^*;\boldsymbol{\theta}^*)$ is the nonnegative definite conditional Fisher information matrix [9, p. 126].

The following lemma links the "size" of $\mathbf{H}_{\mathbf{X}|\mathbf{y}}$ to $\rho_{\mathbf{X}}$.

Lemma 1: *If $\mathbf{H}_1$ and $\mathbf{H}_2$ are nonnegative definite, $\mathbf{L}$ is positive definite, and $\mathbf{H}_1 \geq \mathbf{H}_2$ (i.e. $\mathbf{H}_1 - \mathbf{H}_2$ is nonnegative definite), then*

$$\rho\left(\mathbf{I} - (\mathbf{H}_1 + \mathbf{L})^{-1}\mathbf{L}\right) \geq \rho\left(\mathbf{I} - (\mathbf{H}_2 + \mathbf{L})^{-1}\mathbf{L}\right).$$

Finally, two admissible complete-data spaces $\mathbf{X}_1$ and $\mathbf{X}_2$ can be compared using the following theorem, which is a Fisher Information version of the *data-processing inequality.*

Theorem 2: *If $\mathbf{X}_1$ and $\mathbf{X}_2$ are each an admissible CDS (1), and if their joint density satisfies $f(\mathbf{X}_2,\mathbf{X}_1|\mathbf{y};\boldsymbol{\theta}) = f(\mathbf{X}_2|\mathbf{X}_1,\mathbf{y}) f(\mathbf{X}_1|\mathbf{y};\boldsymbol{\theta})$ where $f(\mathbf{X}_1|\mathbf{X}_2,\mathbf{y})$ is independent of $\boldsymbol{\theta}$, then $\mathbf{H}_{\mathbf{X}_2|\mathbf{y}} \leq \mathbf{H}_{\mathbf{X}_1|\mathbf{y}}$, so from Lemma 1, $\rho_{\mathbf{X}_2} \leq \rho_{\mathbf{X}_1}$.*

In other words, if $\mathbf{X}_2$ is less informative about $\boldsymbol{\theta}$ than $\mathbf{X}_1$, then the EM algorithm for CDS $\mathbf{X}_2$ converges faster.

SPACE-ALTERNATING GEM (SAGE)

The above analysis strongly suggests that minimizing the information of the CDS is essential for improving convergence rate. In many applications, including most penalized-likelihood algorithms, one implements a generalized expectation-maximization (GEM) rather than a pure EM algorithm [1]. GEM methods typically involve updating parameters in small groups while holding the others fixed, rather than updating all parameters simultaneously. The conventional EM or GEM method uses the same CDS for each update. We propose instead to extend the GEM algorithm by relaxing this restriction, allowing the update for each parameter group to correspond to a different CDS. Since the CDS for each group of parameters can often be made smaller than the CDS necessary for the entire parameter space, the resulting algorithms converge faster. Such a space-alternating generalized EM (SAGE) algorithm will also monotonically increase the likelihood.

One application of the SAGE algorithm is in joint estimation of emission and transmission parameters in PET [5]. Here, as a concise illustration of the SAGE method, we consider the following superimposed signal estimation problem:

$$\mathbf{Y} = \mathbf{A}_1\theta_1 + \mathbf{A}_2\theta_2 + \epsilon, \tag{6}$$

where ϵ is additive zero-mean Gaussian noise with covariance $\boldsymbol{\Sigma}$. Let $\mathbf{X}_\beta$ be a family of multivariate Gaussian distributions:

$$\mathbf{X}_\beta \sim N\left(\begin{bmatrix} \mathbf{A}_1\theta_1 \\ \mathbf{A}_2\theta_2 \end{bmatrix}, \begin{bmatrix} \beta\boldsymbol{\Sigma} & 0 \\ 0 & (1-\beta)\boldsymbol{\Sigma} \end{bmatrix}\right), \tag{7}$$

where $\beta \in [0,1]$. Then by letting $\mathbf{Y} = [\mathbf{I}\ \mathbf{I}]\mathbf{X}_\beta$, we see that each $\mathbf{X}_\beta$ is an admissible CDS for (6). Thus there is a continuum of admissible complete-data spaces. The conventional EM algorithm for the CDS $\mathbf{X}_\beta$ can be expressed as the following simultaneous update:

$$\begin{aligned} \hat{\epsilon} &= \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{A}_1\theta_1^i - \mathbf{A}_2\theta_2^i) \\ \theta_1^{i+1} &= \theta_1^i + \beta(\mathbf{A}_1'\boldsymbol{\Sigma}^{-1}\mathbf{A}_1)^{-1}\mathbf{A}_1'\hat{\epsilon} \qquad (8) \\ \theta_2^{i+1} &= \theta_2^i + (1-\beta)(\mathbf{A}_2'\boldsymbol{\Sigma}^{-1}\mathbf{A}_2)^{-1}\mathbf{A}_2'\hat{\epsilon}. \qquad (9) \end{aligned}$$

When deriving the root-convergence factor for this iteration, one finds that the optimal β is $1/2$ (consistent with the intuitive choice made in [2]). In that case $\rho_{\mathbf{X}_\beta} = (1+\cos\phi)/2 = \cos^2(\phi/2)$, where $\cos\phi$ is the cosine of the complementary angle $\phi \in [0, \pi/2]$ between the signal subspaces spanned by $\mathbf{A}_1$ and $\mathbf{A}_2$.

In contrast, our SAGE method alternates between using CDS $\mathbf{X}_1$ for updating θ_1 and CDS $\mathbf{X}_0$ for updating θ_2, which corresponds to using the minimally informative choices. The algorithm is:

$$\begin{aligned} \hat{\epsilon} &= \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{A}_1\theta_1^i - \mathbf{A}_2\theta_2^i) \\ \theta_1^{i+1} &= \theta_1^i + (\mathbf{A}_1'\boldsymbol{\Sigma}^{-1}\mathbf{A}_1)^{-1}\mathbf{A}_1'\hat{\epsilon} \\ \hat{\epsilon} &= \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{A}_1\theta_1^{i+1} - \mathbf{A}_2\theta_2^i) \\ \theta_2^{i+1} &= \theta_2^i + (\mathbf{A}_2'\boldsymbol{\Sigma}^{-1}\mathbf{A}_2)^{-1}\mathbf{A}_2'\hat{\epsilon}. \qquad (10) \end{aligned}$$

Here, the root-convergence factor is $\cos^2\phi$, which is less than $\cos^2\phi$. Comparing (9) and (10), one sees that the step-size for the SAGE algorithm is larger than the conventional EM algorithm since $\beta \in [0,1]$. This is reflected in the root-convergence factors, which are illustrated in Figure 1. The convergence rate for SAGE is siginficantly faster than that of the EM algorithm. For this Gaussian model, the SAGE algorithm is equivalent to alternating projections [10].

GAUSSIAN NON-ASYMPTOTICS

Is our discussion of asymptotic convergence rates relevant to algorithms that are terminated after a finite number of iterations? In this section we show three *non-asymptotic* advantages of using a smaller CDS for the linear Gaussian model:

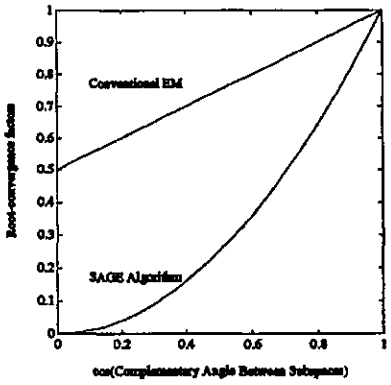

Figure 1: Comparison of root-convergence factors for conventional EM algorithm and proposed SAGE algorithm versus complementary angle between subspaces. The SAGE algorithm has a significantly improved convergence rate.

$$\mathbf{Y} \sim N(\mathbf{A}\theta, \boldsymbol{\Pi}).$$

The goal is to compute the ML estimate of θ from $\mathbf{Y}$ using an iterative method that avoids directly inverting $\mathbf{A}'\boldsymbol{\Pi}^{-1}\mathbf{A}$. Assume $\mathbf{Y} \in \Re^{N_Y}$, $\theta \in \Re^{N_\theta}$, $\mathbf{A} \in \Re^{N_Y \times N_\theta}$, $N_Y \geq N_\theta$, and that $\mathbf{A}$ has full column rank N_θ.

When viewed as an incomplete-data problem, there are a multitude of admissible complete-data spaces that can be applied with different resulting EM iterations. Suppose $\mathbf{X}_1$ and $\mathbf{X}_2$ are each an admissible CDS for this problem, with distributions:

$$\mathbf{X}_k \sim N(\mathbf{C}_k\theta, \boldsymbol{\Sigma}_k),\ k = 1, 2,$$

where $\mathbf{X}_k \in \Re^{N_k}$, $\mathbf{C}_k \in \Re^{N_k \times N_\theta}$. Also, assume $\mathbf{X}_2$ is "smaller" in the sense that

$$\mathbf{X}_2 = \mathbf{G}\,\mathbf{X}_1, \tag{11}$$

where $\mathbf{G} \in \Re^{N_2 \times N_1}$ and $N_2 \leq N_1$. Of course, for $\mathbf{X}_k$ to be "complete," we assume that $N_k \geq N_\theta$, $k = 1, 2$, and that and that $\mathbf{C}_k$ has full column rank N_θ.

For $\mathbf{X}_k$ to be admissible, it suffices for there to be $\mathbf{B}_k$ and $\mathbf{N}_k$ such that

$$\mathbf{Y} = \mathbf{B}_k\mathbf{X}_k + \mathbf{N}_k, \tag{12}$$

where $\mathbf{B}_k \in \Re^{N_Y \times N_k}$ is independent of θ, $\mathbf{N}_k \sim N(0, \boldsymbol{\Pi} - \mathbf{B}_k'\boldsymbol{\Sigma}_k\mathbf{B}_k)$, and $\mathbf{N}_k$ and $\mathbf{X}_k$ are uncorrelated

[6]. Under the additional assumption of linearity between $\mathbf{Y}$, $\mathbf{X}$, and $\boldsymbol{\theta}$, condition (12) is also necessary. Using the properties of conditional normal distributions, one can derive the following EM algorithm:

$$\boldsymbol{\theta}_k^{i+1} = \boldsymbol{\theta}_k^i + (\mathbf{C}_k' \boldsymbol{\Sigma}_k^{-1} \mathbf{C}_k)^{-1} \mathbf{A}' \boldsymbol{\Pi}^{-1} (\mathbf{y} - \mathbf{A}\boldsymbol{\theta}_k^i).$$

First note that it is clear from this recursion that the asymptotic convergence rate is

$$\rho(\mathbf{I} - (\mathbf{F}_{\mathbf{X}_k|\mathbf{y}} + \mathbf{F}_{\mathbf{Y}})^{-1} \mathbf{F}_{\mathbf{Y}}) = \rho(\mathbf{I} - \mathbf{F}_{\mathbf{X}_k}^{-1} \mathbf{F}_{\mathbf{Y}}), \quad (13)$$

where $\mathbf{F}_{\mathbf{X}_k} = \mathbf{C}_k' \boldsymbol{\Sigma}_k^{-1} \mathbf{C}_k$ and $\mathbf{F}_{\mathbf{Y}} = \mathbf{A}' \boldsymbol{\Pi}^{-1} \mathbf{A}$ are the Fisher information matrices of $\mathbf{X}_k$ and $\mathbf{Y}$ respectively. (Note that these matrices are independent of $\boldsymbol{\theta}$ and $\mathbf{Y}$.) From (11) and Theorems 1 and 2 it follows that $\mathbf{F}_{\mathbf{X}_2|\mathbf{y}} \leq \mathbf{F}_{\mathbf{X}_1|\mathbf{y}}$ and the asymptotic convergence rate of the EM algorithm for to $\mathbf{X}_2$ is faster than that of $\mathbf{X}_1$. Note that here we also have $\mathbf{F}_{\mathbf{X}_2} \leq \mathbf{F}_{\mathbf{X}_1}$, so the faster algorithm corresponds to smaller (unconditional) Fisher Information as well. In the remainder we focus on the early iterations.

Theorem 3: $\|\boldsymbol{\theta}_2^1 - \boldsymbol{\theta}^0\| \geq \|\boldsymbol{\theta}_1^1 - \boldsymbol{\theta}^0\|$, *i.e., the EM algorithm for the smaller CDS takes a bigger first step.*

Proof: $\boldsymbol{\theta}_k^1 - \boldsymbol{\theta}^0 = \mathbf{F}_{\mathbf{X}_k}^{-1} [\mathbf{A}' \boldsymbol{\Pi}^{-1} (\mathbf{y} - \mathbf{A}\boldsymbol{\theta}^0)]$, so it suffices to show that $\mathbf{F}_{\mathbf{X}_2}^{-1} \geq \mathbf{F}_{\mathbf{X}_1}^{-1}$. This follows from $\mathbf{F}_{\mathbf{X}_2} \leq \mathbf{F}_{\mathbf{X}_1}$ using positive definiteness. □

So the iterates for the smaller CDS take the lead from the starting line. Do they stay ahead? The next two theorems confirm that they do. Defining the whitened residual: $\boldsymbol{\epsilon}_k^i = \boldsymbol{\Pi}^{-1/2} (\mathbf{y} - \mathbf{A}\boldsymbol{\theta}_k^i)$, it is easily verified that $\boldsymbol{\epsilon}_k^{i+1} = \mathbf{M}_k \boldsymbol{\epsilon}_k^i$, for the "transition matrix"

$$\mathbf{M}_k = \mathbf{I} - \boldsymbol{\Pi}^{-1/2} \mathbf{A} \mathbf{F}_{\mathbf{X}_k}^{-1} \mathbf{A}' \boldsymbol{\Pi}^{-1/2}.$$

For the normal model, smaller residual norm corresponds to higher likelihood.

Theorem 4: *For $\mathbf{X}_1$ and $\mathbf{X}_2$ defined above and leading to the transition matrices $\mathbf{M}_1$ and $\mathbf{M}_2$, $\|\mathbf{M}_2^i \boldsymbol{\epsilon}^0\| \leq \|\mathbf{M}_1^i \boldsymbol{\epsilon}^0\|$, i.e., the likelihood is higher after every iteration, regardless of initial estimate, for the smaller CDS.*

Proof: By symmetry, it suffices to show $\|\mathbf{M}_2\| \leq \|\mathbf{M}_1\|$, i.e. $\mathbf{M}_2 \leq \mathbf{M}_1$. Since

$$\mathbf{M}_1 - \mathbf{M}_2 = \boldsymbol{\Pi}^{-1/2} \mathbf{A} (\mathbf{F}_{\mathbf{X}_2}^{-1} - \mathbf{F}_{\mathbf{X}_1}^{-1}) \mathbf{A}' \boldsymbol{\Pi}^{-1/2},$$

the result follows from $\mathbf{F}_{\mathbf{X}_2}^{-1} \geq \mathbf{F}_{\mathbf{X}_1}^{-1}$. □

Finally, define the distance from the ML estimate by:

$$\mathbf{d}_k^i = \mathbf{F}_{\mathbf{Y}}^{-1/2} (\boldsymbol{\theta}_k^i - \hat{\boldsymbol{\theta}}),$$

where $\hat{\boldsymbol{\theta}}$ denotes the ML estimate.

Theorem 5: *For $\mathbf{d}_1$ and $\mathbf{d}_2$ defined above, $\|\mathbf{d}_2^i\| \leq \|\mathbf{d}_1^i\|$ $\forall i$, i.e., the iterates corresponding to the smaller CDS are closer to the ML estimate at every iteration than those of the larger CDS.*

Proof: Since $\mathbf{d}_k^{i+1} = (\mathbf{I} - \mathbf{F}_{\mathbf{Y}}^{1/2} \mathbf{F}_{\mathbf{X}_k}^{-1} \mathbf{F}_{\mathbf{Y}}^{1/2}) \mathbf{d}_k^i$, again the result follows from $\mathbf{F}_{\mathbf{X}_2}^{-1} \geq \mathbf{F}_{\mathbf{X}_1}^{-1}$. □

We have shown that a smaller CDS not only yields faster asymptotic convergence, but also takes a larger step the first iteration, and yields iterates with higher likelihood and that are closer to the ML estimate every iteration. Theorems 3–5 do not generally follow from the asymptotic results alone; the fact that they are true here is a strong indication of a fundamental link between the size of the CDS and the convergence rate of an EM algorithm. This has significant practical implications for problems such as tomographic reconstruction where the complete data spaces are generally very large.

REFERENCES

[1] 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.

[2] M. Feder and E. Weinstein. Parameter estimation of superimposed signals using the EM algorithm. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(4):477–489, April 1988.

[3] J. A. Fessler. Hidden data spaces for maximum-likelihood PET reconstruction. In *Abstract Book of the 1992 IEEE Nuclear Science Symposium and Medical Imaging Conference*, 1992.

[4] J. A. Fessler, N. H. Clinthorne, and W. L. Rogers. On complete-data spaces for PET reconstruction algorithms, 1992. Submitted to IEEE Transactions on Nuclear Science.

[5] 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 *Conference Record of the 1991 IEEE Nuclear Science Symposium and Medical Imaging Conference*, volume 3, pages 1927–1932, 1991.

[6] A. O. Hero and J. A. Fessler. On the convergence of EM-type algorithms, 1992. In preparation.

[7] S. Kullback. *Information Theory and Statistics*. Dover, 1978.

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

[9] E. L. Lehmann. *Theory of Point Estimation*. Wiley, New York, 1983.

[10] I. Ziskind and M. Wax. Maximum likelihood localization of multiple sources by alternating projection. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(10):1553–1560, October 1988.