MESSAGE PASSING EXPECTATION-MAXIMIZATION ALGORITHMS

Joseph A. O'Sullivan

Electronic Systems and Signals Research Laboratory
Department of Electrical and Systems Engineering
Washington University, Campus Box 1127
St. Louis, MO 63130
jao@wustl.edu

ABSTRACT

Message passing algorithms have had dramatic impacts on important problems in signal processing, learning theory, communication theory, and information theory through their computational efficiency. Expectation-maximization algorithms have had dramatic impacts on problems in estimation and detection theory, but their computational efficiency often limits their applicability. Given a bipartite graphical model for the data, if a set of hidden independent random variables can be associated with the edges, then a resulting expectation-maximization algorithm is message passing on this graph. The algorithms are computationally efficient in the same sense as other message passing algorithms. One example of such algorithms is the standard expectation-maximization algorithm for emission tomography. Another example for a signal in Gaussian noise yields a statistical interpretation to efficient algorithms for sparse linear inverse problems.

1. INTRODUCTION

When attempting to solve problems in statistical signal processing, the performance for the problem must often be balanced against performance of the implementation. Indeed, one of the core challenges in statistical signal processing is the derivation of algorithms that have implementations with fixed or bounded computational and communication complexity, and have provable performance bounds for the statistical signal processing problem. In this paper, we describe a class of algorithms that meet aspects of this challenge. Whenever a problem can be put into the format described in this paper, the resulting algorithms have bounded computational and communication complexity and share the convergence and performance properties of expectation-maximization (EM) algorithms. Several example applications that illustrate the idea are shown.

Message passing algorithms originated in computer science, being used for inference on Bayesian networks. The success of message passing algorithms in iterative decoding, learning theory, communication theory, and signal processing has led to the derivation of general message passing algorithms and the reinterpretation of many existing algorithms (including the Kalman filter) as being message passing [4, 6]. The primary class of message passing algorithms that have been explored in this way can be viewed as computing the marginals of a joint distribution (see [1]). Similarly, important algorithms in image processing, including the EM algorithm for emission tomography, have locality and parallelizability properties (see below) and can be viewed as message passing algorithms. We examine a general class of problems in statistical signal processing that yield message passing EM algorithms.

Heskes, et al. [5] describe a family of approximate expectation maximization algorithms derived by using belief propagation on a graphical model to compute approximations to the expectations needed in the E-step of the EM algorithm. As they note, this is equivalent to using a Kikuchi approximation to a free energy; that is, an approximate probability distribution is used for the posterior on the hidden variables given the measured (incomplete) data and the current estimates of the parameters. Their algorithm is message passing on a graph and therefore efficient. When a model can be derived in the form that we describe here, there is no need to use their approach to approximate the E-step. When models of the form described here are not available, approximate expectation maximization may be an attractive option.

The need for efficient computations is especially important in high dimensional optimization problems such as image reconstruction. Standard tomography algorithms such as filtered backprojection have complexity proportional to the number of projections times the computations per projection; the computations per projection typically involve a Fourier transform and thus have $n \log n$ complexity, with n being the number of samples in one projection.

Iterative algorithms, including those designed to maximize likelihood, usually seek to minimize the complex-

ity per iteration. Algorithms whose complexity per iteration is proportional to the complexity of computing forward and backward projections can be viewed as computing on the graph determined by the forward projection operator. Other researchers refer to this property as a local computation property. For example, Sauer and Bouman [10] describe Gauss-Seidel iterations for transmission tomography that have this locality property. A closely related property is parallelizability: if the computations are local and a large subset of the computations can be performed simultaneously, then the algorithm is parallelizable. The Gauss-Seidel iterations of Sauer and Bouman are not parallelizable. Jeffrey Fessler [3] has described this property, and examined the parallelizability of many algorithms from the literature, including the EM algorithm.

For the problem of estimating a signal in Gaussian noise, the resulting EM algorithms (parameterized by the choice of the hidden variables) efficiently solve linear inverse problems.

In order to emphasize the common properties of the algorithms, a common example graphical model is used, namely a random sparse matrix, with entries either zero or one. This matrix is both the adjacency matrix for the bipartite graph and the matrix parameterizing the measurement system.

2. PROBLEM DEFINITION

Consider a data model of the form

$$p(\mathbf{y}|\mathbf{s}) = \prod_j p(y_j|s_k, k \in \mathcal{K}(j)), \tag{1}$$

where $\mathbf{s}$ is the parameter vector taking values in a finite dimensional space, and $\mathcal{K}(j)$ is a subset of indices such that Y_j is independent of the remaining elements of $\mathbf{s}$ given $\{s_k : k \in \mathcal{K}(j)\}$. We consider both random and nonrandom parameters. For random parameters, we assume that the S_k are independent random variables with known probability density function,

$$p(\mathbf{s}) = \prod_k p(s_k). \tag{2}$$

This data model defines a bipartite graph, with nodes corresponding to variables. One type of node corresponds to the measurements y_j, while another corresponds to the parameters s_k. There is an edge between nodes corresponding to y_j and s_k if and only if $k \in \mathcal{K}(j)$. We define the set of indices for y_j such that there is an edge between y_j and s_k for k fixed as $\mathcal{J}(k)$:

$$\mathcal{J}(k) = \{j : k \in \mathcal{K}(j)\}. \tag{3}$$

The maximum likelihood estimation problem given measurement vector $\mathbf{y}$ is to find the vector $\mathbf{s}$ that maximizes the loglikelihood function $\ln p(\mathbf{y}|\mathbf{s})$; in EM algorithm terminology this is the incomplete data loglikelihood function. The maximum a posteriori estimation problem given measurement vector $\mathbf{y}$ is to find the vector $\mathbf{s}$ that maximizes the loglikelihood function $\ln p(\mathbf{y}|\mathbf{s}) + \ln p(\mathbf{s})$. The following is our *key assumption*.

Assumption 2.1 *There exist random variables X_{jk} associated with each edge such that (for the random parameter case): (i) $\mathbf{Y}$ and $\mathbf{S}$ are independent given $\{X_{jk}\}$*

$$p(\mathbf{y}|\mathbf{x}, \mathbf{s}) = p(\mathbf{y}|\mathbf{x}); \tag{4}$$

(ii) the random variable Y_j depends only on the random variables on edges attached to y_j

$$p(\mathbf{y}|\mathbf{x}) = \prod_j p(y_j|x_{jk}, k \in \mathcal{K}(j)); \tag{5}$$

(iii) the random variables $\{X_{jk}, j \in \mathcal{J}(k)\}$ are conditionally independent given S_k, and are conditionally independent of all other $S_{k'}, k' \neq k$

$$p(\mathbf{x}|\mathbf{s}) = \prod_k \prod_{j \in \mathcal{J}(k)} p(x_{jk}|s_k). \tag{6}$$

In the EM framework, we view $\mathbf{X}$ as the complete data. The basis of the EM algorithm is the equality

$$\begin{aligned} &\ln \left[\int p(\mathbf{y}|\mathbf{x}) p(\mathbf{x}|\mathbf{s}) d\mathbf{x} \right] \\ = \; &- \min_{\Phi \in \mathcal{P}} \int \Phi(\mathbf{x}|\mathbf{y}, \mathbf{s}) \ln \left[\frac{\Phi(\mathbf{x}|\mathbf{y}, \mathbf{s})}{p(\mathbf{y}|\mathbf{x}) p(\mathbf{x}|\mathbf{s})} \right] d\mathbf{x}, \end{aligned} \tag{7}$$

where the minimization is over all conditional probability density functions $\Phi(\mathbf{x}|\mathbf{y}, \mathbf{s})$. The minimum is achieved by the probability density function

$$\Phi(\mathbf{x}|\mathbf{y}, \mathbf{s}) = \frac{p(\mathbf{y}|\mathbf{x}) p(\mathbf{x}|\mathbf{s})}{\int p(\mathbf{y}|\mathbf{x}') p(\mathbf{x}'|\mathbf{s}) d\mathbf{x}'}. \tag{8}$$

The equality (7) is a variational representation of the incomplete data loglikelihood function. Given this variational representation, the EM algorithm is equivalent to the double minimization

$$\min_{\mathbf{s}} \min_{\Phi \in \mathcal{P}} \int \Phi(\mathbf{x}|\mathbf{y}, \mathbf{s}) \ln \left[\frac{\Phi(\mathbf{x}|\mathbf{y}, \mathbf{s})}{p(\mathbf{y}|\mathbf{x}) p(\mathbf{x}|\mathbf{s})} \right] d\mathbf{x}. \tag{9}$$

Alternately minimizing over $\mathbf{s}$ and Φ leads to the EM algorithm:
(1) Set $m = 0$, select an initial guess $\mathbf{s}^{(0)}$;
(2) Minimize over Φ to obtain

$$\Phi^{(m+1)}(\mathbf{x}|\mathbf{y}, \mathbf{s}^{(m)}) = \frac{p(\mathbf{y}|\mathbf{x}) p(\mathbf{x}|\mathbf{s}^{(m)})}{\int p(\mathbf{y}|\mathbf{x}') p(\mathbf{x}'|\mathbf{s}^{(m)}) d\mathbf{x}'}; \tag{10}$$

(3) (E-step) Compute the Q-function

$$Q(\mathbf{s}|\mathbf{s}^{(m)}) = \int \Phi^{(m+1)}(\mathbf{x}|\mathbf{y},\mathbf{s}^{(m)}) \ln p(\mathbf{x}|\mathbf{s}) d\mathbf{x}; \quad (11)$$

(4) (M-step) Maximize $Q(\mathbf{s}|\mathbf{s}^{(m)})$ over $\mathbf{s}$

$$\mathbf{s}^{(m+1)} = \underset{\mathbf{s}}{\operatorname{argmax}}\, Q(\mathbf{s}|\mathbf{s}^{(m)}); \quad (12)$$

(5) Set $m = m + 1$, check for convergence, go to step 2.

Note that in the computation for $\Phi^{(m+1)}$ in step 2, $\mathbf{y}$ and $\mathbf{s}^{(m)}$ are fixed. The minimization over all possible density functions on $\mathbf{x}$ yields this result.

3. MESSAGE PASSING EM ALGORITHM

The key assumption makes this EM algorithm a message passing algorithm on the bipartite graph defined above.

Definition 3.1 *An iterative algorithm is* message passing *on a graph if computations can be associated with nodes, and communication associated with edges. That is, the computations at a given node at a given iteration use only the results of previous computations at that node and information communicated from other nodes connected to the given node by an edge.*

In practice, we assume the nodes have bounded memory and the edges have bounded communication, independent of the problem dimensions. For our bipartite graph, let g be functions associated with nodes corresponding to s_k and f be functions associated with nodes corresponding to y_j. Let ν_{jk} be the message from the node for s_k to the node for y_j and and let μ_{jk} be the message from the node for y_j to the node for s_k. Let τ_k be the value in memory at the node for s_k and let σ_j be in memory at the node for y_j. An iterative algorithm is message passing on this bipartite graph if the computations at the nodes for y_j are of the form

$$\left[\{\mu_{jk}^{(m+1)}, k \in \mathcal{K}(j)\}, \sigma_j^{(m+1)}\right] = f_j(\{\nu_{jk}^{(m)}, k \in \mathcal{K}(j)\}, \sigma_j^{(m)})$$

and the computations at the nodes for s_k are of the form

$$\left[\{\nu_{jk}^{(m+1)}, j \in \mathcal{J}(k)\}, \tau_k^{(m+1)}\right] = g_k(\{\mu_{jk}^{(m+1)}, j \in \mathcal{J}(k)\}, \tau_k^{(m)}).$$

To see that the EM algorithm here is message passing, note that

$$\ln p(\mathbf{x}|\mathbf{s}) = \sum_k \sum_{j \in \mathcal{J}(k)} \ln p(x_{jk}|s_k), \quad (13)$$

so the Q-function is

$$Q(\mathbf{s}|\mathbf{s}^{(m)}) = \sum_k \sum_{j \in \mathcal{J}(k)} E\left[\ln p(X_{jk}|s_k)|\mathbf{y},\mathbf{s}^{(m)}\right]. \quad (14)$$

The marginal probability density functions on x_{jk} given $\mathbf{y}$ and $\mathbf{s}$ are needed; denote these marginals by ϕ, and note that

$$\begin{aligned}
\phi(x_{jk}|\mathbf{y},\mathbf{s}) &= \frac{1}{Z(\mathbf{y},\mathbf{s})} \int \left[\prod_{j'} p(y_{j'}|x_{j'k'}, k' \in \mathcal{K}(j'))\right] \\
&\times \left[\prod_{j'} \prod_{k' \in \mathcal{K}(j')} p(x_{j'k'}|s_{k'})\right] \left[\prod_{k'} p(s_{k'})\right] \\
&\times \left[\prod_{j'} \prod_{k' \in \mathcal{K}(j'),(j',k') \neq (j,k)} dx_{j'k'}\right] \\
&= \frac{1}{Z(y_j, s_{k'}, k' \in \mathcal{K}(j))} \int p(y_j|x_{jk'}, k' \in \mathcal{K}(j)) \\
&\times \prod_{k' \in \mathcal{K}(j)} p(x_{jk'}|s_{k'}) \prod_{k' \in \mathcal{K}(j), k' \neq k} dx_{jk'}, \quad (15)
\end{aligned}$$

where $Z(y_j, s_{k'}, k' \in \mathcal{K}(j))$ normalizes ϕ to be a density on x_{jk}. The important thing to notice is that the computation of this density depends only on $\{s_{k'}, k' \in \mathcal{K}(j)\}$.

The messages sent from nodes corresponding to s_k can be the estimates themselves, $s_k^{(m)}$. The message sent from y_j to s_k is either the posterior density

$$\phi^{(m+1)}(x_{jk}) = \phi(x_{jk}|y_j, s_{k'}^{(m)}, k' \in \mathcal{K}(j)), \quad (16)$$

the corresponding part of the Q-function,

$$E[\ln p(X_{jk}|s_k)|y_j, s_{k'}^{(m)}, k' \in \mathcal{K}(j)], \quad (17)$$

or a sufficient statistic for this computation.

3.1. Gaussian MAP Problems

Let the random parameters to be estimated, S_k, be independent and identically distributed Gaussian random variables with zero mean and variance 1. Suppose that

$$Y_j = \sum_{k \in \mathcal{K}(j)} A_{jk} S_k + W_j, \quad (18)$$

where W_j are zero mean independent Gaussian random variables with variances σ_j^2, independent of $\mathbf{S}$. The entries A_{jk} are known. Define the random variables

$$X_{jk} = A_{jk} S_k + W_{jk}, \quad (19)$$

where W_{jk} is Gaussian, zero mean, variance σ_{jk}^2 such that $\sum_{k \in \mathcal{K}(j)} \sigma_{jk}^2 + \sigma_{j0}^2 = \sigma_j^2$, and all W_{jk} are independent of each other and of $\mathbf{S}$. Then

$$Y_j = \sum_{k \in \mathcal{K}(j)} X_{jk} + W_{j0}, \quad (20)$$

and W_{j0} is Gaussian, zero mean, variance σ_{j0}^2. It is interesting to note that the maximum a posteriori estimation problem is then equivalent to the double minimization

$$\min_{\mathbf{s}} \quad \min_{\mathbf{x}} \quad \sum_j \frac{1}{2\sigma_{j0}^2}(y_j - \sum_{k\in\mathcal{K}(j)} x_{jk})^2 \tag{21}$$

$$+ \quad \sum_j \sum_{k\in\mathcal{K}(j)} \frac{1}{2\sigma_{jk}^2}(x_{jk} - A_{jk}s_k)^2 \tag{22}$$

$$+ \quad \sum_k \frac{1}{2}s_k^2. \tag{23}$$

The EM algorithm is then immediate. Note that the messages from nodes s_k to nodes y_j, are the estimates $s_k^{(m)}$ themselves (or the estimates multiplied by A_{jk}). The messages from y_j to s_k can be just the innovations

$$y_j - \sum_{k\in\mathcal{K}(j)} A_{jk}s_k^{(m)} \tag{24}$$

(or the innovations multiplied by σ_{jk}^2/σ_j^2, that is the expected value of X_{jk} given y_j and the previous estimate for $\mathbf{S}$).

If $\Sigma = \text{diag}(\sigma_j^2)$, the MAP estimate for $\mathbf{s}$ satisfies

$$\mathbf{A}^T\Sigma^{-1}\mathbf{y} \quad = \quad \left(\mathbf{A}^T\Sigma^{-1}\mathbf{A} + \mathbf{I}\right)\mathbf{s}. \tag{25}$$

The EM algorithm is computationally efficient especially when $\mathbf{A}$ is sparse. In the high signal to noise ratio case, with diagonal Σ, the solution is approximately

$$\mathbf{s} \quad = \quad \mathbf{A}^{\#}\mathbf{y}, \tag{26}$$

where $\mathbf{A}^{\#}$ is the pseudoinverse of $\mathbf{A}$.

3.2. Emission Tomography

Suppose that Y_j is Poisson distributed with mean

$$\sum_{k\in\mathcal{K}(j)} H_{jk}s_k. \tag{27}$$

All measurements are independent of each other. In order to simplify the presentation, consider only matrices $\mathbf{H}$ that have entries that are either zero or one, with $H_{jk} = 1$ if and only if $k \in \mathcal{K}(j)$. Define the random variables X_{jk} to be independent Poisson distributed random variables with means s_k. Then

$$Y_j = \sum_{k\in\mathcal{K}(j)} X_{jk}. \tag{28}$$

The complete data loglikelihood function is a sum of terms of the form $x_{jk}\ln s_k - s_k$. The sufficient statistic that must be passed from y_j to s_k is the expected value of X_{jk} given y_j and the last estimate for $\mathbf{s}$. This expected value equals

$$x_{jk}^{(m+1)} = \frac{s_k^{(m)}}{\sum_{k'\in\mathcal{K}(j)} s_{k'}^{(m)}} y_j. \tag{29}$$

The next estimate for s_k is given by

$$s_k^{(m+1)} = \frac{1}{|\mathcal{J}(k)|} \sum_{j\in\mathcal{J}(k)} x_{jk}^{(m+1)}. \tag{30}$$

Clearly this is message passing. This property of the EM algorithm for emission tomography has previously been noted by several authors (see for example Fessler [3]).

3.3. Transmission Tomography

Suppose that Y_j is Poisson distributed with mean

$$I_0 \exp\left(\sum_{k\in\mathcal{K}(j)} A_{jk}s_k\right). \tag{31}$$

All entries of A_{jk} are positive. The EM algorithm derived by O'Sullivan and Benac [9] takes the form

$$s_k^{(m+1)} = s_k^{(m)} - \frac{1}{Z}\ln\left[\frac{\sum_{j\in\mathcal{J}(k)} A_{jk}y_j}{\sum_{j\in\mathcal{J}(k)} A_{jk}q_k^{(m)}}\right], \tag{32}$$

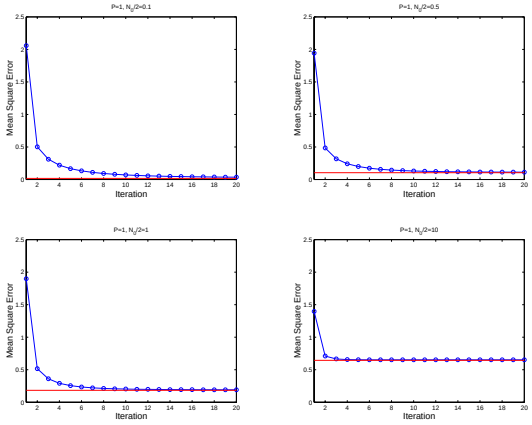
where

$$q_k^{(m)} = I_0 \exp\left(\sum_{k\in\mathcal{K}(j)} A_{jk}s_k^{(m)}\right). \tag{33}$$

Clearly this is message passing. The original derivation by Lange and Carson [7] defines hidden random variables on edges of the corresponding bipartite graph and thus is also message passing.

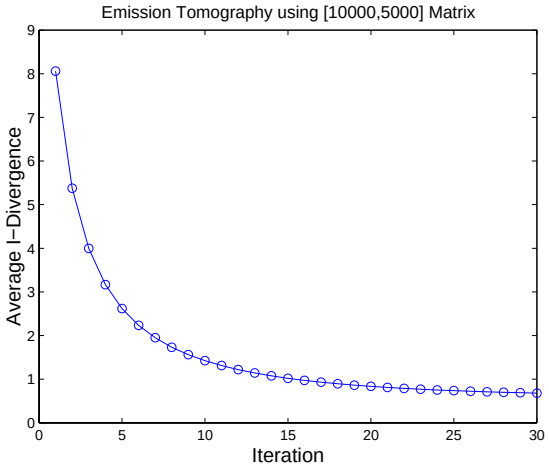
4. COMPUTATIONAL RESULTS

Each of the three algorithms described above was implemented using a common model. The matrix $\mathbf{A}$ is selected as in the emission tomography example to be $\mathbf{A} = \mathbf{H}$, where $\mathbf{H}$ is a matrix of ones and zeros. The matrix $\mathbf{H}$ is chosen using the strategy to design low density parity check codes as a 10000×5000 regular (3,6) matrix. That is, $\mathbf{H}$ has 10000 rows with 3 ones in each row and 5000 columns with 6 ones in each column. Other than those constraints, $\mathbf{H}$ is chosen at random. The results for four noise levels in the Gaussian MAP problem are shown in Figure 1 (with $N_0/2 = \sigma_i^2$, and $N_0/6 = \sigma_{jk}^2$). For the emission and transmission tomography problems, the error is measured in terms of I-divergence. The convergence rates are determined in part by the set of matrices from which $\mathbf{H}$ is drawn and in part by the specific form of the likelihood function.



Gauss-MAP example

Fig. 1. Gaussian maximum a posteriori problem estimation error per component versus iteration.



(b) EMML example

Fig. 2. Emission tomography problem estimation error per component versus iteration.

Acknowledgments

Naveen Singla performed all of the simulations reported here. This work reported was supported in part by the Office of Naval Research contract N000140310110 and National Cancer Institute of the National Institutes of Health under research grant R01CA75371 (J. F. Williamson, P. I.).

5. REFERENCES

[1] Aji and R. McEliece, "The generalized distributive law," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 325-343, March 2000.

[2] I. Csiszár and G. Tusnády, "Information Geometry and Alternating Minimization Procedures," *Statistical Decisions*, Suppl. issue no. 1, pp. 205-207, 1984.

[3] J. A. Fessler, "Statistical image reconstruction methods for transmission tomography," *Handbook of Medical Imaging, Volume 2. Medical Image Processing and Analysis* (M. Sonka and J. M. Fitzpatrick, eds.), SPIE, Bellingham, pp. 1-70, 2000.

[4] B. Frey, *Graphical models for machine learning and digital communication*, MIT Press, Cambridge, MA, 1998.

[5] T. Heskes, O. Zoeter, and W. Wiegerinck, "Approximate Expectation Maximization," in *Advances in Neural Information Processing Systems 16* (S. Thrun, L. Saul, and B. Schölkopf, eds.), MIT Press, Cambridge, MA, 2004.

[6] F. Kschischang, B. Frey, and A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 498-519, Feb. 2001.

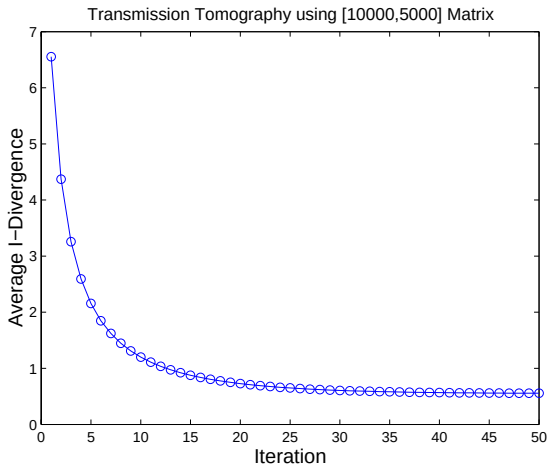
[7] K. Lange and R. Carson, "EM reconstruction algorithms for emission and transmission tomography," *J. Comp. Assisted Tomo.*, vol. 8, pp. 306–16, Apr. 1984.

[8] J. A. O'Sullivan, "Alternating minimization algorithms: from Blahut-Arimoto to expectation-maximization," in: *Codes, Curves, and Signals—Common Threads in Communications* (A. Vardy, ed.), Kluwer Academic Publishers, pp. 173–92, 1998.

[9] J. A. O'Sullivan and J. Benac, "Alternating minimization algorithms for transmission tomography," submitted to *IEEE Trans. Med. Imaging*, in revision.

[10] K. Sauer and C. Bouman, "A Local Update Strategy for Iterative Reconstruction from Projections," *IEEE Trans. on Sig. Proc.*, vol. 41, no. 2, pp. 534-548, Feb. 1993.

[11] D. L. Snyder, T. J. Schulz, and J. A. O'Sullivan, "Deblurring subject to nonnegativity constraints," *IEEE Trans. Signal Process.*, vol. 40, pp. 1143–50, May 1992.



(c) AM example

Fig. 3. Transmission tomography problem estimation error per component versus iteration.