CHOOSING PARAMETERS IN BLOCK-ITERATIVE OR ORDERED SUBSET RECONSTRUCTION

by Charles Byrne (Charles_Byrne@uml.edu),
Department of Mathematical Sciences,
University of Massachusetts Lowell, Lowell, MA 01854

Viewed abstractly, all the algorithms considered here are designed to provide a nonnegative solution x to the system of linear equations $y = Px$, where y is a vector with positive entries and P a matrix whose entries are nonnegative and with no purely zero columns.

The expectation maximization maximum likelihood (EMML) method in emission tomography [1] and the simultaneous multiplicative algebraic reconstruction technique (SMART) [2, 3, 4, 5] are slow to converge on large data sets; accelerating convergence through the use of block-iterative or ordered subset versions of these algorithms is a topic of considerable interest. These block-iterative versions involve relaxation and normalization parameters the correct selection of which may not be obvious to all users. The algorithms are not faster merely by virtue of being block-iterative; the correct choice of the parameters is crucial. Through a detailed discussion of the theoretical foundations of these methods we come to a better understanding of the precise roles these parameters play.

The notion of cross-entropy or the Kullback-Leibler distance is central to our discussion. For positive numbers a and b let

$$KL(a,b) = a\log(a/b) + b - a;$$

also let $KL(a,0) = +\infty$ and $KL(0,b) = b$. It is easily seen that $KL(a,b) > 0$ unless $a = b$. We extend this Kullback-Leibler distance component-wise to vectors x and z with nonnegative entries:

$$KL(x,z) = \sum_{j=1}^{J} KL(x_j, z_j).$$

Note that $KL(x,z)$ and $KL(z,x)$ are generally not the same. While the KL distance is not a metric in the usual sense it does have certain properties involving best approximation that are similar to those of the square of the Euclidean metric.

The methods based on cross-entropy, such as the multiplicative version of the algebraic reconstruction technique (ART), the MART [6], its simultaneous version, SMART, the expectation maximization maximum likelihood method (EMML) and all block-iterative versions of these algorithms apply to nonnegative systems that we denote by $Px = y$, where y is a vector of positive entries, P is a matrix with entries $P_{ij} \geq 0$ such that for each j the sum $s_j = \Sigma_{i=1}^{I} P_{ij}$ is positive and we seek a solution x with nonnegative entries. If no nonnegative x satisfies $y = Px$ we say the system is *inconsistent*.

Simultaneous iterative algorithms employ all of the equations at each step of the iteration; block-iterative methods do not. For the latter methods we assume that the index set $\{i = 1, ..., I\}$ is the (not necessarily disjoint) union of the N sets or *blocks* B_n, $n = 1, ..., N$. We shall require that $s_{nj} = \sum_{i \in B_n} P_{ij} > 0$ for each n and each j. Block-iterative methods like ART and MART for which each block consists of precisely one element are called *row-action* or *sequential* methods. We begin our discussion with the SMART and the EMML method.

Both the SMART and the EMML method provide a solution of $y = Px$ when such exist and (distinct) approximate solutions in the inconsistent case. Both begin with an arbitrary positive vector x^0. Having found x^k the iterative step for the SMART is

SMART:

$$x_j^{k+1} = x_j^k \exp\,(s_j^{-1} \sum_{i=1}^{I} P_{ij} \log \frac{y_i}{(Px^k)_i}) \tag{1}$$

while that for the EMML method is

EMML:

$$x_j^{k+1} = x_j^k s_j^{-1} \sum_{i=1}^{I} P_{ij} \frac{y_i}{(Px^k)_i}. \tag{2}$$

The following theorems summarize what we know of SMART and EMML.

Theorem 1 ***In the consistent case the SMART converges to the unique nonnegative solution of*** $y = Px$ ***for which the distance*** $\Sigma_{j=1}^{J} s_j KL(x_j, x_j^0)$ ***is minimized. In the inconsistent case it converges to the unique nonnegative minimizer of the distance*** $KL(Px, y)$ ***for which*** $\Sigma_{j=1}^{J} s_j KL(x_j, x_j^0)$ ***is minimized; if*** P ***and every matrix derived from*** P ***by deleting columns has full rank then there is a unique nonnegative minimizer of*** $KL(Px, y)$ ***and at most*** $I - 1$ ***of its entries are nonzero.***

Theorem 2 ***In the consistent case the EMML algorithm converges to a nonnegative solution of*** $y = Px$***. In the inconsistent case it converges to a nonnegative minimizer of the distance*** $KL(y, Px)$***; if*** P ***and every matrix derived from*** P ***by deleting columns has full rank then there is a unique nonnegative minimizer of*** $KL(y, Px)$ ***and at most*** $I - 1$ ***of its entries are nonzero.***

In the consistent case there may be multiple nonnegative solutions and the one obtained using the EMML algorithm will depend on the starting vector x^0; how it depends on x^0 is an open question. These theorems are special cases of more general results on block-iterative methods.

Those who have used the SMART or the EMML on sizable problems have certainly noticed that they are both slow to converge. An important issue, therefore, is how to accelerate convergence. One popular method is through the use of block-iterative (or ordered subset) methods. To illustrate block-iterative methods and to motivate our subsequent discussion we consider now the ordered subset EM algorithm (OSEM) [7], which is a popular technique in some areas of medical imaging, as well as an analogous version of SMART, which we shall call here the OSSMART. The OSEM algorithm is now used quite frequently in tomographic image reconstruction, where it is acknowledged to produce usable images significantly faster then EMML method.

The idea behind the OSEM (OSSMART) is simple: the iteration looks very much like the EMML (SMART), but at each step of the iteration the summations are taken only over the current block. The blocks are processed cyclically.

The OSEM iteration is the following: for $k = 0, 1, ...$ and n the index of the current block or subset, having found x^k let

OSEM:

$$x_j^{k+1} = x_j^k s_{nj}^{-1} \sum_{i \in B_n} P_{ij} \frac{y_i}{(Px^k)_i}. \tag{3}$$

The OSSMART has the following iterative step:

OSSMART:

$$x_j^{k+1} = x_j^k \exp\,(s_{nj}^{-1} \sum_{i \in B_n} P_{ij} \log \frac{y_i}{(Px^k)_i}). \tag{4}$$

In general we do not expect block-iterative algorithms to converge in the inconsistent case, but to exhibit subsequential convergence to a limit cycle. We do, however, want them to converge to a solution in the consistent case; in general, the OSEM and OSSMART do not. These two algorithms are known to converge to a solution in the consistent case when the matrix P and the set of blocks satisfy the condition known as subset balance, which means that the sums s_{nj} depend only on j and not on n. While subset balance may be approximately valid in some special cases it is overly restrictive, eliminating, for example, almost every set of blocks whose cardinalities are not all the same. When the OSEM does well in practice in medical imaging it is probably because the N is not large and only a few iterations are carried out.

The experience with the OSEM is encouraging, however, and strongly suggests that an equally fast, but mathematically rigorous, block-iterative version of EMML could be found; this is the *rescaled block-iterative* **EMML (RBI-EMML)[8]. Both RBI-EMML and an analogous corrected version of OSSMART, the RBI-SMART, provide fast convergence to a solution in the consistent case, for any choice of blocks.**

We consider now block-iterative formulations of the SMART and EMML that are general enough to include all of the variants we wish to discuss. In fact, our initial formulations will be too general and will need to be restricted in certain ways to guarantee and to accelerate convergence.

We begin with the block-iterative version of the SMART, which we shall denote BI-SMART. These methods were known prior to the discovery of RBI-EMML and played an important role in that discovery; the importance of rescaling for acceleration was apparently not appreciated, however. The SMART was discovered in 1972, independently, by Darroch and Ratcliff [2], working in statistics, and by Schmidlin [3] in medical imaging. Block-iterative versions of SMART are also treated in [2], but they also insist on subset balance; the inconsistent case was not considered. We start by considering a formulation of BI-SMART that is general enough to include all of the variants we wish to discuss. As we shall see, this formulation is too general and will need to be restricted in certain ways to obtain convergence.

Initially, we let the BI-SMART iterative step be defined as

$$x_j^{k+1} = x_j^k \exp\left(\beta_{nj} \sum_{i \in B_n} \alpha_{ni} P_{ij} \log\left(\frac{y_i}{(Px^k)_i}\right)\right), \tag{5}$$

for $j = 1, 2, ..., J$**,** $n = k(\mathrm{mod}\, N) + 1$ **and** β_{nj} **and** α_{ni} **arbitrary positive weights. Our convergence proof requires that** β_{nj} **be separable, that is,**

$$\beta_{nj} = \gamma_j \delta_n$$

for each j **and** n **so that (5) becomes**

BI-SMART:

$$x_j^{k+1} = x_j^k \exp\left(\gamma_j \delta_n \sum_{i \in B_n} \alpha_{ni} P_{ij} \log\left(\frac{y_i}{(Px^k)_i}\right)\right). \tag{6}$$

We also require

$$\gamma_j \delta_n \sigma_{nj} \leq 1, \tag{7}$$

for $\sigma_{nj} = \sum_{i \in B_n} \alpha_{ni} P_{ij}$**.**

With these conditions satisfied we have the following result.

Theorem 3 ***Let there be nonnegative solutions of*** $y = Px$***. For any positive vector*** x^0 ***and any collection of blocks*** $\{B_n, n = 1, ..., N\}$ ***the BI-SMART sequence*** $\{x^k\}$ ***given by (6) converges to the unique solution of*** $y = Px$ ***for which the weighted cross-entropy, given by*** $\sum_{j=1}^J \gamma_j^{-1} KL(x_j, x_j^0)$***, is minimized.***

We see from the theorem that how we select the γ_j is determined by how we wish to weight the terms in the sum $\Sigma_{j=1}^J \gamma_j^{-1} KL(x_j, x_j^0)$. In some cases we want to minimize the cross-entropy $KL(x, x^0)$ subject to $y = Px$; in this case we would select $\gamma_j = 1$. In other cases we may have some prior knowledge as to the relative sizes of the x_j and wish to emphasize the smaller values more; then we may choose γ_j proportional to our prior estimate of the size of x_j. Having selected the γ_j, convergence will be accelerated if we select δ_n as large as permitted by the condition $\gamma_j \delta_n \sigma_{nj} \leq 1$. This suggests that we take

$$\delta_n = 1/\max\{\sigma_{nj}\gamma_j,\, j = 1, ..., J\}. \tag{8}$$

The *rescaled* **BI-SMART (RBI-SMART) as presented in [8, 9, 10] uses this choice, but with $\alpha_{ni} = 1$ for each n and i.**

Let's look now at some of the other choices for these parameters that have been considered in the literature. First, we notice that the OSSMART does not generally satisfy the requirements, since in (4) the choices are $\alpha_{ni} = 1$ and $\beta_{nj} = s_{nj}^{-1}$; the only times this is acceptable is if the s_{nj} are separable; that is, $s_{nj} = r_j t_n$ for some r_j and t_n. This is slightly more general than the condition of subset balance and is sufficient for convergence of OSSMART, since, for $\gamma_j = \alpha_{ni} = 1$ and δ_n as in (8), the BI-SMART reduces to the OSSMART.

In [4] Censor and Segman make the choices $\beta_{nj} = 1$ and $\alpha_{ni} > 0$ such that $\sigma_{nj} \leq 1$ for all n and j. In those cases in which σ_{nj} is much less than 1 for each n and j their iterative scheme is probably excessively relaxed; it is hard to see how one might improve the rate of convergence by altering only the weights α_{ni}, however. Limiting the choice to $\gamma_j \delta_n = 1$ reduces our ability to accelerate this algorithm.

The original SMART in (1) uses $N = 1$, $\gamma_j = s_j^{-1}$ and $\alpha_{ni} = \alpha_i = 1$. Clearly (7) is satisfied; in fact it becomes an equality now. For the row-action version of SMART, the *multiplicative* ART (MART), due to Gordon, Bender and Herman [6], we take $N = I$ and $B_n = B_i = \{i\}$ for $i = 1, ..., I$.

Darroch and Ratcliff included a discussion of a block-iterative version of SMART in their 1972 paper [2]. Close inspection of their version reveals that they require that $s_{nj} = \sum_{i \in B_n} P_{ij} = 1$ for all j. Since this is unlikely to be the case initially, we might try to rescale the equations or unknowns to obtain this condition. However, unless $s_{nj} = \sum_{i \in B_n} P_{ij}$ depends only on j and not on n, which is the *subset balance* property used in [7], we cannot redefine the unknowns in a way that is independent of n.

The MART begins with a strictly positive vector x^0 and has the iterative step

MART:

$$x_j^{k+1} = x_j^k \left(\frac{y_i}{(Px^k)_i}\right)^{m_i^{-1} P_{ij}}, \tag{9}$$

for $j = 1, 2, ..., J$, $i = k(\bmod I) + 1$ and $m_i > 0$ chosen so that $m_i^{-1} P_{ij} \leq 1$ for all j. Convergence of the MART is generally faster for smaller m_i, so a good choice is $m_i = \max\{P_{ij}|, j = 1, ..., J\}$. Although this particular choice for m_i is not explicitly mentioned in the various discussions of MART, it was used in implementations of MART from the beginning.

The MART fails to converge in the inconsistent case. What is always observed, but for which no proof exists, is that, for each fixed $i = 1, 2, ..., I$, as $m \to +\infty$, the MART subsequences $\{x^{mI+i}\}$ converge to separate limit vectors, say $x^{\infty,i}$. This *limit cycle* LC $= \{x^{\infty,i}|i = 1, ..., I\}$ reduces to a single vector whenever there is a nonnegative solution of $y = Px$. The greater the minimum value of $KL(Px, y)$ the more distinct from one another the vectors of the limit cycle are. An analogous result is observed for BI-SMART.

We turn now to the block-iterative version of the EMML algorithm.

Initially, we let the iterative step of the BI-EMML be defined as

$$x_j^{k+1} = x_j^k(1 - \beta_{nj}\sigma_{nj}) + x_j^k\beta_{nj} \sum_{i \in B_n} \alpha_{ni}P_{ij}\frac{y_i}{(Px^k)_i}, \qquad (10)$$

for $j = 1, 2, ..., J$**,** $n = k(\bmod N) + 1$ **and** β_{nj} **and** α_{ni} **positive. As in the case of BI-SMART, our convergence proof requires that** β_{nj} **be separable, that is,**

$$\beta_{nj} = \gamma_j\delta_n$$

for each j **and** n **and that**

$$\gamma_j\delta_n\sigma_{nj} \leq 1,$$

for $\sigma_{nj} = \sum_{i \in B_n} \alpha_{ni}P_{ij}$**. The BI-EMML then becomes**

BI-EMML:

$$x_j^{k+1} = x_j^k(1 - \gamma_j\delta_n\sigma_{nj}) + x_j^k\gamma_j\delta_n \sum_{i \in B_n} \alpha_{ni}P_{ij}\frac{y_i}{(Px^k)_i}, \qquad (11)$$

With these conditions satisfied we have the following result.
Theorem 4 ***Let there be nonnegative solutions of*** $y = Px$***. For any positive vector*** x^0 ***and any collection of blocks*** $\{B_n, n = 1, ..., N\}$ ***the BI-EMML sequence*** $\{x^k\}$ ***given by (11) converges to a nonnegative solution of*** $y = Px$***.***

When there are multiple nonnegative solutions of $y = Px$ the solution obtained by BI-EMML will depend on the starting point x^0, but precisely how it depends on x^0 is an open question. Also, in contrast to the case of BI-SMART, the solution can depend on the particular choice of the blocks.

Having selected the γ_j, convergence will be accelerated if we select δ_n as large as permitted by the condition $\gamma_j \delta_n \sigma_{nj} \leq 1$. This suggests that once again we take δ_n as in (8). The *rescaled* BI-EMML (RBI-EMML) as presented in [8, 9, 10] uses this choice, but with $\alpha_{ni} = 1$ for each n and i.

Let's look now at some of the other choices for these parameters that have been considered in the literature. First, we notice that the OSEM does not generally satisfy the requirements, since in (3) the choices are $\alpha_{ni} = 1$ and $\beta_{nj} = s_{nj}^{-1}$; the only times this is acceptable is if the s_{nj} are separable; that is, $s_{nj} = r_j t_n$ for some r_j and t_n. This is slightly more general than the condition of subset balance and is sufficient for convergence of OSEM, since, for $\gamma_j = \alpha_{ni} = 1$ and δ_n as in (8), the BI-EMML reduces to the OSEM .

The original EMML in (2) uses $N = 1$, $\gamma_j = s_j^{-1}$ and $\alpha_{ni} = \alpha_i = 1$. Clearly (7) is satisfied; in fact it becomes an equality now. Notice that the calculations required to perform the BI-SMART are somewhat more complicated than those needed in BI-EMML. Because the MART converges rapidly in most cases there is considerable interest in the row-action version of EMML. It was clear from the outset that using the OSEM in a row-action mode does not work. We see from the formula for BI-EMML that the proper row-action version of EMML, which we call the EM-MART, has the iterative step

EM-MART:

$$x_j^{k+1} = (1 - \delta_i \gamma_j \alpha_{ii} P_{ij}) x_j^k + \delta_i \gamma_j \alpha_{ii} P_{ij} \frac{y_i}{(Px^k)_i}, \tag{12}$$

with

$$\gamma_j \delta_i \alpha_{ii} P_{ij} \leq 1$$

for all i and j.

The optimal choice would seem to be to take $\delta_i\alpha_{ii}$ as large as possible; that is, to select

$$\delta_i\alpha_{ii} = 1/\max\{\gamma_j P_{ij}, j = 1, ..., J\}.$$

With this choice the EM-MART is called the *rescaled* **EM-MART (REM-MART). The EM-MART fails to converge in the inconsistent case. What is always observed, but for which no proof exists, is that, for each fixed $i = 1, 2, ..., I$, as $m \to +\infty$, the EM-MART subsequences $\{x^{mI+i}\}$ converge to separate limit vectors, say $x^{\infty,i}$. This** *limit cycle* **LC** $= \{x^{\infty,i} | i = 1, ..., I\}$ **reduces to a single vector whenever there is a nonnegative solution of $y = Px$. The greater the minimum value of $KL(y, Px)$ the more distinct from one another the vectors of the limit cycle are. An analogous result is observed for BI-EMML.**

We must mention a method that closely resembles the REM-MART, the *row-action maximum likelihood algorithm* **(RAMLA), which was discovered independently by Browne and De Pierro [11]. The RAMLA avoids the limit cycle in the inconsistent case by using strong underrelaxation involving a decreasing sequence of relaxation parameters λ_k. The RAMLA has the following iterative step:**

RAMLA:

$$x_j^{k+1} = (1 - \lambda_k \sum_{i \in B_n} P_{ij})x_j^k + \lambda_k x_j^k \sum_{i \in B_n} P_{ij}(\frac{y_i}{(Px^k)_i}), \qquad (13)$$

where the positive relaxation parameters λ_k are chosen to converge to zero and $\sum_{k=0}^{+\infty} \lambda_k = +\infty$.

Acknowledgments

This work was supported in part by grants R01EB001457 and R01EB002798 from the National Institute of Biomedical Imaging and Bioengineering. The views expressed here are those of the author and do not necessarily reflect those of the funder.

The results discussed in this poster are taken from a paper to be published in the IEEE Transactions on Image Processing. The full paper is available on my website (http://faculty.uml.edu/cbyrne/cbyrne.html).

References

[1] Y. Vardi, L.A. Shepp and L. Kaufman, A statistical model for positron emission tomography, *Journal of the American Statistical Association*, **80**, pp. 8–20, 1985.

[2] J. Darroch and D. Ratcliff, Generalized iterative scaling for log-linear models, *The Annals of Mathematical Statistics*, **43 (5)**, pp. 1470–1480, 1972.

[3] P. Schmidlin, Iterative separation of sections in tomographic scintigrams, *Nuclear Medicine*, **15 (1)**, Schatten Verlag, Stuttgart, 1972.

[4] Y. Censor and J. Segman, On block-iterative maximization, *Journal of Information and Optimization Sciences*, **8**, pp. 275-291, 1987.

[5] C. Byrne, Iterative image reconstruction algorithms based on cross-entropy minimization, *IEEE Transactions on Image Processing*, **IP-2**, pp. 96–103, 1993.

[6] R. Gordon, R. Bender and G.T. Herman, Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and x-ray photography, *Journal of Theoretical Biology*, **29**, pp. 471–481, 1970.

[7] H.M. Hudson and R.S. Larkin, Accelerated image reconstruction using ordered subsets of projection data, *IEEE Transactions on Medical Imaging*, **13**, pp. 601-609, 1994.

[8] C. Byrne, Block-iterative methods for image reconstruction from projections, *IEEE Transactions on Image Processing*, **IP-5**, pp. 792-794, 1996.

[9] C. Byrne, Convergent block-iterative algorithms for image reconstruction from inconsistent data, *IEEE Transactions on Image Processing*, **IP-6**, pp. 1296–1304, 1997.

[10] C. Byrne, Accelerating the EMML algorithm and related iterative algorithms by rescaled block-iterative methods, *IEEE Transactions on Image Processing*, **IP-7**, pp. 100-109, 1998.

[11] J. Browne and A. De Pierro, A row-action alternative to the EM algorithm for maximizing likelihoods in emission tomography, *IEEE Transactions on Medical Imaging*, **15**, pp. 687–699, 1996.