On Convergence of the EM-ML Algorithm for PET Reconstruction

Jens Gregor, Søren P. Olesen, and Michael G. Thomason
Department of Electrical Engineering & Computer Science
University of Tennessee
1122 Volunteer Blvd., Suite 203 Knoxville, TN 37996–3450

October 31, 2007

Tech. report UT–CS–07–605

Abstract

The EM-ML (expectation-maximization, maximum-likelihood) algorithm for PET reconstruction is an iterative method. Sequence convergence to a fixed point that satisfies the Karush-Kuhn-Tucker conditions for optimality has previously been established [1, 2, 3]. This correspondence first gives an alternative proof of sequence convergence and optimality based on direct expansion of certain Kullback discrimination functions and a standard result in optimization theory. Using results in series convergence, we then show that several sequences converge to 0 faster than $k \to \infty$, i.e., the sequences are $o(k^{-1})$.

Keywords:	positron emission tomography	expectation-maximization
	iterative image reconstruction	computed imaging

I. EM-ML Iteration Scheme

The EM-ML algorithm for PET reconstruction [1, 2, 3] maximizes the Poisson likelihood $P(\boldsymbol{n}^*|\boldsymbol{\lambda})$ where vectors $\boldsymbol{n^*} = (n_1, \ldots, n_D)$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_B)$, respectively, denote the externally recorded emission counts and the internal emission rates that are to be estimated. The log-likelihood can be expressed as

$$\begin{aligned} \ell(\boldsymbol{\lambda}) &= \log P(\boldsymbol{n}^*|\boldsymbol{\lambda}) && (1) \\ &= \sum_d n_d^* \log \lambda_d^* - \sum_d \log n_d^*! - \sum_d \lambda_d^* && (2) \end{aligned}$$

where $\lambda_d^* = \sum_b \lambda_b p_{bd}$ is the expected number of recorded emission counts, collectively referred to by vector $\boldsymbol{\lambda}^* = (\lambda_1^*, \ldots, \lambda_D^*)$, and p_{bd} is the conditional probability that emission activity at pixel b is detected by tube d. Without loss of generality, normalizations $\sum_d p_{bd} = 1$ for all b and $\sum_d n_d^* = 1$ are assumed [2]. $\ell(\boldsymbol{\lambda})$, and thus $P(\boldsymbol{n}^*|\boldsymbol{\lambda})$, is maximized iff $\boldsymbol{\lambda}^* = \boldsymbol{n^*}$.

The EM-ML algorithm re-estimates the emission rate at pixel b using the multiplicative update scheme

$$\begin{aligned} \lambda_b^{k+1} &= \lambda_b^k (1 + \nabla_b \ell(\boldsymbol{\lambda}^k)) && (3) \\ &= \lambda_b^k \sum_d \frac{n_d^*}{\lambda_d^{*k}} p_{bd} && (4) \end{aligned}$$

where $\lambda_b^0 = 1$ for each b is a viable choice for the initial estimate (because uniform values in $\boldsymbol{\lambda}^0$ cancel out when computing $\boldsymbol{\lambda}^1$). Important features of the algorithm are that $\boldsymbol{\lambda}^k$ is always nonnegative and that the normalizations $\sum_b \lambda_b^k = \sum_d \lambda_d^{*k} = \sum_d n_d^*$ are maintained for $k = 1, 2, \ldots$

Now rewrite the log-likelihood as

$$\ell(\boldsymbol{\lambda}) = f(\boldsymbol{n}^*, \boldsymbol{\lambda}^*) - \mathcal{D}(\boldsymbol{n}^*, \boldsymbol{\lambda}^*) \qquad (5)$$

where

$$\begin{aligned} f(\boldsymbol{n}^*, \boldsymbol{\lambda}^*) &= \sum_d n_d^* \log n_d^* - \sum_d \log n_d^*! - \sum_d \lambda_d^* && (6) \\ &= \sum_d n_d^* \log n_d^* - \sum_d \log n_d^*! - 1 && (7) \end{aligned}$$

and

$$\mathcal{D}(\boldsymbol{n}^*, \boldsymbol{\lambda}^*) = \sum_d n_d^* \log \frac{n_d^*}{\lambda_d^*}. \qquad (8)$$

Note that $f(\boldsymbol{n}^*, \boldsymbol{\lambda}^*)$ is constant given observed data $\boldsymbol{n}^*$ and that $\mathcal{D}(\boldsymbol{n}^*, \boldsymbol{\lambda}^*)$ is a nonnegative Kullback discrimination which equals 0 iff its two arguments are identical [4]. Therefore, $P(\boldsymbol{n}^*|\boldsymbol{\lambda})$ is maximized iff $\mathcal{D}(\boldsymbol{n}^*, \boldsymbol{\lambda}^*)$ is minimized, which in turn is true iff the sequence $\{\boldsymbol{\lambda}^k\}$ converges to a fixed point $\widehat{\boldsymbol{\lambda}}$ for which $\mathcal{D}(\boldsymbol{n}^*, \widehat{\boldsymbol{\lambda}}^*) = 0$. This convergence has previously been established by showing $\boldsymbol{\lambda}^k \to \widehat{\boldsymbol{\lambda}}$ where $\widehat{\boldsymbol{\lambda}}$ satisfies the Karush-Kuhn-Tucker conditions for optimality [1, 2, 3].

It can also be convenient and informative to establish that $\mathcal{D}(\boldsymbol{n}^*, \boldsymbol{\lambda}^{*k}) \to 0$ by direct expansion of certain Kullback discrimination functions together with a result in optimization theory for concave functions and convex sets. After giving this alternative proof of sequence convergence and optimality, we use results in series convergence to show that $k\mathcal{D}(\boldsymbol{n}^*, \boldsymbol{\lambda}^{*k})$ also converges to 0, i.e., $\mathcal{D}(\boldsymbol{n}^*, \boldsymbol{\lambda}^{*k}) \to 0$ faster than $k \to \infty$.

II. Sequence Convergence

Proposition 1 $\mathcal{D}(\boldsymbol{n}^*, \boldsymbol{\lambda}^{*k}) - \mathcal{D}(\boldsymbol{n}^*, \boldsymbol{\lambda}^{*k+1}) \geq \mathcal{D}(\boldsymbol{\lambda}^{k+1}, \boldsymbol{\lambda}^k)$ *for* $k = 0, 1, 2, \ldots$

Proof: Expanding the lefthand side of the inequality we get

$$\begin{aligned}
\mathcal{D}(\boldsymbol{n}^*, \boldsymbol{\lambda}^{*k}) - \mathcal{D}(\boldsymbol{n}^*, \boldsymbol{\lambda}^{*k+1}) &= \sum_d n_d^* \log \frac{n_d^*}{\lambda_d^{*k}} - \sum_d n_d^* \log \frac{n_d^*}{\lambda_d^{*k+1}} \\
&= \sum_d n_d^* \frac{\lambda_d^{*k}}{\lambda_d^{*k}} \log \frac{n_d^*}{\lambda_d^{*k}} \frac{p_{bd}}{p_{bd}} \frac{\lambda_d^{*k+1}}{n_d^*} \\
&= \sum_{b,d} \lambda_b^k \frac{n_d^*}{\lambda_d^{*k}} p_{bd} \log \frac{n_d^* p_{bd}}{\lambda_d^{*k}} \frac{\lambda_d^{*k+1}}{n_d^* p_{bd}} \\
&= \sum_{b,d} \pi_{bd}^{k+1} \log \frac{\pi_{bd}^{k+1}}{\lambda_b^k} \frac{\lambda_b^{k+1}}{\pi_{bd}^{k+2}}
\end{aligned}$$

where $\pi_{bd}^{k+1} = \lambda_b^k (n_d^*/\lambda_d^{*k}) p_{bd}$; note that $\pi_{bd}^{k+1} \geq 0$ and $\sum_{b,d} \pi_{bd}^{k+1} = 1$. The righthand side of the inequality is

$$\begin{aligned}
\mathcal{D}(\boldsymbol{\lambda}^{k+1}, \boldsymbol{\lambda}^k) &= \sum_b \lambda_b^{k+1} \log \frac{\lambda_b^{k+1}}{\lambda_b^k} \\
&= \sum_{b,d} \lambda_b^k \frac{n_d^*}{\lambda_d^{*k}} p_{bd} \log \frac{\lambda_b^{k+1}}{\lambda_b^k} \\
&= \sum_{b,d} \pi_{bd}^{k+1} \log \frac{\lambda_b^{k+1}}{\lambda_b^k}.
\end{aligned}$$

The proposition follows by subtracting this result from the one above since

$$\begin{aligned}
\sum_{b,d} \pi_{bd}^{k+1} \log \frac{\pi_{bd}^{k+1}}{\lambda_b^k} \frac{\lambda_b^{k+1}}{\pi_{bd}^{k+2}} - \sum_{b,d} \pi_{bd}^{k+1} \log \frac{\lambda_b^{k+1}}{\lambda_b^k} &= \sum_{b,d} \pi_{bd}^{k+1} \log \frac{\pi_{bd}^{k+1}}{\pi_{bd}^{k+2}} \\
&= \mathcal{D}(\boldsymbol{\pi}^{k+1}, \boldsymbol{\pi}^{k+2}) \\
&\geq 0
\end{aligned}$$

where we used the fact that $\boldsymbol{\pi}^{k+1}$ and $\boldsymbol{\pi}^{k+2}$ are probability distributions. $\square$

Proposition 2 *The sequence* $\{\boldsymbol{\lambda}^{\boldsymbol{k}}\}$ *converges monotonically to* $\widehat{\boldsymbol{\lambda}}$ *which is a fixed point that maximizes the log-likelihood function* ℓ.

Proof: Since $\mathcal{D}(\boldsymbol{\lambda}^{k+1}, \boldsymbol{\lambda}^k) \geq 0$ with equality iff $\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k$, it follows from Proposition 1 that $\{\mathcal{D}(\boldsymbol{n}^*, \boldsymbol{\lambda}^{*k})\}$ is a monotonically decreasing sequence of real numbers bounded below by 0; hence, it converges [5]. Consequently,

$$\begin{aligned}
\mathcal{D}(\boldsymbol{n}^*, \boldsymbol{\lambda}^{*k}) - \mathcal{D}(\boldsymbol{n}^*, \boldsymbol{\lambda}^{*k+1}) \to 0 \quad &\Rightarrow \quad \mathcal{D}(\boldsymbol{\lambda}^{k+1}, \boldsymbol{\lambda}^k) \to 0 \\
&\Rightarrow \quad \boldsymbol{\lambda}^k \to \widehat{\boldsymbol{\lambda}}.
\end{aligned}$$

The log-likelihood ℓ is a concave function [1, 3]. The set of candidate solutions

$$\mathcal{P}_B = \{(\lambda_1, \ldots, \lambda_B) \mid \lambda_b \geq 0, \ \sum_b \lambda_b = 1\}$$

is the convex hull of the extreme points $\{\lambda_b = 1 \mid 1 \leq b \leq B\}$ and thus a convex set [6]. A standard result in optimization of concave functions on convex sets is that a necessary and sufficient condition for fixed point $\widehat{\boldsymbol{\lambda}}$ to be a maximizer of ℓ is

$$(\boldsymbol{\lambda} - \widehat{\boldsymbol{\lambda}})' \nabla \ell(\widehat{\boldsymbol{\lambda}}) \leq 0$$

for every $\boldsymbol{\lambda} \in \mathcal{P}_B$ [7]. But since $\boldsymbol{\lambda} \geq 0$ and

$$\begin{aligned}
\widehat{\boldsymbol{\lambda}}' \nabla \ell(\widehat{\boldsymbol{\lambda}}) &= \sum_b \widehat{\lambda}_b \nabla_b \ell(\widehat{\boldsymbol{\lambda}}) \\
&= \sum_b \widehat{\lambda}_b \sum_d \frac{n_d^*}{\widehat{\lambda}_d^*} p_{bd} - \sum_b \widehat{\lambda}_b \\
&= \sum_d n_d^* - \sum_b \widehat{\lambda}_b \\
&= 0,
\end{aligned}$$

it suffices to show that $\nabla_b \ell(\widehat{\boldsymbol{\lambda}}) \leq 0$ for every b. By contradiction, suppose that there exists a pixel b for which $\nabla_b \ell(\widehat{\boldsymbol{\lambda}}) > 0$. Then $1 + \nabla_b \ell(\boldsymbol{\lambda}^k)$ converges to a value greater than 1 which in turn implies that $\lambda_b^k \to \infty$ (cf. Equation 3), but this is a contradiction of the convergence $\lambda_b^k \to \widehat{\lambda}_b$ to a finite $\widehat{\lambda}_b$. □

Corollary 1 $\mathcal{D}(\boldsymbol{n}^*, \boldsymbol{\lambda}^{*k}) \to 0.$

Proof: From Proposition 2, $\boldsymbol{\lambda}^k \to \widehat{\boldsymbol{\lambda}}$ which means that $\boldsymbol{\lambda}^{*k} \to \widehat{\boldsymbol{\lambda}}^*$. The corollary then follows since $\widehat{\boldsymbol{\lambda}}$ is a maximizer of $P(\boldsymbol{n}^* | \boldsymbol{\lambda})$ which implies that $\widehat{\boldsymbol{\lambda}}^* = \boldsymbol{n}^*$. □

III. $o(k^{-1})$ Sequences

Proposition 3 $\mathcal{D}(\widehat{\boldsymbol{\lambda}}, \boldsymbol{\lambda}^k) - \mathcal{D}(\widehat{\boldsymbol{\lambda}}, \boldsymbol{\lambda}^{k+1}) \geq \mathcal{D}(\boldsymbol{n}^*, \boldsymbol{\lambda}^{*k})$ *for* $k = 0, 1, 2, \ldots$

Proof: Expanding the lefthand side of the inequality we get

$$\begin{aligned}
\mathcal{D}(\widehat{\boldsymbol{\lambda}}, \boldsymbol{\lambda}^k) - \mathcal{D}(\widehat{\boldsymbol{\lambda}}, \boldsymbol{\lambda}^{k+1}) &= \sum_b \widehat{\lambda}_b \log \frac{\lambda_b^{k+1}}{\lambda_b^k} \\
&= \sum_b \widehat{\lambda}_b \log \sum_d \frac{n_d^*}{\lambda_d^{*k}} p_{bd} \\
&= E_{\widehat{\lambda}_\bullet} \left[\log E_{p_{b\bullet}} \left[\frac{n^*}{\lambda^{*k}} \right] \right]
\end{aligned}$$

where $E_{\widehat{\lambda}_\bullet}$ denotes expectation with respect to fixed point $\widehat{\lambda}_b$ and $E_{p_{b\bullet}}$ denotes conditional expectation with respect to p_{bd}. The proposition follows by applying Jensen's inequality (cf. [8]) to this result since

$$\begin{aligned} E_{\widehat{\lambda}_\bullet}\left[\,\log E_{p_{b\bullet}}\left[\,\frac{n^*}{\lambda^{*k}}\,\right]\,\right] &\geq E_{\widehat{\lambda}_\bullet}\left[\,E_{p_{b\bullet}}\left[\,\log\frac{n^*}{\lambda^{*k}}\,\right]\,\right] \\ &= \sum_b \widehat{\lambda}_b \sum_d p_{bd}\log\frac{n^*_d}{\lambda^{*k}_d} \\ &= \sum_d \sum_b \widehat{\lambda}_b p_{bd}\log\frac{n^*_d}{\lambda^{*k}_d} \\ &= \sum_d n^*_d \log\frac{n^*_d}{\lambda^{*k}_d} \\ &= \mathcal{D}(\boldsymbol{n}^*, \boldsymbol{\lambda}^{*k}) \end{aligned}$$

where we used the fact that $n^*_d = \widehat{\lambda}^*_d = \sum_b \widehat{\lambda}_b p_{bd}$. □

Corollary 2 $\sum_{k=0}^{\infty} \mathcal{D}(\boldsymbol{n}^*, \boldsymbol{\lambda}^{*k}) \leq \mathcal{D}(\widehat{\boldsymbol{\lambda}}, \boldsymbol{\lambda}^0)$.

Proof: Since Kullback discrimination is non-negative, we can take the partial sum on both sides of the inequality of Proposition 3 to obtain

$$\begin{aligned} \sum_{k=0}^{m} \mathcal{D}(\boldsymbol{n}^*, \boldsymbol{\lambda}^{*k}) &\leq \sum_{k=0}^{m} (\mathcal{D}(\widehat{\boldsymbol{\lambda}}, \boldsymbol{\lambda}^k) - \mathcal{D}(\widehat{\boldsymbol{\lambda}}, \boldsymbol{\lambda}^{k+1})) \\ &= \mathcal{D}(\widehat{\boldsymbol{\lambda}}, \boldsymbol{\lambda}^0) - \mathcal{D}(\widehat{\boldsymbol{\lambda}}, \boldsymbol{\lambda}^{m+1}). \end{aligned}$$

The corollary follows because $\boldsymbol{\lambda}^{m+1} \to \widehat{\boldsymbol{\lambda}}$ such that $\mathcal{D}(\widehat{\boldsymbol{\lambda}}, \boldsymbol{\lambda}^{m+1}) \to 0$. □

Corollary 3 *The sequence* $\{\mathcal{D}(\boldsymbol{n}^*, \boldsymbol{\lambda}^{*k})\}$ *is* $o(k^{-1})$.

Proof: From Proposition 2 and corollary 2 we have that $\sum_{k=0}^{\infty} \mathcal{D}(\boldsymbol{n}^*, \boldsymbol{\lambda}^{*k})$ is a convergent series of positive monotonically decreasing terms. Hence, $k\mathcal{D}(\boldsymbol{n}^*, \boldsymbol{\lambda}^{*k}) \to 0$ [5] and the corollary follows. □

It is also true that both $\mathcal{D}(\boldsymbol{\lambda}^{k+1}, \boldsymbol{\lambda}^k)$ and $||\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k||_1^2$ are $o(k^{-1})$ because the inequality $\mathcal{D}(\boldsymbol{\lambda}^{k+1}, \boldsymbol{\lambda}^k) \leq \mathcal{D}(\boldsymbol{n}^*, \boldsymbol{\lambda}^{*k})$ holds by Proposition 1 and the inequality $||\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k||_1^2 \leq C\mathcal{D}(\boldsymbol{\lambda}^{k+1}, \boldsymbol{\lambda}^k)$ for constant $C > 0$ holds between the squared-norm and the Kullback function (cf.[9]).

References

[1] L. A. Shepp and Y. Vardi, "Maximum likelihood reconstruction for emission tomography," *IEEE Trans. Med. Imaging*, vol. 1, pp. 113–122, 1982.

[2] Y. Vardi, L. A. Shepp, and L. Kaufman, "A statistical model of positron emission tomography," *J. Am. Stat. Assoc.*, vol. 80, pp. 8–20, 1985.

[3] K. Lange and R. Carson, "EM reconstruction algorithms for emission and transmission tomography," *J. Comp. Asst. Tomog.*, vol. 8, pp. 306–316, 1984.

[4] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.

[5] K. Knopp, *Theory and Application of Infinite Series*. London and Glasgow: Blackie and Son Limited, 1949.

[6] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming Theory and Algorithms*. New York: Wiley, 1993.

[7] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[8] T. M. Cover, “An algorithm for maximizing expected log investment return,” *IEEE Trans. Info. Theory*, vol. IT-30, pp. 369–373, 1984.

[9] S. Kullback, “A lower bound for discrimination information in terms of variation,” *IEEE Trans. Info. Theory*, vol. IT-13, pp. 126–127, 1967.