# Related Algorithms and An Open Question — Discussion of Lange, Hunter, and Yang Paper

Ying Nian Wu
Dept. of Statistics, UCLA

I have learned a great deal from reading this article, and I believe that many people whose research involves scientific computing will benefit from it. In the following, I will discuss several related algorithms and an open question.

1. *A similar algorithm for global optimization.* Gradual Non-Convexity (GNC) algorithm is an algorithm that shares similar spirit to optimization transfer, but with a rather different goal. It was designed for reconstructing piecewise continuous images from noisy observations (Blake and Zisserman, 1987). To make things simple, let's consider the one-dimensional case. Let $Y = (Y_1, ..., Y_n)$ be a signal observed on a one-dimensional grid $(1, ..., n)$. We want to recover the "true" signal $\theta = (\theta_1, ..., \theta_n)$ by minimizing

$$l(\theta) = \sum_{i=1}^{n} (Y_i - \theta_i)^2 + \lambda \sum_{i=1}^{n-1} g(\theta_{i+1} - \theta_i),$$

where $\lambda$ is a fixed constant, and $g(x)$ is a continuous function that penalizes large $|x|$ when $|x|$ is within a threshold, but becomes constant when $|x|$ is beyond that threshold. This weak continuity constraint is used to accommodate edges. The goal is to find the *global* minimum of $l(\theta)$. This can be difficult because $l(\theta)$ is not convex and can have many local minima, so a gradient descent algorithm can be easily trapped in a local minimum. In the GNC algorithm, a class of functions $g_p(x)$, indexed by $p \in [0, 1]$, is designed to approximate $g(x)$, so that a class of surrogate objective functions

$$l_p(\theta) = \sum_{i=1}^{n} (Y_i - \theta_i)^2 + \lambda \sum_{i=1}^{n-1} g_p(\theta_{i+1} - \theta_i)$$

can be obtained. The $g_p(x)$ are chosen in such a way that when $p = 1$, $l_p(\theta)$ is a convex function, and as $p$ decreases to 0, $g_p(x)$ converges to $g(x)$ and $l_p(\theta)$ gradually becomes non-convex. The algorithm runs as follows. First, use gradient descent to find the minimum, say, $\theta^1$, of the surrogate function $l_1(\theta)$. $\theta^1$ is the global minimum of $l_1(\theta)$ because the latter is convex. After that, lower $p$ a little bit, and run gradient descent from the current value of $\theta$, i.e., $\theta^1$, to find a local minimum of the new surrogate function $l_p(\theta)$. Repeat this process until

$p = 0$. This algorithm proves to be quite effective for finding the near-optimal reconstruction. The reason is that $l_p(\theta)$ for $p = 1$ captures the rough shape of $l(\theta)$, and as $p$ decreases, $l_p(\theta)$ resembles $l(\theta)$ with increasingly higher resolution.

2. *Transformation schemes for accelerating EM.* There are several recent advances on accelerating the EM algorithm within the EM or optimization transfer philosophy. To fix notation, let $Y_{\rm obs}$ be the observed data, $Y_{\rm mis}$ be the missing data, and $p(y_{\rm obs}, y_{\rm mis} \mid \theta)$ be the complete-data model. The goal is to maximize the log-likelihood of the observed data model, i.e., $\log p(Y_{\rm obs} \mid \theta)$. Each iteration of EM maximizes the surrogate function

$$Q(\theta \mid \theta^n) = \mathrm{E}[\log p(Y_{\rm obs}, Y_{\rm mis} \mid \theta) \mid Y_{\rm obs}, \theta^n],$$

where the expectation is with respect to the predictive distribution of the missing data $[Y_{\rm mis} \mid Y_{\rm obs}, \theta^n]$.

Meng and van Dyk (1997) proposed a clever "efficient augmentation" scheme to accelerate the above EM. The basic idea is to introduce a *transformation* $Y_{\rm mis} = t(Z_{\rm mis}, r_a(\theta))$, where for fixed $a$, $r_a$ is a one-to-one mapping in the parameter space, and for a fixed value of $r_a(\theta)$, $t$ is a one-to-one mapping in the space of missing data. Then, under the original model $p(y_{\rm obs}, y_{\rm mis} \mid \theta)$, the above transformation induces a class of "twisted models" $p_a(y_{\rm obs}, z_{\rm mis} \mid \theta)$ indexed by the working parameter "$a$". Therefore, a fixed "$a$" leads to an EM implementation, each iteration of which maximizes the surrogate function

$$Q_a(\theta \mid \theta^n) = \mathrm{E}[\log p_a(Y_{\rm obs}, Z_{\rm mis} \mid \theta) \mid Y_{\rm obs}, \theta^n].$$

An optimal "$a$" can be selected to achieve the best rate of convergence by minimizing the "fraction of missing information". Of course, "efficient augmentation" has a much broader statistical meaning than described above.

Liu, Rubin, and Wu (1998) proposed to replace $r_a(\theta)$ in the above transformation by an independent new parameter $\alpha$, i.e., $Y_{\rm mis} = t(Z_{\rm mis}, \alpha)$. Then, under the original model $p(y_{\rm obs}, y_{\rm mis} \mid \theta)$, this transformation induces an "expanded model" $p(y_{\rm obs}, z_{\rm mis} \mid \theta, \alpha)$. This leads to a "parameter expanded EM (PX-EM) algorithm", each iteration of which maximizes the surrogate function

$$Q(\theta, \alpha \mid \theta^n) = \mathrm{E}[\log p(Y_{\rm obs}, Z_{\rm mis} \mid \theta, \alpha) \mid Y_{\rm obs}, \theta^n, \alpha_0],$$

over both $\theta$ and $\alpha$ to obtain $\theta^{n+1}$. $\alpha_0$ is an arbitrary constant (usually chosen such that $t(\cdot, \alpha_0)$ is an identity transformation). Originally, PX-EM of Liu, Rubin, and Wu (1998) was

motivated by the observation that parameters in a model can become non-identifiable if the data are not fully observed. Statistically, the algorithm can be considered as performing a more "efficient analysis".

From the optimization transfer perspective, "efficient augmentation" designs a class of surrogate functions and choose among them one that best approximates the objective function, while PX-EM designs a surrogate function with auxiliary dimensions to increase the freedom of search.

3. *EM with multiple sources of missing information*? Although optimization techniques used in attacking complex scientific problems rely heavily on Monte Carlo simulation, it is always desirable to avoid simulation for the sake of efficiency (GNC algorithm is one example). The following is a maximum likelihood problem for which I am unable to find an EM or optimization transfer algorithm. The problem is still formulated in the incomplete data context, except that we have two sources of missing information, i.e., $Y_{\rm mis} = (Y_{\rm mis,1}, Y_{\rm mis,2})$. The computational constraint is that expectations with respect to the joint predictive distribution $[Y_{\rm mis} \mid Y_{\rm obs}, \theta]$ cannot be computed in closed form, but expectations with respect to conditional predictive distributions $[Y_{\rm mis,1} \mid Y_{\rm mis,2}, Y_{\rm obs}, \theta]$ and $[Y_{\rm mis,2} \mid Y_{\rm mis,1}, Y_{\rm obs}, \theta]$ are readily available. With a prior on $\theta$, a Gibbs sampler can be easily designed to simulate the posterior $p(\theta \mid Y_{\rm obs})$. However, it is unclear whether there exists a parallel closed-form EM or optimization transfer algorithm for maximizing $\log p(Y_{\rm obs} \mid \theta)$.

In conclusion, I thank the authors for this stimulating paper. I also thank the editor for inviting me to contribute the discussion.

**Additional references**


Black, A. and Zisserman A. (1987) *Visual reconstruction*, MIT press.

Liu, C., Rubin, D. B., and Wu, Y. N. (1998) Parameter expansion to accelerate EM — the PX-EM algorithm, *Biometrika*, **85**, 755-770.

Meng, X.-L. and van Dyk, D. (1997) The EM algorithm - an old folk song sung to a fast tune (with discussion), *J. R. Statist. Soc. B*, **59**, 511-67.