NOISE BENEFITS

IN

EXPECTATION-MAXIMIZATION

ALGORITHMS

by

Osonde Adekorede Osoba

A Dissertation Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(ELECTRICAL ENGINEERING)

August 2013

Chapter 1

Preview of Dissertation Results

The main aim of this dissertation is to demonstrate that noise injection can improve the average speed of Expectation-Maximization (EM) algorithms. The EM discussion in Chapter 2 gives an idea of the power and generality of the EM algorithm schema. But EM algorithms have a key weakness: they converge slowly especially on high-dimensional incomplete data. Noise injection can address this problem. The Noisy Expectation Maximization (NEM) theorem (Theorem 3.1) in Chapter 3 describes a condition under which injected noise causes faster EM convergence on average. This general condition reduces to a simpler condition (Corollary 3.2) for Gaussian mixture models (GMMs). The GMM noise benefit leads to EM speed-ups in clustering algorithms and in the training of hidden Markov models. The general NEM noise benefit also applies to the backpropagation algorithm for training feedforward neural network. This noise benefit relies on the fact that backpropagation is indeed a type of EM algorithm (Theorem 6.1).

The secondary aim of this dissertation is to show that uniform function approximators can expand the set of model functions (likelihood functions, prior pdfs, and hyperprior pdfs) available for Bayesian inference. Bayesian statisticians often limit themselves to a small set of closed-form model functions either for ease of analysis or because they have no robust method for approximating arbitrary model functions. This dissertation shows a simple robust method for uniform model function approximation in Chapters 8 and 9. Theorem 8.2 and Theorem 9.1 guarantee that uniform approximators for model functions lead to uniform approximators for posterior pdfs.

1.1 Noisy Expectation-Maximization

The Noisy Expectation Maximization (NEM) theorem (Theorem 3.1) is the major result in this dissertation.

Theorem. [Noisy Expectation Maximization (NEM)]:
An EM iteration noise benefit occurs on average if

$$\mathbb{E}_{Y,Z,N|\theta_*}\left[\ln\left(\frac{f(Y+N,Z|\theta_k)}{f(Y,Z|\theta_k)}\right)\right] \geq 0\ . \tag{1.1}$$

The theorem gives a sufficient condition under which adding noise N to the observed data Y leads to an increase in the average convergence speed of the EM algorithm. This is the first description of a noise benefit for EM algorithms. It relies on the insight that noise can sometimes perturb the likelihood function favorably. Thus noise injection can lead to better iterative estimates for parameters. This sufficient condition is general and applies to *any* EM data model.

The first major corollary (Corollary 3.2) applies this sufficient condition to EM algorithms on the Gaussian mixture model. This results in a simple quadratic noise screening condition for the average noise benefit.

Corollary. [NEM Condition for GMMs (in 1-D)]:
The NEM sufficient condition holds for a GMM if the additive noise samples n satisfy the following algebraic condition

$$n^2 \leq 2n\left(\mu_j - y\right) \ \textit{for all GMM sub-populations} \quad j\ . \tag{1.2}$$

This quadratic condition defines the geometry (Figure 1.1) of the set of noise samples that can speed up the EM algorithm.

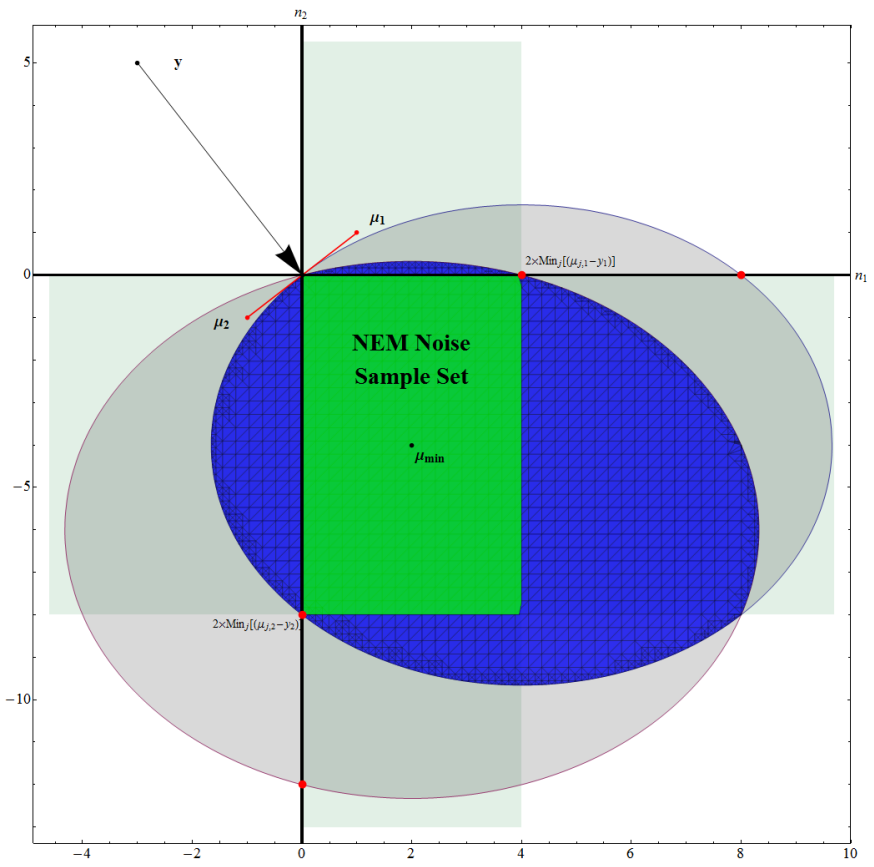

Figure 1.1: Geometry of NEM noise for a GMM. Noise samples in the blue overlapping region satisfy the NEM sufficient condition and lead to faster EM convergence. Noise samples in the green box satisfy a simpler quadratic NEM sufficient condition and also lead to faster EM convergence. Sampling from the green box is easier. This geometry is for a sample $\mathbf{y}$ of a 2-D GMM with sub-populations centered at $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. §3.2.3 and §3.2.4 discuss these geometries in more detail.

Noise injection subject to the NEM condition leads to better EM estimates on average at each iteration and faster EM convergence. Combining NEM noise injection with a noise decay per iteration leads to much faster overall EM convergence. We refer to the combination of NEM noise injection and noise cooling as the *NEM algorithm* (§3.3). A comparison of the evolution of EM and NEM algorithms on a sample estimation problem shows that the NEM algorithm reaches the stationary point of the likelihood function in 30% fewer steps than the EM algorithm (see Figure 1.2).

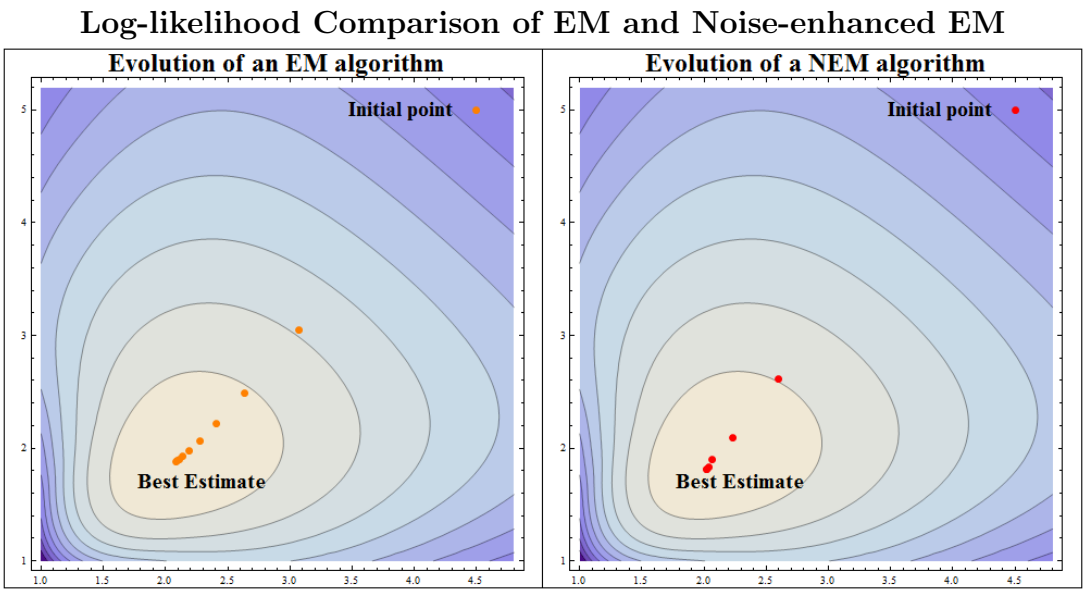

Figure 1.2: NEM noise injection can speed up the convergence of the EM algorithm. The plot shows the evolution of an EM algorithm on a log-likelihood surface with and without noise injection. Both algorithms start at the same initial estimate and converge to the same point on the log-likelihood surface. The EM algorithm converges in 10 iterations while the noise-enhanced algorithm converges in 7 iterations—30% faster than the EM algorithm.

1.2 Applications of Noisy Expectation-Maximization

Finding the NEM noise benefit led to recasting other iterative statistical algorithms as EM algorithms to allow a noise boost. The NEM theorem is a general prescriptive tool for extracting noise benefits from arbitrary EM algorithms. So these reinterpretations serve as a basis for introducing NEM noise benefits into other standard iterative estimation algorithms. This dissertation shows NEM noise benefits in three such algorithms: the k-means clustering algorithm (Chapter 4), the Baum-Welch algorithm (Chapter 5), and the backpropagation algorithm (Chapter 6).

The most important of these algorithms is the backpropagation algorithm for feedforward neural network training. We show for the first time that the backpropagation algorithm is in fact a generalized EM (GEM) algorithm (Theorem 6.1) and thus benefits from proper noise injection:

Theorem. [Backpropagation is a GEM Algorithm]:
The backpropagation update equation for a feedforward neural-network likelihood function equals the GEM update equation. Thus backpropagation is a GEM algorithm.

This theorem illustrates a general theme in recasting estimation algorithms as EM algorithms: iterative estimation algorithms that make use of missing information and increase a data log-likelihood are usually (G)EM algorithms. Chapter 6 provides proof details and simulations of NEM noise benefits for backpropagation. The NEM condition for backpropagation (Theorem 6.3) has interesting geometric properties as the backpropagation noise ball in Figure 1.3 illustrates.

Geometry of NEM Noise for Backpropagation

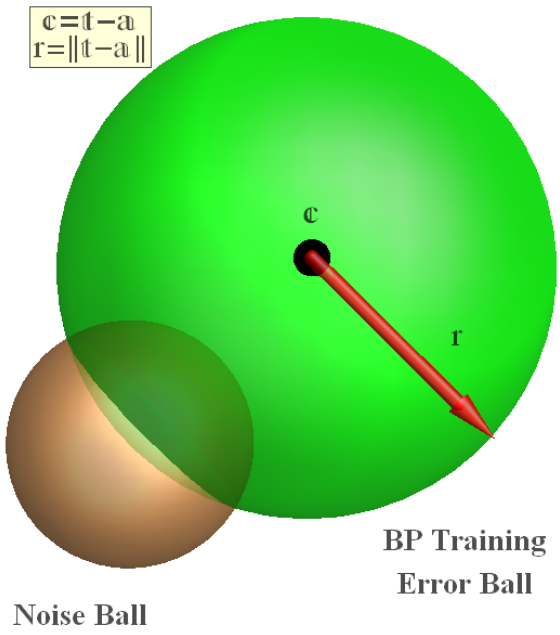

Figure 1.3: NEM noise for faster backpropagation using Gaussian output neurons. The NEM noise must fall inside the backpropagation "training error" sphere. This is the sphere with center $\mathbf{c} = \mathbf{t} - \mathbf{a}$ (the error between the target output $\mathbf{t}$ and the actual output $\mathbf{a}$) with radius $r = \|\mathbf{c}\|$. Noise from the noise ball section that intersects with the error sphere will speed up backpropagation training according to the NEM theorem. The error ball changes at each training iteration.

Deep neural networks can also benefit from the NEM noise benefit. Deep neural networks are "deep" stacks of restricted Boltzmann machines (RBMs). The depth of the network may help the network identify complicated patterns or concepts in complex data like video or speech. These deep networks are in fact bidirectional

associative memories (BAMs). The stability and fast training properties of deep networks are direct consequences of the global stability property of BAMs.

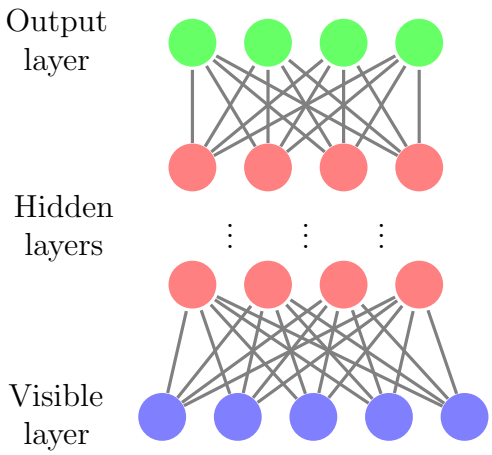

Figure 1.4: A Deep Neural Network consists of a stack of restricted Boltzmann machines (RBMs) or bidirectional associative memories (BAMs).

The so-called Contrastive Divergence algorithm is the current standard algorithm for pre-training deep networks. It is an iterative algorithm for approximate maximum likelihood estimation (§10.2.1). CD is also a GEM algorithm. Theorem 10.1 and Theorem 10.2 give the NEM noise benefit conditions for training the RBMs in a deep network. The NEM condition for RBMs shares many geometrical properties with the NEM condition for backpropagation.

1.3 Results on Bayesian Approximation

The last major results in this dissertation are the Bayesian approximation theorems in Chapters 8 and 9. They address the effects of using approximate model functions for Bayesian inference. Approximate model functions are common in Bayesian statistics because statisticians often have to estimate the true model functions from data or experts. This dissertation presents the first general proof that these model approximations do not degrade the quality of the approximate posterior pdf. Below is a combined statement of the two approximation theorems in this dissertation (Theorem 8.2 and Theorem 9.1):

Theorem. [The Unified Bayesian Approximation Theorem]:
Suppose the model functions (likelihoods g, prior h, and hyperpriors π) for a Bayesian inference problem are bounded and continuous. Suppose also that the joint product of the model functions' uniform approximators $GH\Pi$ is non-zero almost everywhere on the domain of interest $\mathcal{D}$.

Then the posterior pdf approximator $F = \frac{GH\Pi}{\int_{\mathcal{D}} GH\Pi}$ also uniformly approximates the true posterior pdf $f = \frac{gh\pi}{\int_{\mathcal{D}} gh\pi}$

This approximation theorem gives statisticians the freedom to use approximators to approximate arbitrary model functions—even model functions that have no closed functional form—without worrying about the quality of their posterior pdfs.

Statisticians can choose any uniform approximation method to reap the benefits of this theorem. Standard additive model (SAM) fuzzy systems are one such tool for uniform function approximation. Fuzzy systems can use linguistic information to build model functions. Figure 1.5 below shows an example of a SAM system approximating a pdf using 5 fuzzy rules. §8.2 discusses Fuzzy function approximation in detail. Chapter 9 addresses the complexities of approximate Bayesian inference in hierarchical or iterative inference contexts.

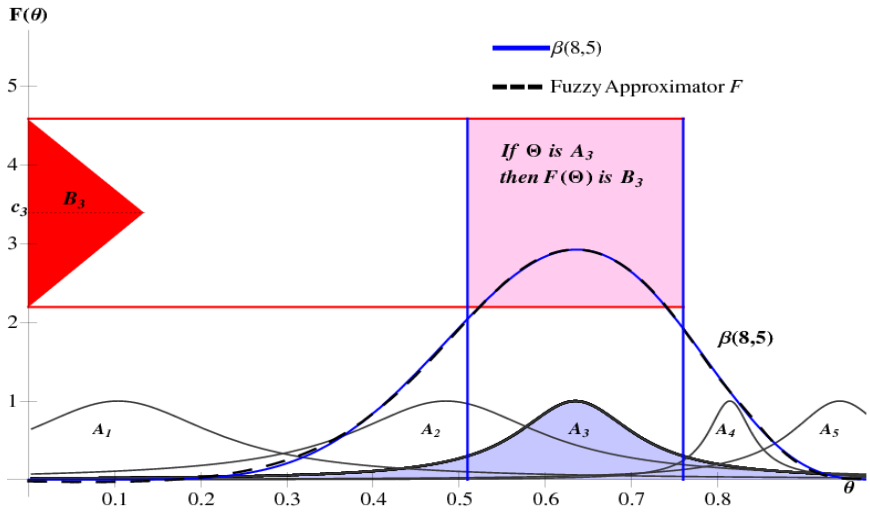

Figure 1.5: A fuzzy function approximation for a $\beta(8, 5)$ prior pdf. An adaptive SAM (standard additive model) fuzzy system tuned five fuzzy sets to give a nearly exact approximation of the beta prior. Each fuzzy rule defines a patch or 3-D surface above the input-output planar state space. The third rule has the form "If $\Theta = A_3$ then B_3" where then-part set B_3 is a fuzzy number centered at centroid c_3. This rule might have the linguistic form "If Θ is *approximately* $\frac{1}{2}$ then $F(\Theta)$ is *large*."

This dissertation also contains other minor results of note including a convergence theorem (Theorem 2.3) for a subset of minorization-maximization (MM) algorithms. MM algorithms generalize EM algorithms. But there are no published proofs of MM convergence. There is an extension of the GMM-NEM condition to mixtures of *jointly* Gaussian sub-populations (Corollary 3.4). There is also an alternate proof showing that the k-means algorithm is a specialized EM algorithm. Other proofs of this subsumption already exists in the literature. This dissertation ends with discussions about ongoing work to establish or demonstrate NEM noise benefits in genomics and medical imaging applications (§10.2.2, §10.2.3, & §10.2.4).