Fluctuation and Noise Letters
Vol. 12, No. 3 (2013) 1350012 (30 pages)

DOI: 10.1142/S0219477513500120

World Scientific
www.worldscientific.com

THE NOISY EXPECTATION–MAXIMIZATION ALGORITHM

OSONDE OSOBA

Signal and Image Processing Institute
Department of Electrical Engineering
University of Southern California
Los Angeles, California 90089-2564, USA

SANYA MITAIM

Department of Electrical and Computer Engineering
Faculty of Engineering, Thammasat University
Pathumthani 12120, Thailand

BART KOSKO

Signal and Image Processing Institute
Department of Electrical Engineering
University of Southern California
Los Angeles, California 90089-2564, USA
kosko@usc.edu

Received 6 November 2012
Accepted 12 June 2013
Published 25 September 2013

Communicated by Natalia B. Janson

We present a noise-injected version of the expectation–maximization (EM) algorithm: the noisy expectation–maximization (NEM) algorithm. The NEM algorithm uses noise to speed up the convergence of the EM algorithm. The NEM theorem shows that additive noise speeds up the average convergence of the EM algorithm to a local maximum of the likelihood surface if a positivity condition holds. Corollary results give special cases when noise improves the EM algorithm. We demonstrate these noise benefits on EM algorithms for three data models: the Gaussian mixture model (GMM), the Cauchy mixture model (CMM), and the censored log-convex gamma model. The NEM positivity condition simplifies to a quadratic inequality in the GMM and CMM cases. A final theorem shows that the noise benefit for independent identically distributed additive noise decreases with sample size in mixture models. This theorem implies that the noise benefit is most pronounced if the data is sparse.

Keywords: Noise benefit; stochastic resonance, expectation maximization algorithm; maximum likelihood; Gaussian mixture model; sparse data.

1. Introduction

The expectation–maximization (EM) algorithm [1–3] is an iterative statistical algorithm that estimates maximum-likelihood parameters from incomplete or corrupted data. This popular algorithm has a wide array of applications that includes data

clustering [4,5], automated speech recognition [6,7], medical imaging [8,9], genome sequencing [10,11], radar denoising [12], and infectious-disease tracking [13,14]. A prominent mathematical modeler has even said that the EM algorithm is "as close as data analysis algorithms come to a free lunch" [15]. But the EM algorithm can converge slowly for high-dimensional parameter spaces or when the algorithm needs to estimate large amounts of missing information [2,16].

We show that careful noise injection can increase the average convergence speed of the EM algorithm. We also derive a general sufficient condition for this EM noise benefit. Simulations show this EM noise benefit in the ubiquitous Gaussian mixture model (Fig. 1), the Cauchy mixture model, and the censored gamma model (Fig. 2). The simulations in Figs. 4 and 5 also show that the noise benefit is faint or absent if the system simply injects blind noise that ignores the sufficient condition. This suggests that the noise-benefit sufficient condition may also be necessary for some data models. The paper concludes with results that show that the noise benefit tends to occur most sharply in sparse data sets.

The EM noise benefit is an example of *stochastic resonance* in statistical signal processing. Stochastic resonance occurs when noise improves a signal system's performance [17–19,70,71]: small amounts of noise improve the performance while too much noise degrades it. Much early work on noise benefits involved natural systems in physics [20], chemistry [21,22] and biology [23–26]. This work inspired the search for noise benefits in nonlinear signal processing and statistical estimation [27–32]. The EM noise benefit does not involve a signal threshold unlike almost all SR noise benefits [18].

The next sections develop theorems and algorithms for noisy expectation–maximization (NEM). Section 2 summarizes the key facts of the EM algorithm. Section 3 introduces the theorem and corollaries that underpin the NEM algorithm. Section 4 presents the NEM algorithm and some of its variants. Section 5 presents a theorem that describes how sample size affects the NEM algorithm for mixture models when the noise is independent and identically distributed (i.i.d.). Section 5 also shows how the NEM positivity condition arises from the central limit theorem and the law of large numbers.

2. The EM Algorithm

The EM algorithm is an iterative maximum-likelihood estimation (MLE) method for estimating probability-density-function (pdf) parameters from incomplete observed data [1–3]. EM compensates for missing information by taking expectations over all missing information conditioned on the observed incomplete information and on current parameter estimates. The goal of the EM algorithm is to find the maximum-likelihood estimate $\hat{\theta}$ for the pdf parameter θ when the data Y has a parametric pdf $f(y \mid \theta)$. The maximum-likelihood estimate $\hat{\theta}$ is

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, \ell(\theta \mid y) \tag{1}$$

where $\ell(\theta \mid y) = \ln f(y \mid \theta)$ is the log-likelihood (the log of the likelihood function).

The EM scheme applies when we observe an incomplete data random variable $Y = r(X)$ instead of the complete data random variable X. The function $r : \mathcal{X} \to \mathcal{Y}$ models data corruption or information loss. $X = (Y, Z)$ can often denote the complete data X where Z is a latent or missing random variable. Z represents any statistical information lost during the observation mapping $r(X)$. This corruption makes the observed data log-likelihood $\ell(\theta \,|\, y)$ complicated and difficult to optimize directly in (1).

The EM algorithm addresses this difficulty by using the simpler complete log-likelihood $\ell(\theta \,|\, y, z)$ to derive a surrogate function $Q(\theta \,|\, \theta_k)$ for $\ell(\theta \,|\, y)$. $Q(\theta \,|\, \theta_k)$ is the average of $\ell(\theta \,|\, y, z)$ over all possible values of the latent variable Z given the observation $Y = y$ and the current parameter estimate θ_k:

$$\begin{aligned} Q(\theta \,|\, \theta_k) &= \mathbb{E}_Z[\ell(\theta \,|\, y, Z) \,|\, Y = y, \theta_k] \\ &= \int_{\mathcal{Z}} \ell(\theta \,|\, y, z) f(z \,|\, y, \theta_k) \mathrm{d}z. \end{aligned} \tag{2}$$

Dempster *et al.* [1] first showed that any θ that increases $Q(\theta \,|\, \theta_k)$ cannot reduce the log-likelihood difference $\ell(\theta \,|\, y) - \ell(\theta_k \,|\, y)$. This "ascent property" led to an iterative method that performs gradient ascent on the log-likelihood $\ell(\theta \,|\, y)$. This result underpins the EM algorithm and its many variants [4, 33–37].

We use the following notation for expectations to avoid cumbersome equations:

$$\begin{aligned} \mathbb{E}_{S|t,\phi}[g(S, t, \theta)] &\equiv \mathbb{E}_S[g(S, T, \theta) \,|\, T = t, \phi] \\ &= \int g(s, t, \theta) f_{S|T}(s \,|\, t, \phi) \mathrm{d}s, \end{aligned}$$

where S and T are random variables, ϕ and θ are deterministic parameters, and g is integrable with respect to the conditional pdf $f_{S|T}$.

A standard EM algorithm performs the following two steps iteratively on a vector $\mathbf{y} = (y_1, \ldots, y_M)$ of observed random samples of Y:

Algorithm 1 $\hat{\theta}_{EM} = \text{EM-Estimate}(\mathbf{y})$

1: **E-Step:** $Q(\theta \,|\, \theta_k) \leftarrow \mathbb{E}_{Z|y,\theta_k}[\ln f(\mathbf{y}, \mathbf{Z} \,|\, \theta)]$
2: **M-Step:** $\theta_{k+1} \leftarrow \text{argmax}_\theta\{Q(\theta \,|\, \theta_k)\}$

The algorithm stops when successive estimates differ by less than a given tolerance: $\|\theta_k - \theta_{k-1}\| < 10^{-\text{tol}}$ or when $\|\ell(\theta_k \,|\, y) - \ell(\theta_{k-1} \,|\, y)\| < \varepsilon$. The EM algorithm converges to a local maximum θ_* [38, 39]: $\theta_k \to \theta_*$.

The EM algorithm is in fact a family of MLE methods for working with incomplete data models. Such incomplete data models include mixture models [40, 41], censored exponential family models [42], and mixtures of censored models [43]. The next subsection describes examples of such incomplete data models.

Users have a good deal of freedom when they specify the EM complete random variables X and latent random variables Z for probabilistic models on the observed

data Y. This freedom in model selection allows users to recast many disparate algorithms as EM algorithms [4,44–46]. Changes to the E and M steps give another degree of freedom for the EM scheme [1,36,37,47,48].

2.1. *Incomplete data models for EM: Mixture and censored gamma models*

We now describe two general examples of incomplete data models: finite mixture models and censored gamma models. Q-functions specify EM algorithms for the data models. We compare the performance of the EM and NEM algorithms on these data models later in the paper.

A finite mixture model [40,49] is a convex combination of a finite set of sub-populations. The sub-population pdfs come from the same parametric family. Mixture models are useful for modeling mixed populations for statistical applications such as clustering, pattern recognition, and acceptance testing. We use the following notation for mixture models. Y is the observed mixed random variable. K is the number of sub-populations. $Z \in \{1, \ldots, K\}$ is the hidden sub-population index random variable. The convex population mixing proportions $\alpha_1, \ldots, \alpha_K$ form a discrete pdf for $Z : P(Z = j) = \alpha_j$. The pdf $f(y \mid Z = j, \theta_j)$ is the pdf of the jth sub-population where $\theta_1, \ldots, \theta_K$ are the pdf parameters for each sub-population. Θ is the vector of all model parameters $\Theta = \{\alpha_1, \ldots, \alpha_K, \theta_1, \ldots, \theta_K\}$. The joint pdf $f(y, z \mid \Theta)$ is

$$f(y, z \mid \Theta) = \sum_{j=1}^{K} \alpha_j \; f(y \mid j, \theta_j)\delta[z - j] \tag{3}$$

where δ is the Kronecker delta function such that $\delta = 1$ if $x = 0$ and such that $\delta = 0$ otherwise. The marginal pdf $f(y \mid \Theta)$ for Y and the conditional pdf $p(j \mid y, \theta)$ for Z given y are

$$f(y \mid \Theta) = \sum_{j} \alpha_j f(y \mid j, \theta_j) \tag{4}$$

$$\text{and} \quad p_Z(j \mid y, \Theta) = \frac{\alpha_j f(y \mid Z = j, \theta_j)}{f(y \mid \Theta)} \tag{5}$$

by Bayes theorem. We rewrite the joint pdf in exponential form for ease of analysis:

$$f(y, z \mid \Theta) = \exp\left[\sum_{j} [\ln(\alpha_j) + \ln f(y \mid j, \theta_j)]\delta[z - j]\right]. \tag{6}$$

Thus

$$\ln f(y, z \mid \Theta) = \sum_{j} \delta[z - j] \ln[\alpha_j f(y \mid j, \theta_j)]. \tag{7}$$

EM algorithms for finite mixture models estimate Θ using the sub-population index Z as the latent variable. An EM algorithm on a finite mixture model uses (5)

to derive the Q-function:

$$Q(\Theta \mid \Theta_k) = \mathbb{E}_{Z|y,\Theta_k}[\ln f(y, Z \mid \Theta)] \tag{8}$$

$$= \sum_z \left(\sum_j \delta[z-j] \ln[\alpha_j f(y \mid j, \theta_j)] \right) p_Z(z \mid y, \Theta_k) \tag{9}$$

$$= \sum_j \ln[\alpha_j f(y \mid j, \theta_j)] p_Z(j \mid y, \Theta_k). \tag{10}$$

Another incomplete data model is the censored gamma model [42, 50]. It produces censored samples y of a gamma complete random variable X. Censorship refers to clipped or interval-limited measurement. Censored gammas can model time-limited medical trials and product reliability [50, 51]. The complete data pdf in this model is the gamma pdf $\gamma(\alpha, \theta)$

$$f(x \mid \theta) = \frac{x^{\alpha-1} \exp(-x/\theta)}{\Gamma(\alpha)\theta^\alpha}. \tag{11}$$

The complete data X does not admit a tractable specification for a latent variable Z. But we can still write a Q-function by taking expectations of the complete X given the observed Y. This is a more general formulation of the Q-function. The E-step for estimating θ is

$$Q(\theta \mid \theta_k) = \mathbb{E}_{X|y,\theta_k}[\ln f(X \mid \theta)] \tag{12}$$

$$= -\ln(\Gamma(\alpha)) - \alpha \ln \theta + (-\theta^{-1} + (\alpha - 1))\mathbb{E}_{X|y,\theta_k}[X] \tag{13}$$

where the samples y are censored observations of X.

2.2. *Noise benefits in the EM algorithm*

Theorem 1 below states a general sufficient condition for a noise benefit in the average convergence time or 'ascent' of the EM algorithm. Figure 1 shows a simulation instance of this theorem for the important EM case of Gaussian mixture densities. Small values of the noise variance reduce convergence time while larger values increase it. This U-shaped noise benefit is the non-monotonic signature of stochastic resonance. The optimal noise speeds convergence by 27.2%. Other simulations on multi-dimensional GMMs have shown speed increases of up to 40%.

The EM noise benefit differs from almost all SR noise benefits because it does not involve the use of a signal threshold [18]. The EM noise benefit also differs from most SR noise benefits because the additive noise can depend on the signal. Independent noise can lead to weaker noise benefits than dependent noise in EM algorithms. This also happens with enhanced convergence in noise-injected Markov chains [32]. Figure 4 shows that the proper dependent noise outperforms independent noise at all tested sample sizes for a Gaussian mixture model. The dependent noise model converges up to 14.5% faster than the independent noise model.

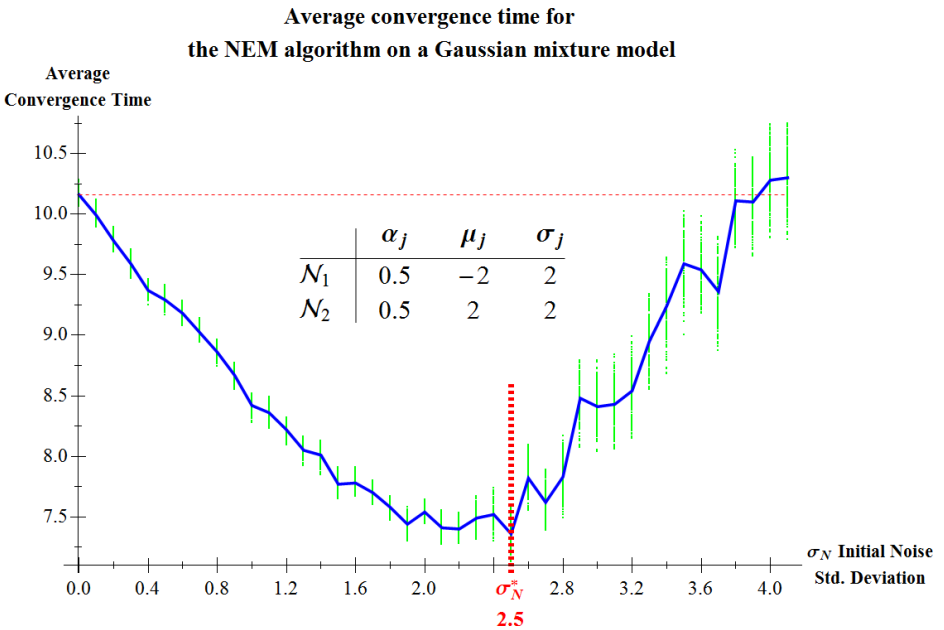

Fig. 1. EM noise benefit for a Gaussian mixture model. The plot used the noise-annealed NEM algorithm. Low intensity initial noise decreased convergence time while higher intensity starting noise increased it. The optimal initial noise level had standard deviation $\sigma_N^* = 2.5$. The average optimal NEM speed-up over the noiseless EM algorithm was 27.2%. This NEM procedure added noise with a cooling schedule. The noise cools at an inverse-square rate. The Gaussian mixture density was a convex combination of two normal sub-populations $\mathcal{N}_1$ and $\mathcal{N}_2$. The simulation used 200 samples of the mixture normal distribution to estimate the standard deviations of the two sub-populations. The additive noise used samples of zero-mean normal noise with standard deviation σ_N screened through the GMM–NEM condition in (42). Each sampled point on the curve is the average of 100 trials. The vertical bars are 95% bootstrap confidence intervals for the mean convergence time at each noise level.

The idea behind the EM noise benefit is that sometimes noise can make the signal data more probable. This occurs at the local level when

$$f(y + n \,|\, \theta) > f(y \,|\, \theta) \tag{14}$$

for pdf f, realization y of random variable Y, realization n of random noise N, and parameter θ. This condition holds if and only if the logarithm of the pdf ratio is positive:

$$\ln\left(\frac{f(y + n \,|\, \theta)}{f(y \,|\, \theta)}\right) > 0. \tag{15}$$

The logarithmic condition (15) in turn occurs much more generally if it holds only on average with respect to all the pdfs involved in the EM algorithm:

$$\mathbb{E}_{Y,Z,N|\theta_*}\left[\ln \frac{f(Y + N, Z \,|\, \theta_k)}{f(Y, Z \,|\, \theta_k)}\right] \geq 0 \tag{16}$$

where random variable Z represents missing data in the EM algorithm and where θ_* is the limit of the EM estimates $\theta_k : \theta_k \to \theta_*$. The positivity condition (16) is precisely the sufficient condition for a noise benefit in Theorem 1 below. We call this theorem the NEM or Noisy EM Theorem.

3. NEM Theorems

We define the EM noise benefit by first defining a modified surrogate log-likelihood function

$$Q_N(\theta \,|\, \theta_k) = \mathbb{E}_{Z|y,\theta_k}[\ln f(y + N, Z \,|\, \theta)] \tag{17}$$

and its maximizer

$$\theta_{k+1,N} = \underset{\theta}{\operatorname{argmax}} \{Q_N(\theta \,|\, \theta_k)\}.$$

The modified surrogate log-likelihood $Q_N(\theta \,|\, \theta_k)$ equals the regular surrogate log-likelihood $Q(\theta \,|\, \theta_k)$ when $N = 0$. $Q(\theta \,|\, \theta_*)$ is the final surrogate log-likelihood given the optimal EM estimate θ_*. So θ_* maximizes $Q(\theta \,|\, \theta_*)$. Thus

$$Q(\theta_* \,|\, \theta_*) \geq Q(\theta \,|\, \theta_*) \quad \text{for all } \theta. \tag{18}$$

An EM noise benefit occurs when the noisy surrogate log-likelihood $Q_N(\theta_k \,|\, \theta_*)$ is closer to the optimal value $Q(\theta_* \,|\, \theta_*)$ than the regular surrogate log-likelihood $Q(\theta_k \,|\, \theta_*)$ is. This holds when

$$Q_N(\theta_k \,|\, \theta_*) \geq Q(\theta_k \,|\, \theta_*) \tag{19}$$

$$\text{or } (Q(\theta_* \,|\, \theta_*) - Q(\theta_k \,|\, \theta_*)) \geq (Q(\theta_* \,|\, \theta_*) - Q_N(\theta_k \,|\, \theta_*)). \tag{20}$$

So the noisy perturbation $Q_N(\theta \,|\, \theta_k)$ of the current surrogate log-likelihood $Q(\theta \,|\, \theta_k)$ is a better log-likelihood function for the data than Q is itself. An average noise benefit results when we take expectations on both sides of inequality (20):

$$\mathbb{E}_N[Q(\theta_* \,|\, \theta_*) - Q(\theta_k \,|\, \theta_*)] \geq \mathbb{E}_N[Q(\theta_* \,|\, \theta_*) - Q_N(\theta_k \,|\, \theta_*)]. \tag{21}$$

The average noise benefit (21) occurs when the final EM pdf $f(y, z \,|\, \theta_*)$ is closer in relative-entropy to the noisy pdf $f(y + N, z \,|\, \theta_k)$ than it is to the noiseless pdf $f(y, z \,|\, \theta_k)$. Define the relative-entropy pseudo-distances

$$c_k(N) = D(f(y, z \,|\, \theta_*) \| f(y + N, z \,|\, \theta_k)) \tag{22}$$

$$c_k = c_k(0) = D(f(y, z \,|\, \theta_*) \| f(y, z \,|\, \theta_k)). \tag{23}$$

Then noise benefits the EM algorithm when

$$c_k \geq c_k(N) \tag{24}$$

holds for the relative-entropy pseudo-distances. The relative entropy itself has the form [52]

$$D(h(u, v) \| g(u, v)) = \int_{\mathcal{U},\mathcal{V}} \ln \left[\frac{h(u, v)}{g(u, v)}\right] h(u, v) du\, dv \tag{25}$$

for positive pdfs h and g over the same support. Convergent sums can replace the integrals as needed.

3.1. *NEM theorem*

The NEM theorem below uses the following notation. The noise random variable N has pdf $f(n \mid y)$. So the noise N can depend on the data Y. Independence implies that the noise pdf becomes $f(n \mid y) = f_N(n)$. $\{\theta_k\}$ is a sequence of EM estimates for θ. $\theta_* = \lim_{k\to\infty} \theta_k$ is the converged EM estimate for θ. Assume that the differential entropy [52] of all random variables is finite. Assume also that the additive noise keeps the data in the likelihood function's support. The appendix gives the proof of the NEM theorem and its three corollaries.

Theorem 1. *Noisy expectation–maximization (NEM). The noise benefit for an EM estimation iteration*

$$(Q(\theta_* \mid \theta_*) - Q(\theta_k \mid \theta_*)) \geq (Q(\theta_* \mid \theta_*) - Q_N(\theta_k \mid \theta_*)) \tag{26}$$

occurs on average if

$$\mathbb{E}_{Y,Z,N|\theta_*}\left[\ln\left(\frac{f(Y+N, Z \mid \theta_k)}{f(Y, Z \mid \theta_k)}\right)\right] \geq 0. \tag{27}$$

The NEM theorem also applies to EM algorithms that use the complete data as their latent random variable. The proofs for these cases follow from the proof in the appendix. The NEM positivity condition in these models changes to

$$\mathbb{E}_{X,Y,N|\theta_*}\left[\ln\left(\frac{f(X+N \mid \theta_k)}{f(X \mid \theta_k)}\right)\right] \geq 0. \tag{28}$$

The NEM theorem implies that each iteration of a suitably noisy EM algorithm moves closer on average toward the EM estimate θ_* than does the corresponding noiseless EM algorithm [53]. This holds because the positivity condition (27) implies that $\mathbb{E}_N[c_k(N)] \leq c_k$ at each step k since c_k does not depend on N from (23). The NEM algorithm produces larger improvements of the estimate on average than does the noiseless EM algorithm for any number k of steps.

The NEM theorem's stepwise noise benefit leads to a noise benefit at any point in the sequence of NEM estimates. This is because we get the following inequalities when the expected value of inequality (19) takes the form

$$Q(\theta_k \mid \theta_*) \leq \mathbb{E}_N[Q_N(\theta_k \mid \theta_*)] \quad \text{for any } k. \tag{29}$$

Thus

$$Q(\theta_* \mid \theta_*) - Q(\theta_k \mid \theta_*) \geq Q(\theta_* \mid \theta_*) - \mathbb{E}_N[Q_N(\theta_k \mid \theta_*)] \quad \text{for any } k. \tag{30}$$

The EM (NEM) sequence converges when the left (right) side of inequality (30) equals zero. Inequality (30) implies that the difference on the right side is closer to zero at any step k.

NEM sequence convergence is even stronger if the noise N_k decays to zero as the iteration count k grows to infinity. This noise annealing implies $N_k \to 0$ with probability one. Continuity of Q as a function of Y implies that $Q_{N_k}(\theta \,|\, \theta_k) \to Q(\theta \,|\, \theta_k)$ as $N_k \to 0$. This holds because $Q(\theta \,|\, \theta_k) = \mathbb{E}_{Z|y,\theta_k}[\ln f(y, Z \,|\, \theta)]$ and because the continuity of Q implies that

$$\begin{aligned} \lim_{N \to 0} Q_N(\theta \,|\, \theta_k) &= \mathbb{E}_{Z|y,\theta_k}\left[\ln f\left(\lim_{N \to 0}(y + N), Z \,|\, \theta\right)\right] \\ &= \mathbb{E}_{Z|y,\theta_k}[\ln f(y, Z \,|\, \theta)] = Q(\theta \,|\, \theta_k). \end{aligned} \tag{31}$$

The evolution of EM algorithms guarantees that $\lim_{k \to \infty} Q(\theta_k \,|\, \theta_*) = Q(\theta_* \,|\, \theta_*)$. This gives the probability-one limit

$$\lim_{k \to \infty} Q_{N_k}(\theta_k \,|\, \theta_*) = Q(\theta_* \,|\, \theta_*). \tag{32}$$

So for any $\epsilon > 0$ there exists a k_0 such that for all $k > k_0$:

$$\begin{aligned} |Q(\theta_k \,|\, \theta_*) - Q(\theta_* \,|\, \theta_*)| < \epsilon \quad &\text{and} \\ |Q_{N_k}(\theta_k \,|\, \theta_*) - Q(\theta_* \,|\, \theta_*)| < \epsilon \quad &\text{with probability one.} \end{aligned} \tag{33}$$

Inequalities (29) and (33) imply that $Q(\theta_k \,|\, \theta_*)$ is ϵ-close to its upper limit $Q(\theta_* \,|\, \theta_*)$ and

$$\mathbb{E}[Q_{N_k}(\theta_k \,|\, \theta_*)] \geq Q(\theta_k \,|\, \theta_*) \quad \text{and} \quad Q(\theta_* \,|\, \theta_*) \geq Q(\theta_k \,|\, \theta_*). \tag{34}$$

So the NEM and EM algorithms converge to the same fixed point by (32). And the inequalities (34) imply that NEM estimates are closer on average to optimal than EM estimates are at any step k.

3.2. *NEM*: *Dominated densities and mixture densities*

The first corollary of Theorem 1 gives a dominated-density condition that satisfies the positivity condition (27) in the NEM theorem. This strong pointwise condition is a direct extension of the pdf inequality in (14) to the case of an included latent random variable Z.

Corollary 1.

$$\mathbb{E}_{Y,Z,N|\theta_*}\left[\ln \frac{f(Y + N, Z \,|\, \theta)}{f(Y, Z \,|\, \theta)}\right] \geq 0 \quad \textit{if } f(y + n, z \,|\, \theta) \geq f(y, z \,|\, \theta) \tag{35}$$

for almost all y, z, and n.

We use Corollary 1 to derive conditions on the noise N that produce NEM noise benefits for mixture models. NEM mixture models use two special cases of Corollary 1. We state these special cases as Corollaries 2 and 3 below. The corollaries use the finite mixture model notation in Sec. 2.1. Recall that the joint pdf of Y and

Z is

$$f(y, z \mid \theta) = \sum_j \alpha_j f(y \mid j, \theta) \delta[z - j]. \tag{36}$$

Define the population-wise noise likelihood difference as

$$\Delta f_j(y, n) = f(y + n \mid j, \theta) - f(y \mid j, \theta). \tag{37}$$

Corollary 1 implies that noise benefits the mixture model estimation if the dominated-density condition holds:

$$f(y + n, z \mid \theta) \geq f(y, z \mid \theta). \tag{38}$$

This occurs if

$$\Delta f_j(y, n) \geq 0 \quad \text{for all } j. \tag{39}$$

The Gaussian mixture model (GMM) uses normal pdfs for the sub-population pdfs [40, 54]. Corollary 2 states a simple quadratic condition that ensures that the noisy sub-population pdf $f(y+n \mid Z = j, \theta)$ dominates the noiseless sub-population pdf $f(y \mid Z = j, \theta)$ for GMMs. The additive noise samples n depend on the data samples y.

Corollary 2. *Suppose* $Y|_{Z=j} \sim \mathcal{N}(\mu_j, \sigma_j^2)$ *and thus* $f(y \mid j, \theta)$ *is a normal pdf. Then*

$$\Delta f_j(y, n) \geq 0 \tag{40}$$

holds if

$$n^2 \leq 2n(\mu_j - y). \tag{41}$$

Now apply the quadratic condition (41) to (39). Then $f(y + n, z \mid \theta) \geq f(y, z \mid \theta)$ holds when

$$n^2 \leq 2n(\mu_j - y) \quad \text{for all } j. \tag{42}$$

The inequality (42) gives the GMM–NEM noise benefit condition (misstated in [55] but corrected in [56]) when the NEM algorithm more quickly estimates the standard deviations σ_j than does noiseless EM. This can also benefit expectation–conditional–maximization [34] methods.

Figure 1 shows a simulation instance of noise benefits for GMM parameter estimation based on the GMM–NEM condition (42). The simulation estimates the sub-population standard deviations σ_1 and σ_2 from 200 samples of a Gaussian mixture of two 1D sub-populations with known means $\mu_1 = -2$ and $\mu_2 = 2$ and mixing proportions $\alpha_1 = 0.5$ and $\alpha_2 = 0.5$. The true standard deviations are $\sigma_1^* = 2$ and $\sigma_2^* = 2$. Each EM and NEM procedure starts at the same initial point with $\sigma_1(0) = 4.5$ and $\sigma_2(0) = 5$. The simulation runs NEM on 100 GMM data sets for each noise level σ_N and counts the number of iterations before convergence for each instance. The average of these iteration counts is the *average convergence time* at that noise level σ_N. The EM and NEM simulations use the *NArgMax* numerical

maximization routine in Mathematica for the M-step. Simulations (not shown) also confirm that both the Cauchy mixture model (CMM) and non-Gaussian noise show a similar pronounced noise benefit.

Corollary 3 gives a similar quadratic condition for the Cauchy mixture model.

Corollary 3. *Suppose* $Y|_{Z=j} \sim \mathcal{C}(m_j, d_j)$ *and thus* $f(y \mid j, \theta)$ *is a Cauchy pdf. Then*

$$\Delta f_j(y, n) \geq 0 \tag{43}$$

holds if

$$n^2 \leq 2n(m_j - y). \tag{44}$$

Again apply the quadratic condition (44) to (39). Then $f(y+n, z \mid \theta) \geq f(y, z \mid \theta)$ holds when

$$n^2 \leq 2n(m_j - y) \quad \text{for all } j. \tag{45}$$

Both quadratic NEM inequality conditions in (42) and (45) reduce to the following inequality (replace μ with m for the CMM case):

$$n[n - 2(\mu_j - y)] \leq 0 \quad \text{for all } j. \tag{46}$$

So the noise n must fall in the set where the *parabola* $n^2 - 2n(\mu_j - y)$ is negative for all j. There are two possible solution sets for (46) depending on the values of μ_j and y. These solution sets are

$$N_j^-(y) = [2(\mu_j - y), 0] \tag{47}$$

$$N_j^+(y) = [0, 2(\mu_j - y)]. \tag{48}$$

The goal is to find the set $N(y)$ of n values that satisfy the inequality in (42) for *all* j:

$$N(y) = \bigcap_j N_j(y) \tag{49}$$

where $N_j(y) = N_j^+(y)$ or $N_j(y) = N_j^-(y)$. $N(y) \neq \{0\}$ holds only when the sample y lies on one side of all sub-population means (or location parameters) μ_j. This holds for

$$y < \mu_j \text{ for all } j \quad \text{or} \quad y < \mu_j \text{ for all } j. \tag{50}$$

The NEM noise N takes values in $\bigcap_j N_j^-$ if the data sample y falls to the right of all sub-population means ($y > \mu_j$ for all j). The NEM noise N takes values in $\bigcap_j N_j^+$ if the data sample y falls to the left of all sub-population means ($y < \mu_j$ for all j). And $N = 0$ is the only valid value for N when y falls between sub-population means. Thus the noise N tends to pull the data sample y away from the tails and towards the cluster of sub-population means (or locations).

3.3. *NEM for log-convex densities*

EM algorithms can satisfy the positivity condition (27) if they use the proper noise N. They can also satisfy the condition if the data model has an amenable complete data pdf $f(x \mid \theta)$. Inequalities (42) and (45) can sculpt the noise N to satisfy (27) for Gaussian and Cauchy mixture models. The next corollary shows how the complete data pdf can induce a noise benefit. The corollary states that a log-convex complete pdf satisfies (27) when the noise is zero-mean. The corollary applies to data models with more general complete random variables X. These include models whose complete random variables X do not decompose into the direct product $X = (Y, Z)$. Examples include censored models that use the unobserved complete random variable as the latent random variable $Z : Z = X$ [3, 43, 50].

Corollary 4. *Suppose that $f(x \mid \theta)$ is log-convex in x, N is independent of X, and $\mathbb{E}_N[N] = 0$. Then*

$$\mathbb{E}_{X,N|\theta_*}\left[\ln \frac{f(X+N \mid \theta_k)}{f(X \mid \theta_k)}\right] \geq 0 \tag{51}$$

and thus the noise benefit $c_k(N) \leq c_k$ holds for all k.

A related corollary gives a similar noise benefit if we replace the zero-mean additive noise with unit-mean multiplicative noise. The noise is also independent of the data.

The right-censored gamma data model gives a log-convex data model when the α-parameter of its complete pdf lies in the interval $(0, 1)$. This holds because the gamma pdf is log-convex when $0 < \alpha < 1$. Log-convex densities often model data with decreasing hazard rates in survival analysis applications [51, 57, 58]. Section 2.1 describes the gamma data model and EM algorithm. Figure 2 shows a simulation instance of noise benefits for a log-convex model. The simulation estimates the θ parameter from right-censored samples of a $\gamma(0.65, 4)$ pdf. Samples are censored to values below a threshold of 4.72. The average optimal NEM speed-up over the noiseless EM algorithm is about 13.3%.

4. The NEM Algorithm

The NEM theorem and its corollaries give a general method for modifying the noiseless EM algorithm. The NEM theorem also implies that on average these NEM variants outperform the noiseless EM algorithm.

Algorithm 2 gives the NEM algorithm schema. The operation NEMNoiseSample($\mathbf{y}$) generates noise samples that satisfy the NEM condition for the current data model. The noise sampling distribution depends on the vector of random samples $\mathbf{y}$ in the Gaussian and Cauchy mixture models. The noise can have any value in the NEM algorithm for censored gamma models. The E-Step takes a conditional expectation of a function of the noisy data samples $\mathbf{y}_\dagger$ given the noiseless data samples $\mathbf{y}$.

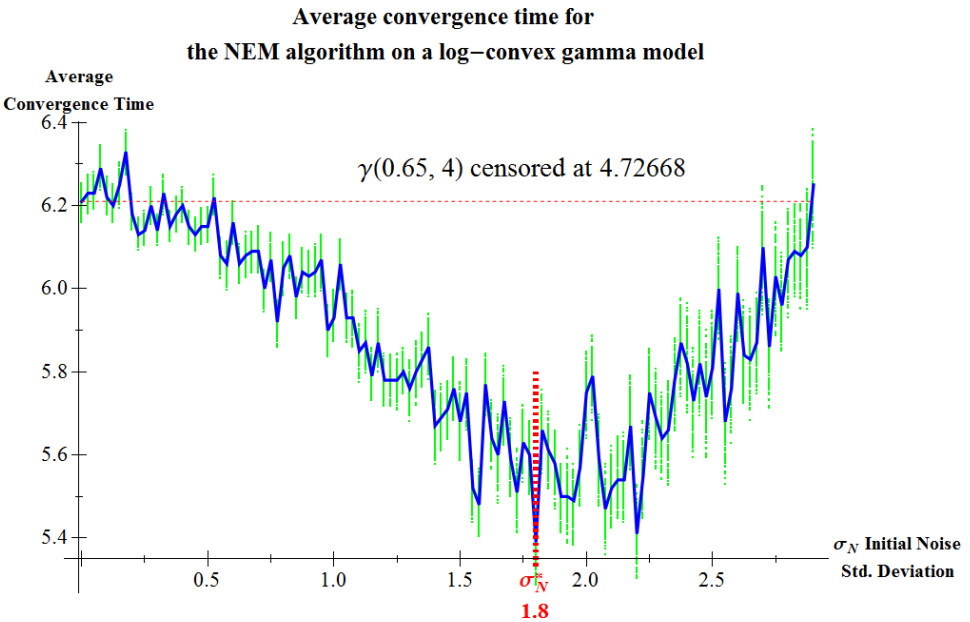

Fig. 2. EM noise benefit for a log-convex censored gamma model. This plot used the annealed-noise NEM algorithm. The average optimal NEM speed-up over the noiseless EM algorithm is about 13.3%. Low intensity initial noise decreased convergence time while higher intensity starting noise increased it. This NEM procedure added cooled i.i.d. normal noise that was *independent* of the data. The noise cooled at an inverse-square rate. The log-convex gamma distribution was a $\gamma(\alpha, \theta)$ distribution with $\alpha < 1$. The censored gamma EM estimated the θ parameter. The model used 375 censored gamma samples. Each sampled point on the curve is the mean of 100 trials. The vertical bars are 95% bootstrap confidence intervals for the mean convergence time at each noise level.

Algorithm 2 $\hat{\theta}_{\text{NEM}} = \text{NEM-Estimate}(\mathbf{y})$

Require: $\mathbf{y} = (y_1, \ldots, y_M)$: vector of observed incomplete data
Ensure: $\hat{\theta}_{NEM}$: NEM estimate of parameter θ

1: **while** $(\|\theta_k - \theta_{k-1}\| \geq 10^{-\text{tol}})$ **do**
2: $\quad$ **N$_\mathbf{S}$-Step:** $\mathbf{n} \leftarrow k^{-\tau} \times \text{NEMNoiseSample}(\mathbf{y})$
3: $\quad$ **N$_\mathbf{A}$-Step:** $\mathbf{y}_\dagger \leftarrow \mathbf{y} + \mathbf{n}$
4: $\quad$ **E-Step:** $Q(\theta \,|\, \theta_k) \leftarrow \mathbb{E}_{Z|y,\theta_k}[\ln f(\mathbf{y}_\dagger, \mathbf{Z} \,|\, \theta)]$
5: $\quad$ **M-Step:** $\theta_{k+1} \leftarrow \underset{\theta}{\text{argmax}} \{Q(\theta \,|\, \theta_k)\}$
6: $\quad$ $k \leftarrow k + 1$
7: **end while**
8: $\hat{\theta}_{\text{NEM}} \leftarrow \theta_k$

A deterministic decay factor $k^{-\tau}$ scales the noise on the kth iteration. τ is the noise decay rate. The decay factor $k^{-\tau}$ reduces the noise at each new iteration. This factor drives the noise N_k to zero as the iteration step k increases. The simulations in this paper use $\tau = 2$ for demonstration. Values between $\tau = 1$ and $\tau = 3$ also work. N_k still needs to satisfy the NEM condition for the data model. The cooling factor $k^{-\tau}$ must not cause the noise samples to violate the NEM condition. This usually means that $0 < k^{-\tau} \leq 1$ and that the NEM condition solution set is closed with respect to contractions.

The decay factor reduces the NEM estimator's jitter around its final value. This is important because the EM algorithm converges to fixed points. So excessive estimator jitter prolongs convergence time even when the jitter occurs near the final solution. The simulations in this paper use polynomial decay factors instead of logarithmic cooling schedules found in annealing applications [59–63].

The NEM algorithm inherits some variants from the classical EM algorithm schema. A NEM adaptation to the generalized expectation–maximization (GEM) algorithm is one of the simpler variations. The GEM algorithm replaces the EM maximization step with a gradient ascent step. The noisy generalized expectation–maximization (NGEM) algorithm (Algorithm 3) uses the same M-step. The NEM algorithm schema also allows for some variations outside the scope of the EM algorithm. These involve modifications to the noise sampling step $\mathbf{N_S}$**-Step** or to the noise addition step $\mathbf{N_A}$**-Step**.

One such modification does not require an additive noise term n_i for each y_i. This is useful when the NEM condition is stringent because then noise sampling can be time intensive. This variant changes the $\mathbf{N_S}$**-Step** by picking a random or deterministic sub-selection of the $\mathbf{y}$ vector that the noise will modify. Then it samples the noise subject to the NEM condition for those sub-selected samples. This is the partial noise addition NEM (PNA-NEM).

The NEM noise generating procedure NEMNoiseSample($\mathbf{y}$) returns a NEM-compliant noise sample n at a given noise level σ_N for each data sample y. This procedure changes with the EM data model. The noise generating procedure for

Algorithm 3 Modified M-Step for NGEM:

1: **M-Step:** $\theta_{k+1} \leftarrow \tilde{\theta}$ such that $Q(\tilde{\theta} \,|\, \theta_k) \geq Q(\theta_k \,|\, \theta_k)$

Algorithm 4 Modified $\mathbf{N_S}$**-Step** for PNA-NEM

$\mathcal{I} \leftarrow \{1 \dots M\}$
$\mathcal{J} \leftarrow$ SubSelection($\mathcal{I}$)
for all $\iota \in \mathcal{J}$ **do**
  $n_\iota \leftarrow k^{-\tau} \times$ NEMNoiseSample(y_ι)
end for

the GMMs and CMMs comes from Corollaries 2 and 3. We used the following 1D noise generating procedure for the GMM simulations:

NEMNoiseSample for GMM–NEM and CMM–NEM

Require: y and σ_N : current data sample and noise level
Ensure: n : noise sample satisfying NEM condition
$N(y) \leftarrow \bigcap_j N_j(y)$
n is a sample from the distribution $TN(0, \sigma_N \,|\, N(y))$

where $TN(0, \sigma_N \,|\, N(y))$ is the normal distribution $N(0, \sigma_N)$ truncated to the support set $N(y)$. The set $N(y)$ is the interval intersection from (49). Multi-dimensional versions of the generator can apply the procedure component-wise.

5. NEM Sample Size Effects: Gaussian and Cauchy Mixture Models

The noise-benefit effect depends on the size of the GMM data set. Analysis of this effect depends on the probabilistic event that the noise satisfies the GMM–NEM condition for the entire sample set. This analysis also applies to the Cauchy mixture model because its NEM condition is the same as the GMM's. Define A_k as the event that the noise N satisfies the GMM–NEM condition for the kth data sample:

$$A_k = \{N^2 \leq 2N(\mu_j - y_k) \,|\, \forall\, j\}. \tag{52}$$

Then define the event A_M that noise random variable N satisfies the GMM–NEM condition for each data sample as

$$A_M = \bigcap_k^M A_k \tag{53}$$

$$= \{N^2 \leq 2N(\mu_j - y_k) \,|\, \forall\, j \text{ and } \forall\, k\}. \tag{54}$$

This construction is useful for analyzing NEM when we use *independent and identically distributed* (*i.i.d.*) noise $N_k \stackrel{d}{=} N$ for all y_k while still enforcing the NEM condition.

5.1. *Large sample size effects*

The next theorem shows that the set A_M shrinks to the singleton set $\{0\}$ as the number M of samples in the data set grows. So the probability of satisfying the NEM condition with i.i.d. noise samples goes to zero as $M \to \infty$ with probability one.

Theorem 2. *Large Sample GMM–NEM and CMM–NEM.*
Assume that the noise random variables are i.i.d. Then the set of noise values

$$A_M = \{N^2 \leq 2N(\mu_j - y_k) \,|\, \forall\, j \text{ and } \forall\, k\} \tag{55}$$

that satisfy the Gaussian NEM condition for all data samples y_k decreases with probability one to the set $\{0\}$ as $M \to \infty$:

$$P\Big(\lim_{M\to\infty} A_M = \{0\}\Big) = 1. \tag{56}$$

The proof shows that larger sample sizes M place tighter bounds on the size of A_M with probability one. The bounds shrink A_M all the way down to the singleton set $\{0\}$ as $M \to \infty$. A_M is the set of values that identically distributed noise N can take to satisfy the NEM condition for all y_k. $A_M = \{0\}$ means that N_k must be zero for all k because the N_k are *identically* distributed. This corresponds to cases where the NEM theorem cannot guarantee improvement over the regular EM using just i.i.d. noise. So identically distributed noise has limited use in the GMM–NEM and CMM–NEM frameworks.

Theorem 2 is a "probability-one" result. But it also implies the following weaker convergence-in-probability result. Suppose $\tilde{N}$ is an arbitrary *continuous* random variable. Then the probability $P(\tilde{N} \in A_M)$ that $\tilde{N}$ satisfies the NEM condition for all samples falls to $P(\tilde{N} \in \{0\}) = 0$ as $M \to \infty$. Figure 3 shows a Monte Carlo simulation of how $P(\tilde{N} \in A_M)$ varies with M.

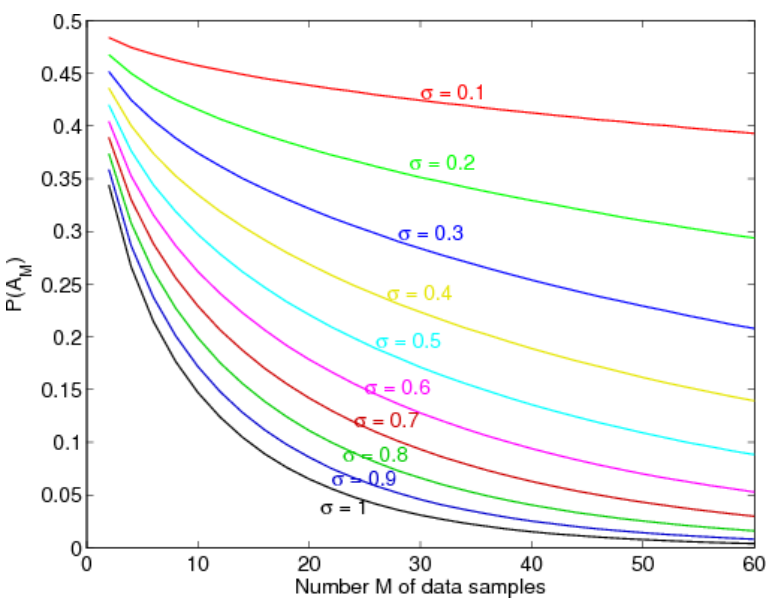

Fig. 3. Probability of satisfying the NEM sufficient condition with different sample sizes M and for different noise standard deviations σ_N. The Gaussian mixture density had mean $\mu = [0, 1]$, standard deviations $\sigma_N = [1, 1]$, and weights $\alpha = [0.5, 0.5]$. The number M of data samples varied from $M = 1$ to $M = 60$. Noise standard deviation varied from $\sigma_N = 0.1$ (top curve) to $\sigma_N = 1.0$ (bottom curve) at 0.1 incremental step. Monte Carlo simulation computed the probability $P(A_M)$ in Eq. (54) from 10^6 samples.

Using non-identically distributed noise N_k avoids the reduction in the probability of satisfying the NEM-condition for large M. The NEM condition still holds when $N_k \in A_k$ for each k even if $N_k \notin A_M = \bigcap_k A_k$. This noise sampling model adapts the kth noise random variable N_k to the kth data sample y_k. This is the general *NEM noise model*. Figures 1 and 2 use the NEM noise model. This model is equivalent to defining the global NEM event $\tilde{A}_M$ as a Cartesian product of sub-events $\tilde{A}_M = \prod_k^M A_k$ instead of the intersection of sub-events $A_M = \bigcap_k A_k$. Thus the bounds of $\tilde{A}_M$ and its coordinate projections no longer depend on sample size M.

Figures 4 and 5 compare the performance of the NEM algorithm with a simulated annealing version of the EM algorithm. This version of EM adds annealed i.i.d. noise to data samples y without screening the noise through the NEM condition. We can thus call it *blind* noise injection. Figure 4 shows that the NEM outperforms blind noise injection at all tested sample sizes M. The average convergence time is about 15% lower for the NEM noise model than for the blind noise

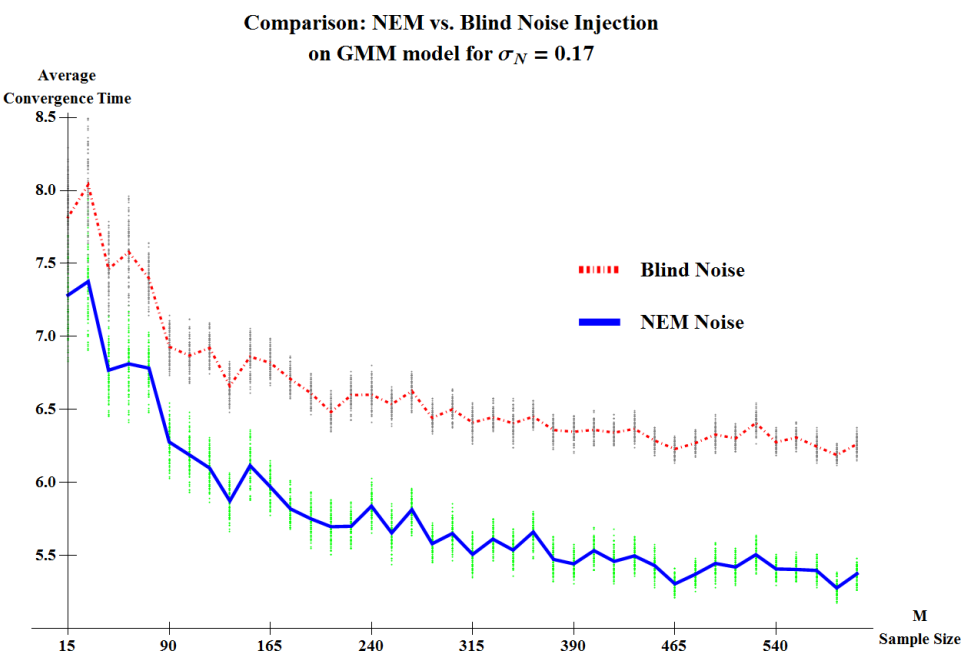

Fig. 4. Plot comparing the effect of the noise sampling model GMM–NEM at different sample sizes M. The NEM noise model used the NEM condition. The blind noise model did not check the NEM condition. So blind noise model had a lower probability of satisfying the NEM condition for all values of M. The plot showed that the NEM noise model outperformed the blind noise model at all sample sizes M. The NEM noise model converged in about 15% fewer steps than the blind noise model for large M. This Gaussian mixture density had sub-population means $\mu = [0, 1]$, standard deviations $\sigma = [1, 1]$, and weights $\alpha = [0.5, 0.5]$. The NEM procedure used the annealed Gaussian noise with initial noise power at $\sigma_N = 0.17$.

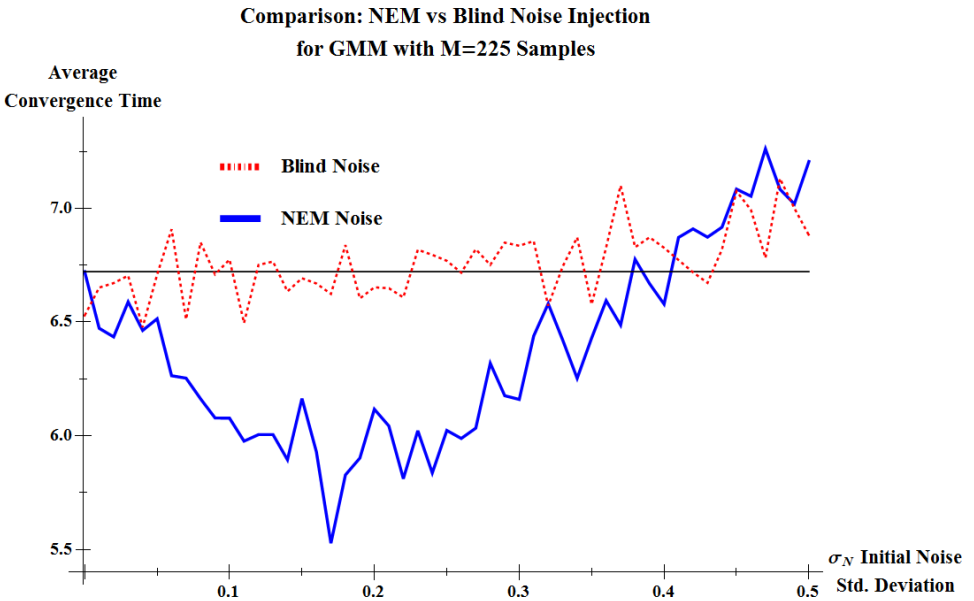

Fig. 5. Comparing the effects of noise injection with and without the NEM sufficient condition. The data model is a GMM with sample size $M = 225$. The blind noise model added annealed noise without checking the NEM condition. The plot shows that NEM noise injection outperformed the blind noise injection. NEM converged up to about 20% faster than the blind noise injection for this model. And blind noise injection produced no reduction in average convergence time. The Gaussian mixture density had mean $\mu = [0, 1]$, standard deviations $\sigma = [1, 1]$, and weights $\alpha = [0.5, 0.5]$ with $M = 225$ samples.

model at large values of M. The two methods are close in performance only at small sample sizes. This is a corollary effect of Theorem 2 from Sec. 5.2. Figure 5 shows that NEM outperforms blind noise injection at a single sample size $M = 225$. But it also shows that blind noise injection may fail to give *any* benefit even when NEM achieves faster average EM convergence for the same set of samples. Thus blind noise injection (or simple simulated annealing) performs worse than NEM and sometimes performs worse than EM itself.

5.2. *Small sample size: Sparsity effect*

The i.i.d. noise model in Theorem 2 has an important corollary effect for sparse data sets. The size of A_M decreases monotonically with M because $A_M = \bigcap_k^M A_k$. Then for $M_0 < M_1$:

$$P(N \in A_{M_0}) \geq P(N \in A_{M_1}) \tag{57}$$

since $M_0 < M_1$ implies that $A_{M_1} \subset A_{M_0}$. Thus arbitrary noise N (i.i.d. *and* independent of Y_k) is more likely to satisfy the NEM condition and produce a noise benefit for smaller samples sizes M_0 than for larger samples sizes M_1. The probability

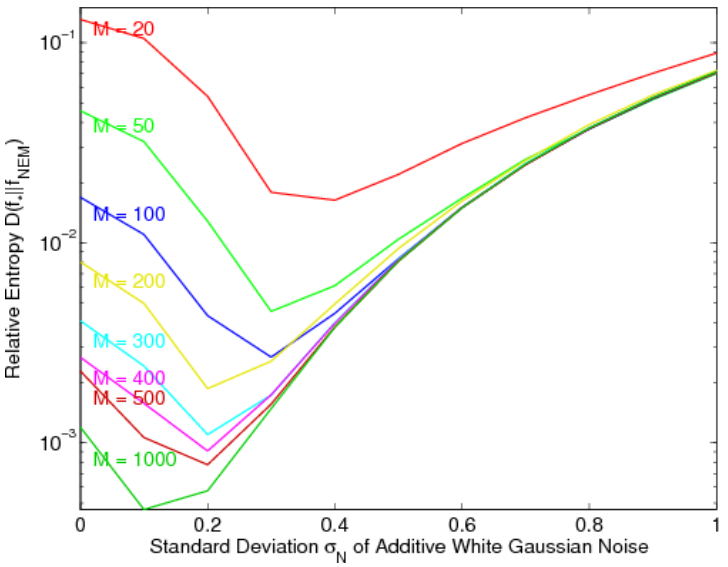

Fig. 6. Noise benefits and sparsity effects in the Gaussian mixture NEM for different sample sizes M. The Gaussian mixture density had sub-populations mean $\mu_1 = 0$ and $\mu_2 = 1$ and standard deviations $\sigma_1 = \sigma_2 = 1$. The number M of data samples varied from $M = 20$ (top curve) to $M = 1000$ (bottom curve). The noise standard deviation varied from $\sigma_N = 0$ (no noise or standard EM) to $\sigma_N = 1$ at 0.1 incremental steps. The plot shows the average relative entropy $D(f^* \| f_{\text{NEM}})$ over 50 trials for each noise standard deviation σ_N. $f_* = f(x \,|\, \theta)$ is the true pdf and $f_{\text{NEM}} = f(x \,|\, \theta_{\text{NEM}})$ is the pdf of NEM-estimated parameters.

that $N \in A_M$ falls to zero as $M \to \infty$. So the strength of the i.i.d. noise benefit falls as $M \to \infty$. Figure 6 shows this sparsity effect. The improvement of relative entropy $D(f * \| f_{\text{NEM}})$ decreases as the number of samples increases: the noise-benefit effect is more pronounced when the data is sparse. Noise appears to act as a type of statistically representative pseudo-data in the sparse case.

5.3. *Asymptotic NEM analysis*

We show last how the NEM noise benefit arises by way of the strong law of large numbers and the central limit theorem. This asymptotic analysis uses the sample mean $\overline{W}_M$:

$$\overline{W}_M = \frac{1}{M} \sum_{m=1}^{M} W_m. \tag{58}$$

The M i.i.d. terms W_m have the logarithmic form

$$W_m = \ln \frac{f(Y_m + N_m, Z_m \,|\, \theta_k)}{f(Y_m, Z_m \,|\, \theta_k)}. \tag{59}$$

The W_m terms are independent because functions of independent random variables are independent. The random sampling framework of the EM algorithm just

means that the underlying random variables are themselves i.i.d. Each W_m term gives a sampling version of the left-hand side of (15) and thus of the condition that the added noise makes the signal value more probable.

We observe first that either the strong or weak law of large numbers [64] applies to the sample mean $\overline{W}_M$. The i.i.d. terms W_m have population mean $\mu_W = \mathbb{E}[W]$ and finite population variance $\sigma_W^2 = V[W]$. Then the strong (weak) law of large numbers states that the sample mean $\overline{W}_M$ converges to the population mean μ_W:

$$\overline{W}_M \to \mu_W \tag{60}$$

with probability one (in probability) [64–66].

The population mean μ_W differs from μ_W^* in general for a given k because θ_k need not equal θ_* until convergence. This difference arises because the expectation μ_W integrates against the pdf $f(y, z, n \,|\, \theta_k)$ while the expectation μ_W^* integrates against the pdf $f(y, z, n \,|\, \theta_*)$. But $\mu_W \to \mu_W^*$ as $\theta_k \to \theta_*$. So the law of large numbers implies that

$$\overline{W}_M \to \mu_W^* \tag{61}$$

with probability one (in probability). So the sample mean converges to the expectation in the positivity condition (27).

The central limit theorem (CLT) applies to the sample mean $\overline{W}_M$ for large sample size M. The CLT states that the standardized sample mean of i.i.d. random variables with finite variance converges in distribution to a standard normal random variable $Z \sim N(0, 1)$ [64]. A noise benefit occurs when the noise makes the signal more probable and thus when $\overline{W}_M > 0$. Then standardizing $\overline{W}_M$ gives the following approximation for large sample size M:

$$P(\overline{W}_M > 0) = P\left(\frac{\overline{W}_M - \mu_W}{\sigma_W/\sqrt{M}} > -\frac{\mu_W}{\sigma_W/\sqrt{M}}\right) \tag{62}$$

$$\approx P\left(Z > -\frac{\sqrt{M}\mu_W}{\sigma_W}\right) \text{ by the CLT} \tag{63}$$

$$= \Phi\left(\frac{\sqrt{M}\mu_W}{\sigma_W}\right) \tag{64}$$

where Φ is the cumulative distribution function of the standard normal random variable Z. So $P(\overline{W}_M > 0) > \frac{1}{2}$ if $\mu_W > 0$ and $P(\overline{W}_M > 0) < \frac{1}{2}$ if $\mu_W < 0$. Suppose the positivity condition (16) holds such that $\mu_W^* > 0$. Then this probability $P(\overline{W}_M > 0)$ goes to one as the sample size M goes to infinity and as θ_k converges to θ_*:

$$\lim_{M\to\infty} P(\overline{W}_M > 0) = 1. \tag{65}$$

The same argument and (64) show that

$$\lim_{M\to\infty} P(\overline{W}_M > 0) = 0 \tag{66}$$

if the positivity condition (16) fails such that $\mu_W^* < 0$.

6. Conclusion

Careful noise injection can speed up the average convergence time of the EM algorithm. The various sufficient conditions for such a noise benefit involve a direct or average effect where the noise makes the signal data more probable. Special cases include mixture density models and log-convex probability density models. Noise injection for the Gaussian and Cauchy mixture models improves the average EM convergence speed when the noise satisfies a simple quadratic condition. Even blind noise injection can sometimes benefit these systems when the data set is sparse. But NEM noise injection still outperforms blind noise injection in all data models tested. An asymptotic argument also shows that the sample-mean version of the EM noise-benefit condition obeys a similar positivity condition. Future work should assess noise benefits in other EM variants and develop methods for finding optimal noise levels for NEM algorithms.

Appendix A. Proof of Theorems

Theorem 1. *Noisy Expectation–Maximization (NEM).*
An EM estimation iteration noise benefit

$$(Q(\theta_* \,|\, \theta_*) - Q(\theta_k \,|\, \theta_*)) \geq (Q(\theta_* \,|\, \theta_*) - Q_N(\theta_k \,|\, \theta_*)) \tag{A.1}$$

occurs on average if

$$\mathbb{E}_{Y,Z,N|\theta_*}\left[\ln\left(\frac{f(Y+N, Z \,|\, \theta_k)}{f(Y, Z \,|\, \theta_k)}\right)\right] \geq 0. \tag{A.2}$$

Proof. We show first that each expectation of Q-function differences in (21) is a distance pseudo-metric. Rewrite Q as an integral:

$$\int_Z \ln[f(y, z \,|\, \theta)] f(z \,|\, y, \theta_k) dz. \tag{A.3}$$

$c_k = D(f(y, z|\theta_*) \| f(y, z \,|\, \theta_k))$ is the expectation over Y because

$$c_k = \iint [\ln(f(y, z \,|\, \theta_*)) - \ln f(y, z \,|\, \theta_k)] f(y, z \,|\, \theta_*) dz\, dy \tag{A.4}$$

$$= \iint [\ln(f(y, z \,|\, \theta_*)) - \ln f(y, z \,|\, \theta_k)] f(z \,|\, y, \theta_*) f(y \,|\, \theta_*) dz\, dy \tag{A.5}$$

$$= \mathbb{E}_{Y|\theta_k}[Q(\theta_* \,|\, \theta_*) - Q(\theta_k \,|\, \theta_*)]. \tag{A.6}$$

$c_k(N)$ is likewise the expectation over Y:

$$c_k(N) = \iint [\ln(f(y, z \,|\, \theta_*)) - \ln f(y + N, z \,|\, \theta_k)] f(y, z \,|\, \theta_*) dz\, dy \tag{A.7}$$

$$= \iint [\ln(f(y, z \,|\, \theta_*)) - \ln f(y + N, z \,|\, \theta_k)] f(z \,|\, y, \theta_*) f(y \,|\, \theta_*) dz\, dy \tag{A.8}$$

$$= \mathbb{E}_{Y|\theta_k}[Q(\theta_* \,|\, \theta_*) - Q_N(\theta_k \,|\, \theta_*)]. \tag{A.9}$$

Take the noise expectation of c_k and $c_k(N)$:

$$\mathbb{E}_N[c_k] = c_k \tag{A.10}$$

$$\mathbb{E}_N[c_k(N)] = \mathbb{E}_N[c_k(N)]. \tag{A.11}$$

So the distance inequality

$$c_k \geq \mathbb{E}_N[c_k(N)] \tag{A.12}$$

guarantees that noise benefits occur on average:

$$\mathbb{E}_{N,Y|\theta_k}[Q(\theta_* \,|\, \theta_*) - Q(\theta_k \,|\, \theta_*)] \geq \mathbb{E}_{N,Y|\theta_k}[Q(\theta_* \,|\, \theta_*) - Q_N(\theta_k \,|\, \theta_*)]. \tag{A.13}$$

We use the inequality condition (A.12) to derive a more useful sufficient condition for a noise benefit. Expand the difference of relative entropy terms $c_k - c_k(N)$:

$$\begin{aligned}
&c_k - c_k(N) \\
&\quad = \iint_{Y,Z} \left(\ln \left[\frac{f(y,z \,|\, \theta_*)}{f(y,z \,|\, \theta_k)} \right] - \ln \left[\frac{f(y,z \,|\, \theta_*)}{f(y+N,z \,|\, \theta_k)} \right] \right) f(y,z \,|\, \theta_*) dy\, dz && \text{(A.14)} \\
&\quad = \iint_{Y,Z} \left(\ln \left[\frac{f(y,z \,|\, \theta_*)}{f(y,z \,|\, \theta_k)} \right] + \ln \left[\frac{f(y+N,z \,|\, \theta_k)}{f(y,z \,|\, \theta_*)} \right] \right) f(y,z \,|\, \theta_*) dy\, dz && \text{(A.15)} \\
&\quad = \iint_{Y,Z} \ln \left[\frac{f(y,z \,|\, \theta_*) f(y+N,z \,|\, \theta_k)}{f(y,z \,|\, \theta_k) f(y,z \,|\, \theta_*)} \right] f(y,z \,|\, \theta_*) dy\, dz && \text{(A.16)} \\
&\quad = \iint_{Y,Z} \ln \left[\frac{f(y+N,z \,|\, \theta_k)}{f(y,z \,|\, \theta_k)} \right] f(y,z \,|\, \theta_*) dy\, dz. && \text{(A.17)}
\end{aligned}$$

Take the expectation with respect to the noise term N:

$$\begin{aligned}
\mathbb{E}_N[c_k - c_k(N)] &= c_k - \mathbb{E}_N[c_k(N)] && \text{(A.18)} \\
&= \int_N \iint_{Y,Z} \ln \left[\frac{f(y+n,z \,|\, \theta_k)}{f(y,z \,|\, \theta_k)} \right] f(y,z \,|\, \theta_*) f(n \,|\, y) dy\, dz\, dn && \text{(A.19)} \\
&= \iint_{Y,Z} \int_N \ln \left[\frac{f(y+n,z \,|\, \theta_k)}{f(y,z \,|\, \theta_k)} \right] f(n \,|\, y) f(y,z \,|\, \theta_*) dn\, dy\, dz && \text{(A.20)} \\
&= \mathbb{E}_{Y,Z,N|\theta_*} \left[\ln \frac{f(Y+N,Z \,|\, \theta_k)}{f(Y,Z \,|\, \theta_k)} \right]. && \text{(A.21)}
\end{aligned}$$

The assumption of finite differential entropy for Y and Z ensures that $\ln f(y,z \,|\, \theta) f(y,z \,|\, \theta_*)$ is integrable. Thus the integrand is integrable. So Fubini's theorem [67] permits the change in the order of integration in (A.21):

$$c_k \geq \mathbb{E}_N[c_k(N)] \quad \text{iff } \mathbb{E}_{Y,Z,N|\theta_*} \left[\ln \left(\frac{f(Y+N,Z \,|\, \theta_k)}{f(Y,Z \,|\, \theta_k)} \right) \right] \geq 0. \tag{A.22}$$

Then an EM noise benefit occurs on average if

$$\mathbb{E}_{Y,Z,N|\theta_*}\left[\ln\left(\frac{f(Y+N,Z\,|\,\theta_k)}{f(Y,Z\,|\,\theta_k)}\right)\right] \geq 0. \tag{A.23}$$

□

Corollary 1. $\mathbb{E}_{Y,Z,N|\theta_*}[\ln \frac{f(Y+N,Z\,|\,\theta)}{f(Y,Z\,|\,\theta)}] \geq 0$ *if*

$$f(y+n,z\,|\,\theta) \geq f(y,z\,|\,\theta) \tag{A.24}$$

for almost all y, z, *and* n.

Proof. The following inequalities need hold only for almost all y, z and n:

$$f(y+n,z\,|\,\theta) \geq f(y,z\,|\,\theta) \tag{A.25}$$

$$\text{iff } \ln[f(y+n,z\,|\,\theta)] \geq \ln[f(y,z\,|\,\theta)] \tag{A.26}$$

$$\text{iff } \ln[f(y+n,z\,|\,\theta)] - \ln[f(y,z\,|\,\theta)] \geq 0 \tag{A.27}$$

$$\text{iff } \ln[\frac{f(y+n,z\,|\,\theta)}{f(y,z\,|\,\theta)}] \geq 0. \tag{A.28}$$

Thus

$$\mathbb{E}_{Y,Z,N|\theta_*}\left[\ln\frac{f(Y+N,Z\,|\,\theta)}{f(Y,Z\,|\,\theta)}\right] \geq 0. \tag{A.29}$$

□

Corollary 2. *Suppose* $Y|_{Z=j} \sim \mathcal{N}(\mu_j, \sigma_j^2)$ *and thus* $f(y\,|\,j,\theta)$ *is a normal pdf. Then*

$$\Delta f_j(y,n) \geq 0 \tag{A.30}$$

holds if

$$n^2 \leq 2n(\mu_j - y). \tag{A.31}$$

Proof. The proof compares the noisy and noiseless normal pdfs. The normal pdf is

$$f(y\,|\,\theta) = \frac{1}{\sigma_j\sqrt{2\pi}}\exp\left[-\frac{(y-\mu_j)^2}{2\sigma_j^2}\right]. \tag{A.32}$$

So $f(y+n\,|\,\theta) \geq f(y\,|\,\theta)$

$$\text{iff } \exp\left[-\frac{(y+n-\mu_j)^2}{2\sigma_j^2}\right] \geq \exp\left[-\frac{(y-\mu_j)^2}{2\sigma_j^2}\right] \tag{A.33}$$

$$\text{iff } -\left(\frac{y+n-\mu_j}{\sigma_j}\right)^2 \geq -\left(\frac{y-\mu_j}{\sigma_j}\right)^2 \tag{A.34}$$

$$\text{iff } -(y-\mu_j+n)^2 \geq -(y-\mu_j)^2. \tag{A.35}$$

Inequality (A.35) holds because σ_j is strictly positive. Expand the left-hand side to get (A.31):

$$(y-\mu_j)^2 + n^2 + 2n(y-\mu_j) \leq (y-\mu_j)^2 \tag{A.36}$$

$$\text{iff } n^2 + 2n(y-\mu_j) \leq 0 \tag{A.37}$$

$$\text{iff } n^2 \leq -2n(y-\mu_j) \tag{A.38}$$

$$\text{iff } n^2 \leq 2n(\mu_j - y). \tag{A.39}$$

□

Corollary 3. *Suppose* $Y|_{Z=j} \sim \mathcal{C}(m_j, d_j)$ *and thus* $f(y \mid j, \theta)$ *is a Cauchy pdf. Then*

$$\Delta f_j(y, n) \geq 0 \tag{A.40}$$

holds if

$$n^2 \leq 2n(m_j - y). \tag{A.41}$$

Proof. The proof compares the noisy and noiseless Cauchy pdfs. The Cauchy pdf is

$$f(y \mid \theta) = \frac{1}{\pi d_j [1 + (\frac{y-m_j}{d_j})^2]}. \tag{A.42}$$

Then $f(y+n \mid \theta) \geq f(y \mid \theta)$

$$\text{iff } \frac{\frac{1}{\pi d_j}}{[1 + (\frac{y+n-m_j}{d_j})^2]} \geq \frac{\frac{1}{\pi d_j}}{[1 + (\frac{y-m_j}{d_j})^2]} \tag{A.43}$$

$$\text{iff } \left[1 + \left(\frac{y-m_j}{d_j}\right)^2\right] \geq \left[1 + \left(\frac{y+n-m_j}{d_j}\right)^2\right] \tag{A.44}$$

$$\text{iff } \left(\frac{y-m_j}{d_j}\right)^2 \geq \left(\frac{y+n-m_j}{d_j}\right)^2. \tag{A.45}$$

Proceed as in the last part of the Gaussian case:

$$\left(\frac{y-m_j}{d_j}\right)^2 \geq \left(\frac{y-m_j+n}{d_j}\right)^2 \tag{A.46}$$

$$\text{iff } (y-m_j)^2 \geq (y-m_j+n)^2 \tag{A.47}$$

$$\text{iff } (y-m_j)^2 \geq (y-m_j)^2 + n^2 + 2n(y-m_j) \tag{A.48}$$

$$\text{iff } 0 \geq n^2 + 2n(y-m_j) \tag{A.49}$$

$$\text{iff } n^2 \leq 2n(m_j - y). \tag{A.50}$$

□

Corollary 4. *Suppose that $f(y, z \mid \theta)$ is log-convex in y and N is independent of Y. Suppose also that $\mathbb{E}_N[N] = 0$. Then*

$$\mathbb{E}_{Y,Z,N|\theta_*}\left[\ln \frac{f(Y+N, Z \mid \theta_k)}{f(Y, Z \mid \theta_k)}\right] \geq 0. \tag{A.51}$$

Proof. (The same argument applies if we use $f(x \mid \theta)$ instead of $f(y, z \mid \theta)$ and if $f(x \mid \theta)$ is log-convex in x.)
$f(y, z \mid \theta)$ is log-convex in y and $\mathbb{E}_N[y + N] = y$. So

$$\mathbb{E}_N[\ln f(y+N, z \mid \theta_k)] \geq \ln f(\mathbb{E}_N[y+N], z \mid \theta_k). \tag{A.52}$$

The right-hand side becomes

$$\ln f(\mathbb{E}_N[y+N], z \mid \theta_k) = \ln f(y + \mathbb{E}_N[N], z \mid \theta_k) \tag{A.53}$$

$$= \ln f(y, z \mid \theta_k) \tag{A.54}$$

because $\mathbb{E}[N] = 0$. So

$$\mathbb{E}_N[\ln f(y+N, z \mid \theta_k)] \geq \ln f(y, z \mid \theta_k) \tag{A.55}$$

$$\text{iff } (\mathbb{E}_N[\ln f(y+N, z \mid \theta_k)] - \ln f(y, z \mid \theta_k)) \geq 0 \tag{A.56}$$

$$\text{iff } (\mathbb{E}_N[\ln f(y+N, z \mid \theta_k) - \ln f(y, z \mid \theta_k))] \geq 0 \tag{A.57}$$

$$\text{iff } \mathbb{E}_{Y,Z|\theta_*}[\mathbb{E}_N[\ln f(Y+N, Z \mid \theta_k) - \ln f(Y, Z \mid \theta_k)]] \geq 0 \tag{A.58}$$

$$\text{iff } \mathbb{E}_{Y,Z,N|\theta_*}\left[\ln \frac{f(Y+N, Z \mid \theta_k)}{f(Y, Z \mid \theta_k)}\right] \geq 0. \tag{A.59}$$

Inequality (A.59) follows because N is independent of θ_*. □

Theorem 2. *Large Sample GMM– and CMM–NEM.*
The set A_M of i.i.d. noise values that satisfy the Gaussian (Cauchy) NEM condition for all data samples y_k decreases with probability one to the set $\{0\}$ as $M \to \infty$:

$$P\left(\lim_{M\to\infty} A_M = \{0\}\right) = 1. \tag{A.60}$$

Proof. Define the NEM-condition event A_k for a single sample y_k as

$$A_k = \{N^2 \leq 2N(\mu_j - y_k) \mid \forall j\}. \tag{A.61}$$

$N^2 \leq 2N(\mu_j - y_k)$ for all j if N satisfies the NEM condition ($N \in A_k$). So

$$N^2 - 2N(\mu_j - y_k) \leq 0 \quad \text{for all } j \tag{A.62}$$

$$\text{and} \quad N(N - 2(\mu_j - y_k)) \leq 0 \quad \text{for all } j. \tag{A.63}$$

This quadratic inequality's solution set (a_j, b_j) for j is

$$I_j = [a_j, b_j] = \begin{cases} [0, 2(\mu_j - y_k)] & \text{if } y_k < \mu_j \\ [2(\mu_j - y_k), 0] & \text{if } y_k > \mu_j \\ \{0\} & \text{if } y_k \in [\min \mu_j, \max \mu_j] \end{cases}. \tag{A.64}$$

Define b_k^+ and b_k^- as $b_k^+ = 2\min_j(\mu_j - y_k)$ and $b_k^- = 2\max_j(\mu_j - y_k)$. Then the maximal solution set $A_k = [a, b]$ over all j is

$$A_k = \bigcap_j^J I_j = \begin{cases} [0, b_k^+] & \text{if } y_k < \mu_j \ \forall j \\ [b_k^-, 0] & \text{if } y_k > \mu_j \ \forall j \\ \{0\} & \text{if } y_k \in [\min \mu_j, \max \mu_j] \end{cases} \tag{A.65}$$

where J is the number of sub-populations in the mixture density. There is a sorting such that the I_j are nested for each sub-case in (A.65). So the nested interval theorem [68] (or Cantor's intersection theorem [69]) implies that A_k is not empty because it is the intersection of nested bounded closed intervals.

$A_k = \{0\}$ holds if the NEM condition fails for that value of y_k This happens when some I_j sets are positive and other I_j sets are negative. The positive and negative I_j sets intersect only at zero. No non-zero value of N will produce a positive average noise benefit. The additive noise N must be zero.

Write A_M as the intersection of the A_k sub-events:

$$A_M = \{N^2 \leq 2N(\mu_j - y_k) \,|\, \forall j \text{ and } \forall k\} \tag{A.66}$$

$$= \bigcap_k^M A_k \tag{A.67}$$

$$= \begin{cases} [0, \min_k b_k^+] & \text{if } y_k < \mu_j \ \forall j, k \\ [\max_k b_k^-, 0] & \text{if } y_k > \mu_j \ \forall j, k \\ \{0\} & \text{if } \exists\, k : y_k \in [\min \mu_j, \max \mu_j] \end{cases} . \tag{A.68}$$

Thus a second application of the nested interval property implies that A_M is not empty.

We now characterize the asymptotic behavior of the set A_M. A_M depends on the locations of the samples y_k relative to the sub-population means μ_j. Then $A_M = \{0\}$ if there exists some k_0 such that $\min \mu_j \leq y_{k_0} \leq \max \mu_j$. Define the set $S = [\min \mu_j, \max \mu_j]$. Then by the Lemma below $\lim_{M\to\infty} \#_M(Y_k \in S) > 0$ holds with probability one. So there exists with probability one a $k_0 \in \{1 \dots M\}$ such that $y_{k_0} \in S$ as $M \to \infty$. Then $A_{k_0} = \{0\}$ by equation (A.68). Then with probability one:

$$\lim_{M\to\infty} A_M = A_{k_0} \cap \lim_{M\to\infty} \bigcap_{k\neq k_0}^M A_k \tag{A.69}$$

$$= \{0\} \cap \lim_{M\to\infty} \bigcap_{k\neq k_0}^M A_k. \tag{A.70}$$

So

$$\lim_{M\to\infty} A_M = \{0\} \quad \text{with probability one} \tag{A.71}$$

since A_M is not empty by the nested intervals property and since $0 \in A_k$ for all k.

□

Lemma 1. *Suppose that $S \subset \mathbb{R}$ is Borel-measurable and that $\mathbb{R}$ is the support of the pdf of the random variable Y. Let M be the number of random samples of Y. Then as $M \to \infty$:*

$$\frac{\#_M(Y_k \in S)}{M} \to P(Y \in S) \quad \textit{with probability one} \tag{A.72}$$

where $\#_M(Y_k \in S)$ is of the number of random samples $y_1, \ldots, y_M$ of Y that fall in S.

Proof. Define the indicator function random variable $\mathbb{I}_S(Y)$ as

$$\mathbb{I}_S(Y) = \begin{cases} 1 & Y \in S \\ 0 & Y \notin S \end{cases}. \tag{A.73}$$

The strong law of large numbers implies that the sample mean $\overline{\mathbb{I}}_S$

$$\overline{\mathbb{I}}_S = \frac{\sum_k^M \mathbb{I}_S(Y_k)}{M} = \frac{\#_M(Y_k \in S)}{M} \tag{A.74}$$

converges to $\mathbb{E}[\mathbb{I}_S]$ with probability one. Here $\#_M(Y_k \in S)$ is the number of random samples $Y_1, \ldots, Y_M$ that fall in the set S. But $\mathbb{E}[\mathbb{I}_S] = P(Y \in S)$. So with probability one:

$$\frac{\#_M(Y_k \in S)}{M} \to P(Y \in S) \tag{A.75}$$

as claimed.

Then $P(S) > 0$ implies that

$$\lim_{M \to \infty} \frac{\#_M(Y_k \in S)}{M} > 0 \tag{A.76}$$

and $\lim_{M \to \infty} \#_M(Y_k \in S) > 0$ with probability one since $M > 0$. □

References

[1] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. R. Stat. Soc. Ser. B* **39** (1977) 1–38.

[2] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions* (John Wiley and Sons, 2007).

[3] M. R. Gupta and Y. Chen, Theory and use of the EM algorithm, *Foundations Trends Signal Process.* **4** (2010) 223–296.

[4] G. Celeux and G. Govaert, A classification EM algorithm for clustering and two stochastic versions, *Comput. Stat. Data Anal.* **14** (1992) 315–332.

[5] C. Ambroise, M. Dang and G. Govaert, Clustering of spatial data by the EM algorithm, *Quant. Geol. Geostat.* **9** (1997) 493–504.

[6] L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proc. IEEE* **77** (1989) 257–286.

[7] B. H. Juang and L. R. Rabiner, Hidden Markov models for speech recognition, *Technometrics* **33** (1991) 251–272.

[8] L. A. Shepp and Y. Vardi, Maximum likelihood reconstruction for emission tomography, *IEEE Trans. Med. Imag.* **1** (1982) 113–122.

[9] Y. Zhang, M. Brady and S. Smith, Segmentation of brain MR Images through a hidden Markov random field model and the expectation–maximization algorithm, *IEEE Trans. Med. Imag.* **20** (2001) 45–57.
[10] C. E. Lawrence and A. A. Reilly, An expectation–maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences, *Proteins*: *Struct. Function Bioinf.* **7** (1990) 41–51.
[11] T. L. Bailey and C. Elkan, Unsupervised learning of multiple motifs in biopolymers using expectation–maximization, *Machine Learning* **21** (1995) 51–80.
[12] J. Wang, A. Dogandzic and A. Nehorai, Maximum likelihood estimation of compound-gaussian clutter and target parameters, *IEEE Trans. Signal Process.* **54** (2006) 3884–3898.
[13] M. Reilly and E. Lawlor, A likelihood-based method of identifying contaminated lots of blood product, *Int. J. Epidemiol.* **28** (1999) 787–792.
[14] P. Bacchetti, Estimating the incubation period of AIDS by comparing population infection and diagnosis patterns, *J. Am. Stat. Assoc.* **85** (1990) 1002–1008.
[15] N. A. Gershenfeld, *The Nature of Mathematical Modeling* (Cambridge University Press, 1999), p. 177.
[16] M. A. Tanner, *Tools for Statistical Inference*: *Methods for the Exploration of Posterior Distributions and Likelihood Functions*, Springer Series in Statistics (Springer, 1996).
[17] A. R. Bulsara and L. Gammaitoni, Tuning in to noise, *Phys. Today* (1996) 39–45.
[18] L. Gammaitoni, P. Hänggi, P. Jung and F. Marchesoni, Stochastic resonance, *Rev. Mod. Phys.* **70** (1998) 223–287.
[19] B. Kosko, *Noise* (Viking, 2006).
[20] J. J. Brey and A. Prados, Stochastic resonance in a one-dimension ising model, *Phys. Lett. A* **216** (1996) 240–246.
[21] H. A. Kramers, Brownian motion in a field of force and the diffusion model of chemical reactions, *Physica* **VII** (1940) 284–304.
[22] A. Förster, M. Merget and F. W. Schneider, Stochastic resonance in chemistry 2: The peroxidase-oxidase reaction, *J. Phys. Chem.* **100** (1996) 4442–4447.
[23] F. Moss, A. Bulsara and M. Shlesinger (eds.), *Journal of Statistical Physics*, *Special Issue on Stochastic Resonance in Physics and Biology* (*Proceedings of the NATO Advanced Research Workshop*), Vol. 70, No. 1/2 (Plenum Press, 1993).
[24] P. Cordo, J. T. Inglis, S. Vershueren, J. J. Collins, D. M. Merfeld, S. Rosenblum, S. Buckley and F. Moss, Noise in human muscle spindles, *Nature* **383** (1996) 769–770.
[25] R. K. Adair, R. D. Astumian and J. C. Weaver, Detection of weak electric fields by sharks, rays and skates, *Chaos*: *Focus Issue Constructive Role Noise Fluctuation Driven Transport Stochastic Resonance* **8** (1998) 576–587.
[26] P. Hänggi, Stochastic resonance in biology, *Chem. Phys. Chem.* **3** (2002) 285–290.
[27] A. R. Bulsara and A. Zador, Threshold detection of wideband signals: A noise-induced maximum in the mutual information, *Phys. Rev. E* **54** (1996) R2185–R2188.
[28] F. Chapeau-Blondeau and D. Rousseau, Noise-enhanced performance for an optimal bayesian estimator, *IEEE Trans. Signal Process.* **52** (2004) 1327–1334.
[29] M. McDonnell, N. Stocks, C. Pearce and D. Abbott, *Stochastic Resonance*: *From Suprathreshold Stochastic Resonance to Stochastic Signal Quantization* (Cambridge University Press, 2008).
[30] H. Chen, P. Varshney, S. Kay and J. Michels, Noise enhanced nonparametric detection, *IEEE Trans. Inf. Theory* **55** (2009) 499–506.
[31] A. Patel and B. Kosko, Noise benefits in quantizer-array correlation detection and watermark decoding, *IEEE Trans. Signal Process.* **59** (2011) 488–505.

[32] B. Franzke and B. Kosko, Noise can speed convergence in Markov chains, *Phys. Rev. E* **84** (2011) 041112.
[33] G. Celeux and J. Diebolt, The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem, *Comput. Stat. Q.* **2** (1985) 73–82.
[34] X. L. Meng and D. B. Rubin, Maximum likelihood estimation via the ECM algorithm: A general framework, *Biometrika* **80** (1993) 267.
[35] C. Liu and D. B. Rubin, The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence, *Biometrika* **81** (1994) 633.
[36] J. A. Fessler and A. O. Hero, Space-alternating generalized expectation–maximization algorithm, *IEEE Trans. Signal Process.* **42** (1994) 2664–2677.
[37] H. M. Hudson and R. S. Larkin, Accelerated image reconstruction using ordered subsets of projection data, *IEEE Trans. Med. Imag.* **13** (1994) 601–609.
[38] C. F. J. Wu, On the convergence properties of the EM algorithm, *Ann. Stat.* **11** (1983) 95–103.
[39] R. A. Boyles, On the convergence of the EM algorithm, *J. R. Stat. Soc. Ser. B* (*Methodological*) **45** (1983) 47–50.
[40] R. A. Redner and H. F. Walker, Mixture densities, maximum likelihood and the EM algorithm, *SIAM Rev.* **26** (1984) 195–239.
[41] L. Xu and M. I. Jordan, On convergence properties of the EM algorithm for Gaussian mixtures, *Neural Comput.* **8** (1996) 129–151.
[42] R. Sundberg, Maximum likelihood theory for incomplete data from an exponential family, *Scandinavian J. Stat.* **1** (1974) 49–58.
[43] D. Chauveau, A stochastic EM algorithm for mixtures with censored data, *J. Stat. Plan. Inf.* **46** (1995) 1–25.
[44] R. J. Hathaway, Another interpretation of the EM algorithm for mixture distributions, *Statist. Probab. Lett.* **4** (1986) 53–56.
[45] J. P. Delmas, An equivalence of the EM and ICE algorithm for exponential family, *IEEE Trans. Signal Process.* **45** (1997) 2613–2615.
[46] M. Á. Carreira-Perpiñán, Gaussian mean shift is an EM algorithm, *IEEE Trans. Pattern Anal. Machine Intel.* **29** (2005) 2007.
[47] X. L. Meng and D. B. Rubin, Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm, *J. Am. Stat. Assoc.* **86** (1991) 899–909.
[48] G. Celeux, S. Chrétien, F. Forbes and A. Mkhadri, A component-wise EM algorithm for mixtures, *J. Comput. Graph. Stat.* **10** (2001) 697–712.
[49] G. J. McLachlan and D. Peel, *Finite Mixture Models*, Wiley series in probability and statistics: Applied probability and statistics (Wiley, 2004).
[50] M. T. Tan, G. Tian and K. W. Ng, *Bayesian Missing Data Problems*: *EM*, *Data Augmentation and Noniterative Computation* (CRC Press, 2010).
[51] F. Proschan, Theoretical explanation of observed decreasing failure rate, *Technometrics* **5** (1963) 375–383.
[52] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley & Sons, New York, 1991).
[53] O. Osoba, S. Mitaim and B. Kosko, Noise benefits in the expectation–maximization algorithm: NEM theorems and models, in *The Int. Joint Conf. Neural Networks* (*IJCNN*) (IEEE, 2011), pp. 3178–3183.
[54] V. Hasselblad, Estimation of parameters for a mixture of normal distributions, *Technometrics* **8** (1966) 431–444.
[55] O. Osoba and B. Kosko, Noise-enhanced clustering and competitive learning algorithms, *Neural Netw.* **37** (2013) 132–140.

[56] O. Osoba and B. Kosko, Erratum to "Noise enhanced clustering and competitive learning algorithms" [Neural Netw. **37** (2013) 132–140], *Neural Netw.* (2013).
[57] R. C. Dahiya and J. Gurland, Goodness of fit tests for the gamma and exponential distributions, *Technometrics* **14** (1972) 791–801.
[58] M. Bagnoli and T. Bergstrom, Log-concave probability and its applications, *Econ. Theory* **26** (2005) 445–469.
[59] S. Kirkpatrick, C. Gelatt Jr. and M. Vecchi, Optimization by simulated annealing, *Science* **220** (1983) 671–680.
[60] V. Černỳ, Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm, *J. Opt. Theory Appl.* **45** (1985) 41–51.
[61] S. Geman and C. Hwang, Diffusions for global optimization, *SIAM J. Control Opt.* **24** (1986) 1031–1043.
[62] B. Hajek, Cooling schedules for optimal annealing, *Math. Operations Res.* **13** (1988) 311–329.
[63] B. Kosko, *Neural Networks and Fuzzy Systems*: *A Dynamical Systems Approach to Machine Intelligence* (Prentice Hall, 1991).
[64] P. Billingsley, *Probability and Measure*, 3rd edn. (John Wiley & Sons, 1995).
[65] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. II (John Wiley & Sons, 1966).
[66] R. Durrett, *Probability*: *Theory and Examples*, 4th edn. (Cambridge University Press, 2010).
[67] G. B. Folland, *Real Analysis*: *Modern Techniques and Their Applications*, 2nd edn. (Wiley-Interscience, 1999).
[68] M. Spivak, *Calculus* (Cambridge University Press, 2006).
[69] W. Rudin, *Principles of Mathematical Analysis*, 3rd edn. (McGraw-Hill, New York, 1976).
[70] S. Mitaim and B. Kosko, Adaptive stochastic resonance, *Proceedings of the IEEE* **86**(11) (1998) 2152–2183.
[71] B. Kosko and S. Mitaim, Stochastic resonance in noisy threshold neurons, *Neural Networks* **16**(5) (2003) 755–761.