Fluctuation and Noise Letters
Vol. 15, No. 1 (2016) 1650007 (23 pages)

DOI: 10.1142/S0219477516500073

The Noisy Expectation-Maximization Algorithm for Multiplicative Noise Injection

Osonde Osoba[*,†,‡] and Bart Kosko[†,§]

[*] *RAND Corporation*
Santa Monica, CA 90401-3208, USA

[†] *Signal and Image Processing Institute*
Electrical Engineering Department
University of Southern California
Los Angeles, CA 90089-2564, USA

[‡] *oosoba@rand.org*
[§] *kosko@usc.edu*

Received 17 May 2015
Accepted 2 December 2015
Published 17 March 2016

Communicated by Igor Khovanov

We generalize the noisy expectation-maximization (NEM) algorithm to allow arbitrary modes of noise injection besides just adding noise to the data. The noise must still satisfy a NEM positivity condition. This generalization includes the important special case of multiplicative noise injection. A generalized NEM theorem shows that all measurable modes of injecting noise will speed the average convergence of the EM algorithm if the noise satisfies a generalized NEM positivity condition. This noise-benefit condition has a simple quadratic form for Gaussian and Cauchy mixture models in the case of multiplicative noise injection. Simulations show a multiplicative-noise EM speed-up of more than 27% in a simple Gaussian mixture model. Injecting blind noise only slowed convergence. A related theorem gives a sufficient condition for an average EM noise benefit for arbitrary modes of noise injection if the data model comes from the general exponential family of probability density functions. A final theorem shows that injected noise slows EM convergence on average if the NEM inequalities reverse and the noise satisfies a negativity condition.

Keywords: Expectation maximization algorithm; noise benefit; stochastic resonance; maximum likelihood estimation.

1. Noise Boosting the Expectation-Maximization Algorithm

We show how carefully chosen and injected multiplicative noise can speed convergence of the popular expectation-maximization (EM) algorithm. Multiplicative noise [1] occurs in many applications in signal processing and communications. These include synthetic aperture radar [2–5], sonar imaging [6, 7], photonics [8], and random amplitude modulation [9]. A more general theorem allows *arbitrary*

modes of combining signal and noise. The result still speeds EM convergence on average at each iteration so long as the injected noise satisfies a positivity condition.

The EM algorithm generalizes maximum-likelihood estimation to the case of missing or corrupted data [10, 11]. Maximum likelihood maximizes the conditional signal probability density function (pdf) $f(y|\theta)$ for a random signal variable Y given a vector of parameters θ. It equally maximizes the log-likelihood $\ln f(y|\theta)$ since the logarithm is monotone increasing. So the maximum-likelihood estimate θ_* is

$$\theta_* = \underset{\theta}{\operatorname{argmax}} \ \ln f(y|\theta). \tag{1}$$

The parameter vector θ can contain means or covariances or mixture weights or any other terms that parametrize the pdf $f(y|\theta)$. The data itself consists of observations or realizations y of the signal random variable Y. The data can be speech samples or image vectors or any type of measured numerical quantity. The EM framework allows for missing or hidden data or so-called latent variables. The random variable Z denotes all such latent variables. These latent variables can describe unseen states in a hidden Markov model or hidden neurons in a multilayer neural network. Then Z appears in the log-likelihood $\ln f(y|\theta)$ through the pdf identity $f(y|\theta) = \frac{f(y,z|\theta)}{f(z|y,\theta)}$. This gives the key EM log-likelihood equality $\ln f(y|\theta) = \ln f(y,z|\theta) - \ln f(z|y,\theta)$.

The EM algorithm estimates the missing information in Z by iteratively maximizing the probability of Z given both the observed data y and the current parameter estimate θ_k [12]. This involves averaging the log-likelihood $\ln f(y,z|\theta_k)$ over the conditional pdf $f(z|y,\theta_k)$ to form the surrogate likelihood function $Q(\theta|\theta_k)$:

$$Q(\theta|\theta_k) = \mathbb{E}_Z[\ln f(y,Z|\theta)\big|Y = y,\theta_k] \tag{2}$$

$$= \int_{\mathcal{Z}} \ln[f(y,z|\theta)] f(z|y,\theta_k)\ \mathrm{d}z. \tag{3}$$

Then EM's "ascent property" [10] uses Jensen's inequality [13] and the above EM log-likelihood equality to ensure that any θ that increases the surrogate likelihood function $Q(\theta|\theta_k)$ can only increase the log-likelihood difference $\ln f(y|\theta) - \ln f(y|\theta_k)$ or leave it unchanged: $\ln \frac{f(y|\theta)}{f(y|\theta_k)} \geq Q(\theta|\theta_k) - Q(\theta_k|\theta_k)$. The result is that EM is a hill–climbing algorithm that can never decrease the log-likelihood $\ln f(y|\theta)$ at any step. The algorithm can at most increase the log-likelihood.

The EM algorithm iteratively climbs the closest hill of probability or log-likelihood until it reaches the peak of maximum likelihood. The peak or mode corresponds to the locally maximal parameter θ_*. So the EM algorithm converges to the local likelihood maximum θ_*: $\theta_k \to \theta_*$. The EM algorithm halts in practice when its successive estimates θ_k differ by less than a given tolerance level: $\|\theta_k - \theta_{k-1}\| < 10^{-\mathrm{tol}}$ or when $|\ln f(y|\theta_k) - \ln f(y|\theta_{k-1})| < \varepsilon$ for some small positive ε.

The EM algorithm generalizes many popular algorithms. These include the k-means clustering algorithm [14] used in pattern recognition and big-data analysis, the backpropagation algorithm used to train deep feedforward and convolutional neural networks [15–17], and the Baum–Welch algorithm used to train hidden

Markov models [18, 19]. But the EM algorithm can converge slowly if the amount of missing data is high or if the number of estimated parameters is large [11, 20]. It can also get stuck at local probability maxima. Users can run the EM algorithm from several starting points to mitigate the problem of convergence to local maxima.

The Noisy EM (NEM) algorithm [14, 21–23] is a noise-enhanced version of the EM algorithm that carefully selects noise and then *adds* it to the data. NEM converges faster on average than EM does because on average it takes larger steps up the same hill of probability or of log-likelihood. NEM never takes shorter steps on average. The largest noise gains tend to occur in the first few steps. So NEM enhances the ascent property at each iteration. This is a type of nonlinear noise benefit or *stochastic resonance* [24–35] that does not depend on a threshold [36, 37].

NEM adds noise N to the data Y if the noise satisfies the NEM average positivity (nonnegativity) condition:

$$\mathbb{E}_{Y,Z,N|\theta_*}\left[\ln\left(\frac{f(Y+N,Z|\theta_k)}{f(Y,Z|\theta_k)}\right)\right] \geq 0. \tag{4}$$

The NEM positivity condition (4) holds when the noise-perturbed complete likelihood $f(y+N,z|\theta_k)$ is larger on average than the noiseless likelihood $f(y,z|\theta_k)$ at the kth step of the algorithm [21, 23]. This noise-benefit condition has a simple quadratic form when the data or signal model is a mixture of Gaussian pdfs.

The NEM positivity inequality (4) is not vacuous. This holds because the expectation conditions on the converged parameter vector θ_*. Consider instead what happens in the generic case of averaging a log-likelihood ratio [12]. Take the expectation of the log-likelihood ratio $\ln\frac{f(y|\theta)}{g(y|\theta)}$ with respect to the pdf $g(y|\theta)$. This gives the expectation $\mathbb{E}_g[\ln\frac{f(y|\theta)}{g(y|\theta)}]$. But the logarithm is concave. So Jensen's inequality gives $\mathbb{E}_g[\ln\frac{f(y|\theta)}{g(y|\theta)}] \leq \ln\mathbb{E}_g[\frac{f(y|\theta)}{g(y|\theta)}]$. Then the pdf $g(y|\theta)$ cancels out of the latter expectation: $\ln\mathbb{E}_g[\frac{f(y|\theta)}{g(y|\theta)}] = \ln\int_Y \frac{f(y|\theta)}{g(y|\theta)}g(y|\theta)dy = \ln\int_Y f(y|\theta)dy = \ln 1 = 0$ since the pdf $f(y|\theta)$ integrates to one over the whole sample space. So $\mathbb{E}_g[\ln\frac{f(y|\theta)}{g(y|\theta)}] \leq 0$. Then a strict positivity condition is impossible. The expectation in (4) does not lead to such a pdf cancellation in general because the integrating density depends on θ_* rather than on θ_k. Cancellation occurs only when the NEM algorithm has converged because then $\theta_k = \theta_*$.

A simple argument gives the intuition behind the NEM positivity condition (4) for additive noise. This argument holds in much greater generality and underlies much of the theory of noise-boosting both the EM algorithm and Markov chain Monte Carlo algorithms [37]. Consider a noise sample or realization n that makes a signal y more probable: $f(y+n|\theta) \geq f(y|\theta)$ for some parameter θ. The value y is a realization of the signal random variable Y. The value n is a realization of the noise random variable N. Then this pdf inequality holds if and only if $\ln\frac{f(y+n|\theta)}{f(y|\theta)} \geq 0$. Averaging over the signal and noise random variables gives the basic expectation form of the NEM positivity condition. Averaging implies that the pdf inequality need hold only almost everywhere. It need not hold on sets of zero probability.

This allows the user to ignore particular values when using continuous probability models.

Particular choices of the signal conditional probability $f(y|\theta)$ can greatly simplify the NEM sufficient condition. This signal probability is the so-called "data model" in the EM context of maximum likelihood estimation. We show below that Gaussian and Cauchy choices lead to simple quadratic NEM conditions when injecting multiplicative noise. An exponential data model leads to an even simpler linear NEM condition.

The same argument for multiplicative noise suggests that a similar positivity condition should hold for a noise benefit. This will hold given the corresponding pdf inequality $f(yn|\theta) \geq f(y|\theta)$. This inequality is equivalent to $\ln \frac{f(yn|\theta)}{f(y|\theta)} \geq 0$. Then averaging again gives a NEM positivity condition. There is nothing unique about the operations of addition or multiplication in this signal-noise context. So a noise benefit should hold for *any* measurable function $\phi(y, n)$ that combines the signal y and noise n if $f(\phi(y, n)|\theta) \geq f(y|\theta)$. The four theorems below show that this is the case.

Theorem 1 generalizes the NEM Theorem from additive noise injection $Y + N$ to arbitrary measurable noise injection $\phi(Y, N)$. Theorem 2 states the NEM sufficient condition for the special case where the noise-injection mode is multiplicative: $\phi(Y, N) = YN$. We call this new condition the m-NEM condition or the multiplicative-NEM condition. Corollary 1 shows that a mixture of such pdfs satisfies the general NEM property if all the mixed pdfs do. Corollary 2 derives the specific form for Gaussian and Cauchy signal pdfs. Theorem 3 states a sufficient NEM condition for arbitrary noise injection when the signal model comes from the important class of exponential family pdfs. Theorem 4 states the dual NEM *negativity* condition for a noise *harm* or slow-down in EM convergence.

Figure 1 shows an EM speed-up of 27.6% for m-NEM noise injection in the generic case of a bimodal mixture of two Gaussian pdfs. Sampling from the mixture corresponds to sampling from two subpopulations that have the same variance but different means. This structure can arise when sampling from a population that consists of two unknown bipolar signals. The task is threefold: Estimate the unknown means of the two mixed Gaussian densities. Estimate the unknown variances of the mixed densities. And estimate the unknown mixture weights. The mixture weights are non-negative and sum to unity.

The noise-injected EM algorithm estimated all these parameters of the equally weighted two-pdf Gaussian mixture model. Suppose random variable Y_j is Gaussian or normal with mean μ_j and variance σ_j^2: $Y_j \sim N(\mu_j, \sigma_j^2)$ with pdf $f_j(y|\mu_j, \sigma_j^2)$. Then the two-mixture density in Fig. 1 had the form $f(y) = \alpha f_1(y|\mu_1, \sigma_1^2) + (1 - \alpha) f_2(y|\mu_2, \sigma_2^2) = \frac{1}{2} f_1(y| - 2, 4) + \frac{1}{2} f_2(y|2, 4)$. The data itself came from Mathematica's routine for randomly sampling from a Gaussian mixture. The noise-boosted EM algorithm took on average only seven iterations to estimate the Gaussian mixture parameters $\alpha, \mu_1, \mu_2, \sigma_1^2$, and σ_2^2 while the noiseless EM algorithm took on average 10 steps. The optimal initial noise standard deviation was $\sigma_N^* = 0.44$. The simulations "cooled" or "annealed" the noise by multiplying the starting noise

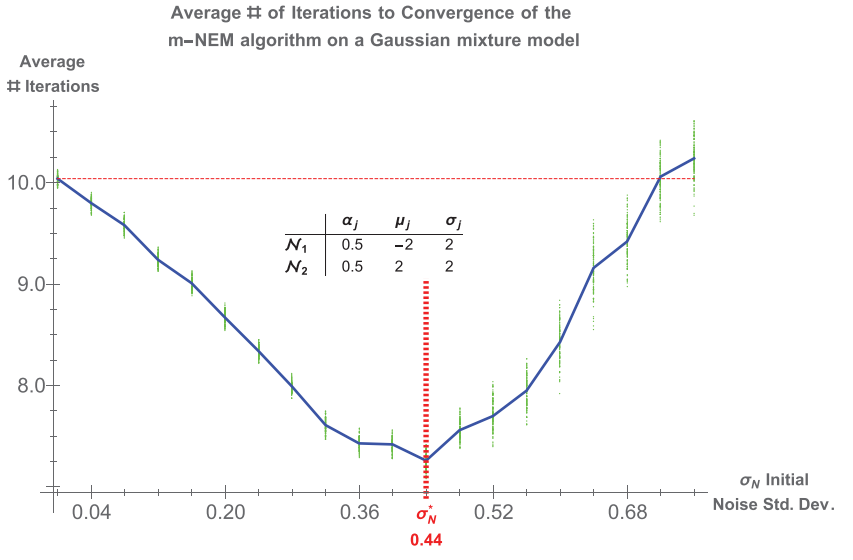

Fig. 1. Multiplicative noise benefit when estimating the parameters of a sampled Gaussian mixture model. The mixture density f equally weighted two Gaussian probability density functions with the same variance of 4: $f(y) = \frac{1}{2}N_1(-2, 4) + \frac{1}{2}N_2(2, 4)$. The EM algorithm estimated the mixing weights, the means, and the variances of the two Gaussian densities. Low intensity starting noise decreased the EM convergence time while higher intensity starting noise increased it. The multiplicative noise had unit mean with different but decaying standard deviations. The optimal initial noise standard deviation was $\sigma^* = 0.44$. It gave a 27.6% speed-up over the noiseless EM algorithm. Optimal m-NEM needed only seven iterations on average to converge to the correct mixture parameters while noiseless EM needed 10 iterations on average. The m-NEM procedure injected multiplicative noise that decayed at an inverse-square rate with the iterations.

standard deviation σ_N with the inverse-square term k^{-2} at each iteration k. This gradually shut off the noise injection as we discuss below when we present the details of the n-NEM algorithm. The far right of Fig. 1 shows a type of swamping effect where too much injected noise begins to hurt performance compared with noiseless EM. This appears to be an artifact of injecting such noise into EM's fixed-point structure.

Figure 2 shows that ordinary or *blind* multiplicative noise (not subject to the m-NEM condition) only slowed EM convergence for the same Gaussian-mixture problem as in Fig. 1. Blind noise was just noise drawn at random or uniformly from the set of all possible noise. It was not subject to the m-NEM condition or to any other condition.

The optimal speed-up using additive noise on the same data model was 30.5% at an optimal noise power of $\sigma^* = 1.9$. This speed-up was slightly better than the m-NEM speed-up for the same mixture model of two Gaussian pdfs.

A statistical test for the difference in the averaged optimal convergence times found that this difference was not statistically significant at the standard 0.05 significance level. Nor was it significant at the 0.10 or 0.01 levels. The hypothesis test for

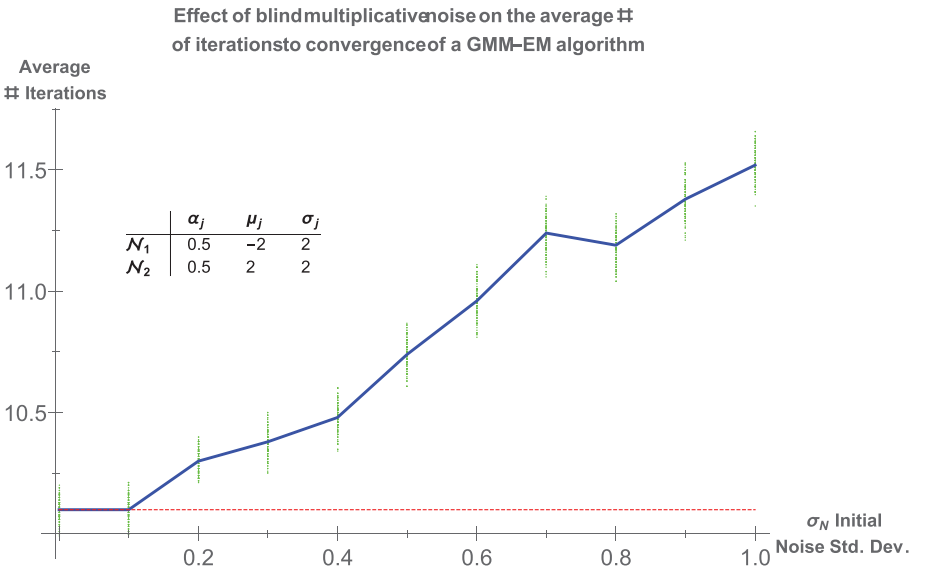

Fig. 2. Blind multiplicative noise did not improve convergence time when using the EM algorithm to estimate the parameters of the two-pdf Gaussian mixture model $f(y) = \frac{1}{2}N_1(-2,4) + \frac{1}{2}N_2(2,4)$. Such blind noise only increased the average number of iterations that it took the EM algorithm to converge. This increase in convergence time occurred even when (as in this case) the multiplicative noise used a cooling schedule to gradually shut off the noise injection.

the difference of means gave the very large bootstrap p-value (achieved significance level [12]) of 0.492 based on 10,000 bootstraps. That large p-value argues strongly against rejecting the null hypothesis that there was no statistically significant difference in the optimal average convergence times of the additive and multiplicative NEM speed-ups.

A 95%-bootstrap confidence interval for the average difference in optimal convergence time was $(-0.44, 0.06)$. The confidence interval contained zero. So we cannot reject the null hypothesis that the difference in optimal average convergence times for the two noise-injection modes was statistically insignificant at the 0.05 level. Nor can we reject the null hypothesis at the 0.10 and 0.01 significance levels because their respective 90% and 99% bootstrap confidence intervals were $(-0.40, 0.02)$ and $(-0.52, 0.13)$. So there was no statistically significant difference in the noise speed-ups of the additive and multiplicative cases. An open and important research question is whether there are general conditions under which one of these noise injection modes outperforms the other.

2. General Noise Injection for a NEM Benefit

We next generalize the original proof for additive NEM [21, 23] to NEM that uses an arbitrary mode of noise injection. The metrical idea behind the proof remains

the same: a noise benefit occurs on average at an iteration if the noisy pdf is closer to the optimal pdf than the noiseless pdf is.

The relative entropy $D(h||g)$ measures the pseudo-distance between two pdfs h and g in a topological space of pdfs. The literature sometimes refers to the relative entropy as the Kullback–Leibler divergence [13]. An EM noise benefit occurs at iteration k if the noise-injected pdf f_N is closer to the optimal or maximum-likelihood pdf than the noiseless pdf f is:

$$D(f(y,z|\theta_*)\|f_N(y,z|\theta_k)) \leq D(f(y,z|\theta_*)\|f(y,z|\theta_k)), \tag{5}$$

where

$$f_N(y,z|\theta_k) = f(\phi(y,N),z|\theta_k), \tag{6}$$

is the noise-injected pdf for arbitrary measurable function ϕ.

The relative entropy is asymmetric and has the form of an average logarithm

$$D(h(u,v)||g(u,v)) = \int_{\mathcal{U}}\int_{\mathcal{V}} \ln\left[\frac{h(u,v)}{g(u,v)}\right] h(u,v)\mathrm{d}u\mathrm{d}v, \tag{7}$$

for positive pdfs h and g over the same support [13]. Convergent sums can replace the integrals in the discrete case. We follow convention in calling the relative entropy a pseudo-metric. It is technically only a pre-metric because the relative entropy between two pdfs is always non-negative. The relative entropy is zero if and only if the two pdfs are equal almost everywhere. This yields the proof strategy of reducing the relative entropy with respect to the optimal pdf at each iteration k.

The key point is that the noise-injection mode $\phi(y,N)$ need be neither addition $\phi(y,N) = y + N$ nor multiplication $\phi(y,N) = yN$. It can be any measurable function ϕ of the data y and the noise N. This generality does not affect the main proofs for a noise benefit. The proof of Theorem 1 below demonstrates this.

The above relative entropy inequality (5) is logically equivalent to the EM noise-benefit condition at iteration k if we cast the noise benefit in terms of expectations [21]:

$$\mathbb{E}[Q(\theta_*|\theta_*) - Q_N(\theta_k|\theta_*)] \leq \mathbb{E}[Q(\theta_*|\theta_*) - Q(\theta_k|\theta_*)], \tag{8}$$

where Q_N is the noise-perturbed surrogate likelihood function

$$Q_N(\theta|\theta_k) = \mathbb{E}_{Z|Y,\theta_k}[\ln f_N(y,Z|\theta)]. \tag{9}$$

Any noise N that satisfies this EM noise-benefit condition (8) will on average give better parameter estimates at each iteration than will noiseless estimates or those that use blind noise. The relative-entropy version of the noise-benefit condition allows the same derivation of the generalized NEM condition as in the original case of additive noise. The result is Theorem 1. The proof assumes finite differential entropies.

Theorem 1 (Arbitrary Noise Injection NEM Theorem). *Let $\phi(Y,N)$ be an arbitrary measurable mode of combining the signal Y with the noise N. Suppose the*

NEM average positivity condition holds at iteration k:

$$\mathbb{E}_{Y,Z,N|\theta_*}\left[\ln\left(\frac{f(\phi(Y,N),Z|\theta_k)}{f(Y,Z|\theta_k)}\right)\right] \geq 0. \tag{10}$$

Then the EM noise benefit

$$Q(\theta_k|\theta_*) \leq Q_N(\theta_k|\theta_*) \tag{11}$$

holds on average at iteration k:

$$\mathbb{E}_{N,Y|\theta_k}[Q(\theta_*|\theta_*) - Q_N(\theta_k|\theta_*)] \leq \mathbb{E}_{Y|\theta_k}[Q(\theta_*|\theta_*) - Q(\theta_k|\theta_*)]. \tag{12}$$

Proof. The proof shows that the noisy pdf (or likelihood) $f(\phi(y,N),z|\theta_k)$ is closer on average to the optimal pdf $f(y,z|\theta_*)$ than the noiseless pdf $f(y,z|\theta_k)$ is. We use relative entropy for this pdf comparison.

Let c_k denote the relative entropy between the optimal likelihood and the noiseless likelihood at iteration k:

$$c_k = D(f(y,z|\theta_*)\|f(y,z|\theta_k)). \tag{13}$$

Let $c_k(N)$ denote the relative entropy between the optimal likelihood and the noisy likelihood:

$$c_k(N) = D(f(y,z|\theta_*)\|f(\phi(y,N),z|\theta_k)). \tag{14}$$

The notation $c_k(N)$ uses upper-case N rather than lower-case n to emphasize that this relative entropy is a random variable because the included noise term N is a random variable. Then the proof derives the average noise-benefit inequality $c_k \geq \mathbb{E}_N[c_k(N)]$ at iteration k.

We first show that the expectation of the Q-function differences in (8) inherits the pseudo-metrical structure of relative entropy. Write the Q-function expectation as an integral over Z:

$$Q(\theta|\theta_k) = \int_Z \ln[f(y,z|\theta)]f(z|y,\theta_k)dz. \tag{15}$$

Then the relative-entropy term $c_k = D(f(y,z|\theta_*)\|f(y,z|\theta_k))$ is the expectation over Y given the current parameter value θ_k of the difference of Q-functions. This holds because factoring the conditional pdf $f(y,z|\theta_*)$ gives $f(y,z|\theta_*) = f(z|y,\theta_*)f(y|\theta_*)$:

$$c_k = \iint \ln\left[\frac{f(y,z|\theta_*)}{f(y,z|\theta_k)}\right] f(y,z|\theta_*)dz\,dy \tag{16}$$

$$= \iint [\ln(f(y,z|\theta_*)) - \ln f(y,z|\theta_k)]f(y,z|\theta_*)dz\,dy \tag{17}$$

$$= \iint [\ln(f(y,z|\theta_*)) - \ln f(y,z|\theta_k)]f(z|y,\theta_*)f(y|\theta_*)dz\,dy \tag{18}$$

$$= \int_{Y|\theta_k} \left[\int_Z \ln[(f(y,z|\theta_*)] f(z|y,\theta_*) dz \right.$$

$$\left. - \int_Z \ln[f(y,z|\theta_k)] f(z|y,\theta_*) dz \right] f(y|\theta_*) dy \tag{19}$$

$$= \int_{Y|\theta_k} [Q(\theta_*|\theta_*) - Q(\theta_k|\theta_*)] f(y|\theta_*) dy \tag{20}$$

$$= \mathbb{E}_{Y|\theta_k}[Q(\theta_*|\theta_*) - Q(\theta_k|\theta_*)]. \tag{21}$$

The noise-injected term $c_k(N)$ similarly equals the expectation over Y given θ_k:

$$c_k(N) = \iint [\ln(f(y,z|\theta_*)) - \ln f(\phi(y,N),z|\theta_k)] f(y,z|\theta_*) dz\, dy \tag{22}$$

$$= \iint [\ln(f(y,z|\theta_*)) - \ln f(\phi(y,N),z|\theta_k)] f(z|y,\theta_*) f(y|\theta_*) dz\, dy \tag{23}$$

$$= \mathbb{E}_{Y|\theta_k}[Q(\theta_*|\theta_*) - Q_N(\theta_k|\theta_*)]. \tag{24}$$

So the expected Q-difference is equivalent to relative entropy. And so it has the same pseudo-metrical structure. We note also that c_k is a constant at each iteration k. But $c_k(N)$ is a random variable since the expectation in (24) does not average out the noise N.

Take noise expectations over both terms c_k and $c_k(N)$:

$$\mathbb{E}_N[c_k] = c_k, \tag{25}$$

$$\mathbb{E}_N[c_k(N)] = \mathbb{E}_N[c_k(N)]. \tag{26}$$

Then the pseudo-metrical inequality

$$c_k \geq \mathbb{E}_N[c_k(N)] \tag{27}$$

ensures an average noise benefit at iteration k:

$$\mathbb{E}_{N,Y|\theta_k}[Q(\theta_*|\theta_*) - Q_N(\theta_k|\theta_*)] \leq \mathbb{E}_{N,Y|\theta_k}[Q(\theta_*|\theta_*) - Q(\theta_k|\theta_*)]. \tag{28}$$

We use the inequality condition (27) above to derive a more useful sufficient condition for a noise benefit. Expand the difference of the relative-entropy terms $c_k - c_k(N)$:

$$c_k - c_k(N)$$

$$= \iint_{Y,Z} \left(\ln \left[\frac{f(y,z|\theta_*)}{f(y,z|\theta_k)} \right] - \ln \left[\frac{f(y,z|\theta_*)}{f(\phi(y,N),z|\theta_k)} \right] \right) f(y,z|\theta_*) dy\, dz \tag{29}$$

$$= \iint_{Y,Z} \left(\ln \left[\frac{f(y,z|\theta_*)}{f(y,z|\theta_k)} \right] + \ln \left[\frac{f(\phi(y,N),z|\theta_k)}{f(y,z|\theta_*)} \right] \right) f(y,z|\theta_*) dy\, dz \tag{30}$$

$$= \iint_{Y,Z} \ln \left[\frac{f(y,z|\theta_*) f(\phi(y,N),z|\theta_k)}{f(y,z|\theta_k) f(y,z|\theta_*)} \right] f(y,z|\theta_*) dy\, dz \tag{31}$$

$$= \iint_{Y,Z} \ln \left[\frac{f(\phi(y,N),z|\theta_k)}{f(y,z|\theta_k)} \right] f(y,z|\theta_*) dy\, dz. \tag{32}$$

Then take the expectation with respect to the noise random variable N:

$$\begin{aligned}
&\mathbb{E}_N[c_k - c_k(N)] \\
&\quad = c_k - \mathbb{E}_N[c_k(N)] && (33)\\
&\quad = \int_N \iint_{Y,Z} \ln\left[\frac{f(\phi(y,n),z|\theta_k)}{f(y,z|\theta_k)}\right] f(y,z|\theta_*) f(n|y) dy\, dz\, dn && (34)\\
&\quad = \iint_{Y,Z} \int_N \ln\left[\frac{f(\phi(y,n),z|\theta_k)}{f(y,z|\theta_k)}\right] f(n|y) f(y,z|\theta_*) dn\, dy\, dz && (35)\\
&\quad = \iint_{Y,Z} \int_N \ln\left[\frac{f(\phi(y,n),z|\theta_k)}{f(y,z|\theta_k)}\right] f(n|y,z,\theta_*) f(y,z|\theta_*) dn\, dy\, dz && (36)\\
&\quad = \iint_{Y,Z} \int_N \ln\left[\frac{f(\phi(y,n),z|\theta_k)}{f(y,z|\theta_k)}\right] f(n,y,z|\theta_*) dn\, dy\, dz && (37)\\
&\quad = \mathbb{E}_{Y,Z,N|\theta_*}\left[\ln \frac{f(\phi(Y,N),Z|\theta_k)}{f(Y,Z|\theta_k)}\right]. && (38)
\end{aligned}$$

The assumption of finite differential entropy for Y and Z ensures that $\ln f(y,z|\theta) f(y,z|\theta_*)$ is integrable. So the integrand is integrable. Then Fubini's theorem [38] permits the change in the order of integration in the above multiple integral. The pdf equality $f(n|y,z,\theta_*) = f(n|y)$ holds because the noise random variable N does not depend on the latent variable Z or on the optimal parameter value θ_*. N does depend on the signal random variable Y in general. Then factorization gives the pdf equality $f(n|y,z,\theta_*) f(y,z|\theta_*) = \frac{f(n,y,z|\theta_*)}{f(y,z|\theta_*)} f(y,z|\theta_*) = f(n,y,z|\theta_*)$.

The result is the logical equivalence

$$\begin{aligned}
&c_k \geq \mathbb{E}_N[c_k(N)] \quad \text{if and only if} \\
&\mathbb{E}_{Y,Z,N|\theta_*}\left[\ln\left(\frac{f(\phi(Y,N),Z|\theta_k)}{f(Y,Z|\theta_k)}\right)\right] \geq 0. && (39)
\end{aligned}$$

Then an EM noise benefit occurs on average at iteration k if

$$\mathbb{E}_{Y,Z,N|\theta_*}\left[\ln\left(\frac{f(\phi(Y,N),Z|\theta_k)}{f(Y,Z|\theta_k)}\right)\right] \geq 0. \quad (40)$$

□

3. The Special Case of Multiplicative NEM

Theorem 1 allows a direct proof that properly chosen multiplicative noise can speed average EM convergence. The proof requires only that the noise-injection mode ϕ be m-NEM:

$$\phi(Y,N) = YN. \quad (41)$$

Then Theorem 1 gives the following special case for multiplicative noise. We state this result as Theorem 2 because of the importance of mulitplicative noise injection.

Theorem 2 (m-NEM Theorem). *Suppose the average posiitivity condition holds for multiplicative noise injection at iteration k:*

$$\mathbb{E}_{Y,Z,N|\theta_*}\left[\ln\left(\frac{f(YN,Z|\theta_k)}{f(Y,Z|\theta_k)}\right)\right] \geq 0. \tag{42}$$

Then the EM noise benefit

$$Q(\theta_k|\theta_*) \leq \; Q_N(\theta_k|\theta_*) \tag{43}$$

holds on average at iteration k:

$$\mathbb{E}_{N,Y|\theta_k}[Q(\theta_*|\theta_*) - Q_N(\theta_k|\theta_*)] \leq \mathbb{E}_{Y|\theta_k}[Q(\theta_*|\theta_*) - Q(\theta_k|\theta_*)]. \tag{44}$$

The next section develops the important case of a mixture data model. A key result is that we can derive mixture NEM benefits by deriving such benefits for the individual mixed pdfs.

4. Noise-Boosting Mixture Models

Mixture models are by far the most common data models in EM applications. Mixtures allow a user to create a tunable multimodal pdf by mixing unimodal pdfs. So this section develops their NEM noise-boosting at some length.

Many of the additive-NEM mixture results apply to the generalized NEM condition without change. Corollary 2 from [21] leads to an m-NEM condition for a Gaussian mixture model (GMM) because the noise condition applies to each mixed normal pdf in the mixture. An identical multiplicative noise-benefit condition holds for a Cauchy mixture model. We state and prove the m-NEM GMM result as a separate corollary. The resulting quadratic m-NEM condition depends only on the Gaussian means and not on their variances.

We first review mixture models. This will generalize the simple two-Gaussian GMM in Figs. 1 and 2 where $f(y) = \alpha f_1(y|\mu_1, \sigma_1^2) + (1-\alpha) f_2(y|\mu_2, \sigma_2^2)$. The EM algorithm offers a practical way to estimate the parameters of a mixture model. The parameters consist of the convex or probability mixing weights α_j as well as the individual parameters of the mixed pdfs.

A finite mixture model [12, 39–41] is a convex combination of a finite number of similar pdfs. So a mixture is a convex or probabilistic combination of similar sub-populations. The sub-population pdfs are similar in the sense that they all come from the same parametric family. Mixture models apply to a wide range of statistical problems in pattern recognition and machine learning [42, 43]. A Gaussian mixture consists of convex-weighted normal pdfs. The EM algorithm estimates the mixture weights as well as the means and variances of each mixed normal pdf.

A Cauchy mixture consists likewise of convex-weighted Cauchy pdfs. The GMM is by far the most common mixture model in practice [44].

Let Y be the observed mixed random variable. Let K be the number of sub-populations. Let $Z \in \{1, \ldots, K\}$ be the hidden sub-population index random variable. The convex population mixing proportions $\alpha_1, \ldots, \alpha_K$ define a discrete pdf for Z: $P(Z = j) = \alpha_j$. The pdf $f(y|Z = j, \theta_j)$ is the pdf of the jth sub-population where $\theta_1, \ldots, \theta_K$ are the pdf parameters for each sub-population. We can also denote this jth mixed density as $f_j(y|\theta_j)$ as in Figs. 1 and 2. The sub-population parameter θ_j can represent the mean or variance of a normal pdf or both. It can represent any number of quantities that parametrize the pdf.

Let Θ denote the vector of all model parameters: $\Theta = \{\alpha_1, \ldots, \alpha_K, \theta_1, \ldots, \theta_K\}$. The mixing weights $\alpha_1, \ldots, \alpha_K$ are convex coefficients. This means that they are non-negative and add to unity. So again they define a discrete probability density for latent or hidden variable Z. Then the joint or "complete" pdf $f(y, z|\Theta)$ is

$$f(y, z|\Theta) = \sum_{j=1}^{K} \alpha_j \; f(y|j, \theta_j) \; \delta[z - j], \tag{45}$$

where $\delta[z - j] = 1$ if $z = j$ and $\delta[z - j] = 0$ otherwise. The K likelihoods $f(y|j, \theta_j)$ are the mixed pdfs in the finite mixture. Their structure determines the sufficient condition for an m-NEM noise benefit.

The marginal pdf for Y and the conditional posterior pdf for Z given y are

$$f(y|\Theta) = \sum_j \alpha_j f(y|j, \theta_j), \tag{46}$$

$$\text{and} \quad p_Z(j|y, \Theta) = \frac{\alpha_j f(y|Z = j, \theta_j)}{f(y|\Theta)}. \tag{47}$$

The marginal $f(y|\Theta)$ has the sum structure (46) after summing over all z terms on both sides of (45) because of the delta term $\delta[z - j]$. The ratio form (47) of the posterior $p_Z(j|y, \Theta)$ follows from Bayes theorem. This holds because (46) is just the theorem on total probability since the convex mixing weights α_j are the prior probabilities $P(Z = j)$ and since the pdfs $f(y|Z = j, \theta_j)$ are likelihoods. These posterior pdfs or "responsibilities" [43] are crucial in the EM update equations below for a Gaussian mixture model.

Rewrite the joint pdf as an exponential (since $\delta(z - j) = 0$ unless $z = j$):

$$f(y, z|\Theta) = \exp\left[\sum_j [\ln(\alpha_j) + \ln f(y|j, \theta_j)]\delta[z - j]\right]. \tag{48}$$

This gives a simple linear form for the log-likelihood:

$$\ln f(y, z|\Theta) = \sum_j \delta[z - j] \ln[\alpha_j f(y|j, \theta_j)]. \tag{49}$$

EM algorithms for finite mixture models estimate Θ using the sub-population index Z as the latent variable. An EM algorithm uses (47) to derive the Q-function

$$Q(\Theta|\Theta_k) = \mathbb{E}_{Z|y,\Theta_k}[\ln f(y, Z|\Theta)] \tag{50}$$

$$= \sum_z \left(\sum_j \delta[z-j] \ln[\alpha_j f(y|j,\theta_j)] \right) p_Z(z|y,\Theta_k) \tag{51}$$

$$= \sum_j \ln[\alpha_j f(y|j,\theta_j)] p_Z(j|y,\Theta_k). \tag{52}$$

The EM algorithm has an especially simple form for estimating the parameters Θ_k of a Gaussian mixture model [42, 43]. EM estimates the K mixing probabilities α_j, the K sub-population means μ_j, and the K sub-population variances σ_j^2. Θ_k gives the current estimate of the GMM parameters: $\Theta_k = \{\alpha_1(k), \ldots, \alpha_K(k), \mu_1(k), \ldots, \mu_K(k), \sigma_1^2(k), \ldots, \sigma_K^2(k)\}$. Then the iterations of the GMM–EM algorithm reduce to the following update equations based on the K posterior pdfs $p_Z(j|y,\Theta)$ in (47):

$$\alpha_j(k+1) = \frac{1}{N} \sum_{i=1}^N p_Z(j|y_i,\Theta_k), \tag{53}$$

$$\mu_j(k+1) = \frac{\sum_{i=1}^N p_Z(j|y_i,\Theta_k) y_i}{\sum_{i=1}^N p_Z(j|y_i,\Theta_k)}, \tag{54}$$

$$\sigma_j^2(k+1) = \frac{\sum_{i=1}^N p_Z(j|y_i,\Theta_k)(y_i - \mu_j(k))^2}{\sum_{i=1}^N p_Z(j|y_i,\Theta_k)}. \tag{55}$$

These equations update the parameters α_j, μ_j, and σ_j^2 with coordinate values that maximize the Q function in (52). These equations also updated the parameters in the GMMs in Figs. 1 and 2.

The updates effectively combine both the E-steps and M-steps of the EM procedure. We can alternatively view the E-step as computing the K Bayesian posteriors $p_Z(j|y,\Theta_k)$ that appear in the Q-function in (52). Then the M-step corresponds to computing the above three updates for $\alpha_j(k+1)$, $\mu_j(k+1)$, and $\sigma_j^2(k+1)$.

We turn now to noise-boosting a mixture model. Corollary 1 from [21] gives a simple pdf-inequality condition when additive noise satisfies the additive NEM condition (4) for almost all samples y:

$$f(y+n, z|\theta) \geq f(y, z|\theta). \tag{56}$$

We can derive similar NEM conditions for mixture models. The complete data likelihood of a mixture model

$$f(y, z|\Theta) = \sum_j \alpha_j f(y|j,\theta_j) \delta[z-j] \tag{57}$$

allows us to rewrite (56) as

$$f(y+n,z|\Theta) - f(y,z|\Theta) = \sum_j \alpha_j \delta[z-j](f(y+n|j,\theta_j) - f(y|j,\theta_j)) \quad (58)$$

$$= \sum_j \alpha_j \delta[z-j]\Delta f_j(y,n), \quad (59)$$

where $\Delta f_j(y,n) = f(y+n|j,\theta_j) - f(y|j,\theta_j)$.

Suppose that $\Delta f_j(y,n) \geq 0$ holds for all j. Then (56) holds. So the condition gives a mixture noise benefit. Corollary 2 from [21] gives the quadratic condition when $\Delta f_j(y,n) \geq 0$ holds for all j for the case of GMM-NEM with additive noise.

We next extend the above argument to a mixture-model noise benefit condition that applies to arbitrary noise-signal combinations and arbitrary finite mixture data models.

Corollary 1 (Generalized NEM Condition for Arbitrary Mixture Models). *Suppose that $Y|_{Z=1},\ldots,Y|_{Z=K}$ are K arbitrary sub-population random variables with K corresponding sub-population pdfs $f(y|1,\theta_1),\ldots,f(y|K,\theta_K)$ and convex mixture weights $\alpha_1,\ldots,\alpha_K$. Define the mixture-model complete pdf $f(y,z|\Theta)$ as*

$$f(y,z|\Theta) = \sum_j \alpha_j f(y|j,\theta_j)\delta[z-j], \quad (60)$$

so that summing over the hidden-variable values z gives the marginal mixture density

$$f(y|\Theta) = \sum_j \alpha_j f(y|j,\theta_j). \quad (61)$$

Let $\phi(Y,N)$ be an arbitrary measurable mode of combining the signal Y with the noise N. Then the mixture-pdf NEM noise benefit for general noise injection

$$f(\phi(y,n),z|\Theta) \geq f(y,z|\Theta), \quad (62)$$

holds if

$$\Delta f_j(y,n) \geq 0 \quad \textit{for all } j, \quad (63)$$

where $\Delta f_j(y,n) = f(\phi(y,n)|j,\theta_j) - f(y|j,\theta_j)$.

Proof. We first show that the mixture inequality (62) invokes the NEM noise benefit of Theorem 1. The complete mixture-pdf noise-benefit inequality

$$f(\phi(y,n),z|\Theta) \geq f(y,z|\Theta), \quad (64)$$

holds if and only if

$$\ln \frac{f(\phi(y,n),z|\Theta)}{f(y,z|\Theta)} \geq 0. \quad (65)$$

Take expectations on both sides of this inequality to get

$$\mathbb{E}_{Y,Z,N|\theta_*}\left[\ln\left(\frac{f(\phi(Y,N),Z|\theta_k)}{f(Y,Z|\theta_k)}\right)\right] \geq 0. \tag{66}$$

This is just the sufficient NEM condition of Theorem 1 at iteration k.

We show next that the mixture inequality (62) holds if $\Delta f_j(y,n) \geq 0$ holds for all j. Then the expansion (60) implies that the mixture inequality (62) holds if and only if (iff)

$$\sum_j \alpha_j f(\phi(y,n)|j,\theta_j)\delta[z-j] \geq \sum_j \alpha_j f(y|j,\theta_j)\delta[z-j], \tag{67}$$

$$\text{iff } \sum_j \alpha_j \delta[z-j][f(\phi(y,n)|j,\theta_j) - f(y|j,\theta_j)] \geq 0, \tag{68}$$

$$\text{iff } \sum_j \alpha_j \delta[z-j][\Delta f_j(y,n)] \geq 0. \tag{69}$$

Then the mixture inequality (62) holds if

$$\Delta f_j(y,n) \geq 0, \quad \text{for } j = 1,\ldots,K, \tag{70}$$

since the mixing weights α_j and delta functions $\delta[z-j]$ are non-negative. So the finite mixture model enjoys a NEM noise benefit if each mixed density obeys $f(\phi(y,n)|j,\theta_j) \geq f(y|j,\theta_j)$. □

Corollary 1 allows us to ignore the mixture structure of arbitrary mixture models. We can derive NEM sufficient conditions by just focusing on the K subpopulation pdfs.

We next state and prove a sufficient NEM condition for the special case of multiplicative NEM in a Gaussian mixture model: $\Delta f_j(y,n) = f(yn|j,\theta_j) - f(y|j,\theta_j) \geq 0$ in this case. We use the pdf condition (70) instead of the sum condition. The resulting pdf NEM condition has a simple quadratic form that depends only on the noise terms and the Gaussian population means μ_j.

Corollary 2 (m-NEM Condition for Gaussian Mixture Models). *Suppose that $Y|_{Z=j} \sim \mathcal{N}(\mu_j, \sigma_j^2)$ in the finite mixture of K Gaussian pdfs $f(y|1,\theta_1),\ldots,f(y|K,\theta_K)$. Then the mixture-pdf NEM noise benefit for multiplicative noise*

$$f(yn|\theta_j) \geq f(y|\theta_j), \tag{71}$$

holds for each mixed pdf if

$$y(n-1)\left[y(n+1) - 2\mu_j\right] \leq 0, \tag{72}$$

for $j = 1,\ldots,K$.

Proof. Corollary 1 allows us to prove this mixture result by just proving the pdf noise-benefit inequality for all K mixed pdfs $f(y|\theta_j)$. So compare the noise-injected normal pdf $f(yn|\theta_j)$ with the noiseless normal pdf $f(y|\theta_j)$. The normal pdf is

$$f(y|\theta_j) = \frac{1}{\sigma_j\sqrt{2\pi}} \exp\left[-\frac{(y-\mu_j)^2}{2\sigma_j^2}\right]. \tag{73}$$

So $f(yn|\theta_j) \geq f(y|\theta_j)$ holds

$$\text{iff } \exp\left[-\frac{(yn-\mu_j)^2}{2\sigma_j^2}\right] \geq \exp\left[-\frac{(y-\mu_j)^2}{2\sigma_j^2}\right] \tag{74}$$

$$\text{iff } -\left(\frac{yn-\mu_j}{\sigma_j}\right)^2 \geq -\left(\frac{y-\mu_j}{\sigma_j}\right)^2 \tag{75}$$

$$\text{iff } (yn-\mu_j)^2 \leq (y-\mu_j)^2 \tag{76}$$

$$\text{iff } y^2n^2 + \mu_j^2 - 2\mu_j yn \leq y^2 + \mu_j^2 - 2y\mu_j \tag{77}$$

$$\text{iff } y^2n^2 - 2\mu_j yn \leq y^2 - 2y\mu_j \tag{78}$$

$$\text{iff } y^2(n^2-1) - 2y\mu_j(n-1) \leq 0 \tag{79}$$

$$\text{iff } y(n-1)\,[y(n+1) - 2\mu_j] \leq 0. \tag{80}$$

So (72) and Corollary 1 imply the n-NEM mixture condition (71). □

The identical quadratic m-NEM noise-benefit condition (72) holds for a Cauchy mixture model. Suppose that $Y|_{Z=j} \sim \mathcal{C}(m_j, d_j)$. So $f(y|j,\theta_j)$ is a Cauchy pdf with median m_j and dispersion d_j. The median controls the location of the Cauchy bell curve. The dispersion controls its width. A Cauchy random variable has no mean. It does have finite lower-order fractional moments. But its variance and all its higher-order moments are either infinite or not defined. The Cauchy pdf $f(y|j,\theta_j)$ has the form

$$f(y|\theta_j) = \frac{1}{\pi d_j\left[1 + \left(\frac{y-m_j}{d_j}\right)^2\right]}. \tag{81}$$

Then the mixed-pdf inequality $f(yn|\theta_j) \geq f(y|\theta_j)$ is equivalent to the same quadratic inequality as in the above derivation of the Gaussian m-NEM condition. This gives (72) as the Cauchy m-NEM noise-benefit condition with the median m_j replacing the mean μ_j: $y(n-1)[y(n+1) - 2m_j] \leq 0$ for all j mixed pdfs.

5. The m-NEM Algorithm

The m-NEM Theorem and its corollaries give a general method for noise-boosting the EM algorithm. Theorem 1 implies that on average these NEM variants outperform the noiseless EM algorithm.

Algorithm 1 gives the multiplicative-NEM algorithm schema. The operation MNEMNOISESAMPLE($\mathbf{y},\ k^{-\tau}\sigma_N$) generates noise samples that satisfy the m-NEM condition for the current data model. The noise sampling pdf depends on the vector of random samples $\mathbf{y}$ in the Gaussian and Cauchy mixture models. The noise can have any value in the m-NEM algorithm for censored gamma models that are log-convex [21]. Censorship means setting a finite upper bound for gamma random samples because the gamma pdf is an infinite right-sided density.

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

Require: $\mathbf{y} = (y_1, \ldots, y_M)$: vector of observed incomplete data
Ensure: $\hat{\theta}_{m\text{NEM}}$: m-NEM estimate of parameter θ
1: **while** $(\|\theta_k - \theta_{k-1}\| \geq 10^{-\text{tol}})$ **do**
2: $\quad$ **N$_\mathbf{S}$-Step:** $\mathbf{n} \leftarrow$ mNEMNoiseSample($\mathbf{y},\ k^{-\tau}\sigma_N$)
3: $\quad$ **N$_\mathbf{M}$-Step:** $\mathbf{y}_\dagger \leftarrow \mathbf{yn}$
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: $\hat{\theta}_{m\text{NEM}} \leftarrow \theta_k$

The E-Step takes a conditional expectation of a function of the noisy data samples $\mathbf{y}_\dagger$ given the noiseless data samples $\mathbf{y}$. The M-Step maximizes the resulting surrogate likelihood function over all parameters θ.

A deterministic decay factor $k^{-\tau}$ scaled the noise on the kth iteration. It did this by replacing the fixed standard deviation σ_N of the noise with the weighted standard deviation $k^{-\tau}\sigma_N$. So the m-NEM noise had slightly smaller standard deviation with each successive iteration. τ was the noise decay rate [21]. The decay factor drove the noise N_k to zero as the iteration step k increased. This eventually shut off the noise injection. We found that the value $\tau = 2$ worked best in the simulations and thus we used an inverse-square scaling k^{-2}.

The inverse-square decay factor reduced the NEM estimator's jitter around its final value. This was important because the EM algorithm converges to fixed-points. Excessive estimator jitter prolongs convergence time even when the jitter occurs near the final solution. Our simulations used the inverse-square (hence polynomial) decay factor instead of the logarithmic cooling schedules found in most applications of simulated annealing [37, 45–49].

The NEM noise generating procedure MNEMNOISESAMPLE($\mathbf{y}, k^{-\tau}\sigma_N$) returned an m-NEM-compliant noise sample n at a given noise level σ_N for each data sample y. This procedure changed with the EM data model. The following noise generating procedure applied to GMMs in accord with the above corollary for m-NEM GMMS. We used the following 1D noise generating procedure for the

GMM simulations:

Algorithm 2* mNEMNoiseSample for GMM-m-NEM

Require: y and σ_N : current data sample and noise level
Ensure: n : noise sample satisfying NEM condition
$N(y) \leftarrow \{n \in \mathcal{R} : y(n-1)\,[y(n+1) - 2\mu_j] \leq 0\}$
n is a sample from the distribution $TN(1, \sigma_N|N(y))$

The term $TN(1, \sigma_N|N(y))$ denotes a truncated Gaussian pdf over some finite-length support. The term $N(y)$ denotes the corresponding NEM noise support that depends on the data sample y.

Figure 1 displays a noise benefit for a m-NEM algorithm on the GMM that evenly mixes two Gaussian pdfs: $f(y) = \frac{1}{2}N_1(-2, 4) + \frac{1}{2}N_2(2, 4)$. The injected noise is subject to the Gaussian m-NEM condition in (72).

The next section develops the NEM theory for the important case of exponential family pdfs. The corresponding theorem states the generalized NEM condition for arbitrary modes of noise injection. The theorem also applies to mixtures of exponential family pdfs because of Corollary 1.

6. NEM Noise Benefits for Exponential Family Probabilities

This section derives the NEM condition for the general exponential family of pdfs. Exponential family pdfs include such popular densities as the normal, exponential, gamma, and Poisson [12]. A member of this exponential family has a pdf $f(y|\theta)$ of the exponential form

$$f(y|\theta) = \exp[a(\theta)K(y) + b(y) + c(\theta)], \tag{82}$$

if the density's domain does not include the parameter θ. This latter condition bars the uniform pdf from the exponential family. The exponential family also excludes Cauchy and Student-t pdfs.

Direct substitutions show that the Gaussian or normal pdf belongs to the exponential family. The normal pdf $f(y|\theta) = \frac{1}{\sigma\sqrt{2\pi}}\exp[-\frac{(y-\mu)^2}{2\sigma^2}]$ has the exponential-family form given the substitutions $a(\theta) = \frac{\mu}{\sigma^2}$, $K(y) = y$, $b(y) = -\frac{y^2}{2\sigma^2} - \ln\sqrt{2\pi\sigma^2}$, and $c(\theta) = -\frac{\mu^2}{2\sigma^2}$.

The next theorem states the NEM condition for an exponential-family pdf and arbitrary combination of signal and noise. The result shows that the noise benefit does not depend on the term $c(\theta)$.

Theorem 3 (Arbitrary Noise Injection NEM Condition for Exponential Family Probability Density Functions). *Suppose the signal Y has an exponential family pdf*:

$$f(y|\theta) = \exp[a(\theta)K(y) + b(y) + c(\theta)]. \tag{83}$$

Let $\phi(Y, N)$ be an arbitrary measurable mode of combining Y with the noise N. Then an EM noise benefit occurs if

$$a(\theta)[K(\phi(y,n)) - K(y)] + b(\phi(y,n)) - b(y) \geq 0. \tag{84}$$

Proof. Compare the noisy pdf $f(\phi(x,n)|\theta)$ with the noiseless pdf $f(x|\theta)$. The noise benefit occurs if

$$\ln f(\phi(y,n)|\theta) \geq \ln f(y|\theta), \tag{85}$$

since the logarithm is a monotone increasing function. This inequality holds

$$\text{iff } a(\theta)K(\phi(y,n)) + b(\phi(y,n)) + c(\theta) \geq a(\theta)K(y) + b(y) + c(\theta) \tag{86}$$

$$\text{iff } a(\theta)K(\phi(y,n)) + b(\phi(y,n)) \geq a(\theta)K(y) + b(y) \tag{87}$$

$$\text{iff } a(\theta)[K(\phi(y,n)) - K(y)] + b(\phi(y,n)) - b(y) \geq 0. \tag{88}$$

The last inequality is just (84). □

The exponential-family noise-benefit condition reduces to

$$a(\theta)[K(y+n)) - K(y)] + b(y+n)) - b(y) \geq 0 \tag{89}$$

in the additive noise case when $\phi(y,n) = y + n$. It reduces to

$$a(\theta)[K(yn)) - K(y)] + b(yn)) - b(y) \geq 0 \tag{90}$$

in the multiplicative noise case when $\phi(y,n) = yn$. The $c(\theta)$ term does not appear in the NEM conditions.

Consider the exponential signal pdf $f(y|\theta) = \frac{1}{\theta}e^{-\frac{y}{\theta}}$. It is an exponential-family pdf because $a(\theta) = -\frac{1}{\theta}$, $K(y) = y$, $b(y) = 0$, and $c(\theta) = -\ln\theta$. So the condition for an additive NEM noise benefit becomes

$$-\frac{1}{\theta}[y + n - y] \geq 0. \tag{91}$$

This gives a simple negative noise condition for an additive NEM benefit:

$$n \leq 0. \tag{92}$$

So the NEM condition does not depend on the parameter θ. The condition for a multiplicative NEM benefit likewise becomes

$$-\frac{1}{\theta}[yn - y] \geq 0. \tag{93}$$

It gives a similar linear NEM condition:

$$n \leq 1. \tag{94}$$

We conclude with a dual theorem that guarantees that some noise will *harm* the EM algorithm by slowing its average convergence. Both the theorem statement and its proof simply reverse all pertinent inequalities in Theorem 1.

Theorem 4 (The Generalized EM Noise-Harm Theorem). *Let $\phi(Y, N)$ be an arbitrary measurable mode of combining the signal Y and the noise N. Suppose the average negativity condition holds at iteration k:*

$$\mathbb{E}_{Y,Z,N|\theta_*}\left[\ln\left(\frac{f(\phi(Y,N),Z|\theta_k)}{f(Y,Z|\theta_k)}\right)\right] \leq 0. \tag{95}$$

Then the EM noise harm

$$Q(\theta_k|\theta_*) \geq \; Q_N(\theta_k|\theta_*), \tag{96}$$

holds on average at iteration k:

$$\mathbb{E}_{N,Y|\theta_k}[Q(\theta_*|\theta_*) - Q_N(\theta_k|\theta_*)] \geq \mathbb{E}_{Y|\theta_k}[Q(\theta_*|\theta_*) - Q(\theta_k|\theta_*)]. \tag{97}$$

Proof. The proof follows from the same argument that proves Theorem 1 if we reverse all inequalities. This applies specifically to the logical equivalence in (39). Then

$$c_k \leq \mathbb{E}_N[c_k(N)] \text{ if and only if } \mathbb{E}_{Y,Z,N|\theta_*}\left[\ln\left(\frac{f(\phi(Y,N),Z|\theta_k)}{f(Y,Z|\theta_k)}\right)\right] \leq 0. \tag{98}$$

So an EM noise harm occurs on average at iteration k if

$$\mathbb{E}_{Y,Z,N|\theta_*}\left[\ln\left(\frac{f(\phi(Y,N),Z|\theta_k)}{f(Y,Z|\theta_k)}\right)\right] \leq 0. \tag{99}$$

□

This general noise-harm result leads to corollary noise-harm conditions for the additive and multiplicative GMM-NEM models by reversing all pertinent inequalities. A similar inequality reversal gives a noise-harm condition for all pdfs from the exponential family.

Such harmful GMM noise *increased* EM convergence by 35% in the multiplicative-noise case and by 40% in the additive-noise case for the problem of estimating the parameters of the two mixed Gaussian pdfs in Figs. 1 and 2. No noise benefit or harm occurs on average if equality replaces all pertinent inequalities. We state this NEM GMM noise-harm result as a corollary for both additive and multiplicative noise injection.

Corollary 3 (Noise-Harm Conditions for GMM NEM). *The noise-harm condition in Theorem 4 holds for the GMM-NEM algorithm if*

$$n^2 \geq 2n(\mu_j - y), \tag{100}$$

for additive noise. It also holds if

$$y(n-1)[y(n+1) - 2\mu_j] \geq 0, \tag{101}$$

for multiplicative noise.

7. Conclusion

The original NEM additive noise model extends to arbitrary combinations of noise and signal. The multiplicative NEM theorem specifically gives a sufficient positivity condition such that multiplicative noise reduces the average number of iterations that the EM algorithm takes to converge. The multiplicative-noise NEM condition for the GMM and exponential family models are only slightly more complex than their respective additive-noise NEM conditions. An open research question is whether there are general conditions when either multiplicative or additive noise outperforms the other. We would also expect that data sparsity affects more general noise-injection modes as it does the additive case [21].

References

[1] L. Rudin, P.-L. Lions and S. Osher, Multiplicative denoising and deblurring: Theory and algorithms, in *Geometric Level Set Methods in Imaging, Vision, and Graphics* (Springer, 2003), pp. 103–119.
[2] J. Ash, E. Ertin, L. Potter and E. Zelnio, Wide-angle synthetic aperture radar imaging: Models and algorithms for anisotropic scattering, *IEEE Signal Process. Mag.* **31** (2014) 16–26.
[3] S. Chen, Y. Li, X. Wang, S. Xiao and M. Sato, Modeling and interpretation of scattering mechanisms in polarimetric synthetic aperture radar: Advances and perspectives, *IEEE Signal Process. Mag.* **31** (2014) 79–89.
[4] G. Aubert and J.-F. Aujol, A variational approach to removing multiplicative noise, *SIAM J. Appl. Math.* **68** (2008) 925–946.
[5] M. Tur, K. C. Chin and J. W. Goodman, When is speckle noise multiplicative?, *Appl. Opt.* **21** (1982) 1157–1159.
[6] J. M. Bioucas-Dias and M. A. Figueiredo, Multiplicative noise removal using variable splitting and constrained optimization, *IEEE Trans. Image Process.* **19** (2010) 1720–1730.
[7] J. Ringelstein, A. B. Gershman and J. F. Böhme, Direction finding in random inhomogeneous media in the presence of multiplicative noise, *IEEE Signal Process. Lett.* **7** (2000) 269–272.
[8] T. Yilmaz, C. M. Depriest, A. Braun, J. H. Abeles and P. J. Delfyett, Noise in fundamental and harmonic modelocked semiconductor lasers: Experiments and simulations, *IEEE J. Quantum Electron.* **39** (2003) 838–849.
[9] A. Swami, Multiplicative noise models: Parameter estimation using cumulants, *Signal Process.* **36** (1994) 355–373.
[10] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete Data via the EM algorithm (with discussion), *J. R. Statist. Soc. Ser. B* **39** (1977) 1–38.
[11] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions* (John Wiley & Sons, 2007).
[12] R. V. Hogg, J. McKean and A. T. Craig, *Introduction to Mathematical Statistics* (Pearson, 2013).
[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley & Sons, New York, 1991).
[14] O. Osoba and B. Kosko, Noise-enhanced clustering and competitive learning algorithms, *Neural Netw.* **37** (2013) 132–140.

[15] K. Audhkhasi, O. Osoba and B. Kosko, Noise benefits in backpropagation and deep bidirectional pre-training, in *The 2013 Int. J. Conf. on Neural Networks* (*IJCNN*) (IEEE, 2013), pp. 1–8.
[16] K. Audhkhasi, O. Osoba and B. Kosko, Noise benefits in convolutional neural networks, in *Proc. 2014 Int. Conf. on Advances in Big Data Analytics* (2014), pp. 73–80.
[17] Y. LeCun, Y. Bengio and G. Hinton, Deep learning, *Nature* **521** (2015) 436–444.
[18] K. Audhkhasi, O. Osoba and B. Kosko, Noisy hidden Markov models for speech recognition, in *Int. Joint Conf. on Neural Networks* (*IJCNN*) (IEEE, 2013), pp. 2738–2743.
[19] L. R. Welch, Hidden Markov models and the Baum-Welch algorithm, *IEEE Inf. Theory Soc. Newslett.* **53** (2003) 1–14.
[20] M. A. Tanner, *Tools for Statistical Inference*: *Methods for the Exploration of Posterior Distributions and Likelihood Functions*, Springer Series in Statistics (Springer, 1996).
[21] O. Osoba, S. Mitaim and B. Kosko, The noisy expectation-maximization algorithm, *Fluct. Noise Lett.* **12** (2013) 1350012.
[22] O. Osoba, S. Mitaim and B. Kosko, Noise benefits in the expectation-maximization algorithm: NEM theorems and models, in *The Int. Joint Conf. on Neural Networks* (*IJCNN*) (IEEE, 2011), pp. 3178–3183.
[23] O. A. Osoba, Noise benefits in expectation-maximization algorithms, Ph.D. thesis, University of Southern California (2013).
[24] K. Wiesenfeld, F. Moss *et al.*, Stochastic resonance and the benefits of noise: From ice ages to crayfish and squids, *Nature* **373** (1995) 33–36.
[25] A. R. Bulsara and L. Gammaitoni, Tuning in to noise, *Phys. Today* **49** (1996) 39–47.
[26] L. Gammaitoni, P. Hänggi, P. Jung and F. Marchesoni, Stochastic resonance, *Rev. Mod. Phys.* **70** (1998) 223.
[27] S. Mitaim and B. Kosko, Adaptive stochastic resonance, in *Proc. IEEE*: *Special Issue Intelligent Signal Process.* **86** (1998) 2152–2183.
[28] F. Chapeau-Blondeau and D. Rousseau, Noise-enhanced performance for an optimal bayesian estimator, *IEEE Trans. Signal Process.* **52** (2004) 1327–1334.
[29] I. Lee, X. Liu, C. Zhou and B. Kosko, Noise-enhanced detection of subthreshold signals with carbon nanotubes, *IEEE Trans. Nanotechnol.* **5** (2006) 613–627.
[30] B. Kosko, *Noise* (Penguin, 2006).
[31] M. McDonnell, N. Stocks, C. Pearce and D. Abbott, *Stochastic Resonance*: *From Suprathreshold Stochastic Resonance to Stochastic Signal Quantization* (Cambridge University Press, 2008).
[32] A. Patel and B. Kosko, Optimal mean-square noise benefits in quantizer-array linear estimation, *IEEE Signal Process. Lett.* **17** (2010) 1005–1009.
[33] A. Patel and B. Kosko, Noise benefits in quantizer-array correlation detection and watermark decoding, *IEEE Trans. on Signal Process.* **59** (2011) 488–505.
[34] H. Chen, L. R. Varshney and P. K. Varshney, Noise-enhanced information systems, *Proc. IEEE* **102** (2014) 1607–1621.
[35] S. Mitaim and B. Kosko, Noise-benefit forbidden-interval theorems for threshold signal detectors based on cross correlations, *Phys. Rev. E* **90** (2014) 052124.
[36] B. Franzke and B. Kosko, Noise can speed convergence in markov chains, *Phys. Rev. E* **84** (2011) 041112.
[37] B. Franzke and B. Kosko, Using noise to speed up markov chain monte carlo estimation, *Procedia Comput. Sci.* **53** (2015) 113–120.
[38] G. B. Folland, *Real Analysis*: *Modern Techniques and Their Applications*, 2nd edn. (Wiley-Interscience, 1999).

[39] R. A. Redner and H. F. Walker, Mixture densities, maximum likelihood and the EM algorithm, *SIAM Rev.* **26** (1984) 195–239.
[40] G. J. McLachlan and D. Peel, *Finite Mixture Models*, Wiley Series in Probability and Statistics: Applied Probability and Statistics (Wiley, 2004).
[41] R. V. Hogg and E. A. Tanis, *Probability and Statistical Inference* (Prentice Hall, 2006), 7th edition.
[42] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, 2nd edn. (Wiley-Interscience, 2001).
[43] C. Bishop, *Pattern Recognition and Machine Learning*, Information Science and Statistics (Springer, 2006).
[44] N. A. Gershenfeld, *The Nature of Mathematical Modeling* (Cambridge University Press, 1999).
[45] S. Kirkpatrick, C. Gelatt Jr and M. Vecchi, Optimization by simulated annealing, *Science* **220** (1983) 671–680.
[46] V. Černỳ, Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm, *J. Optim. Theory Appl.* **45** (1985) 41–51.
[47] S. Geman and C. Hwang, Diffusions for global optimization, *SIAM J. Control Optim.* **24** (1986) 1031–1043.
[48] B. Hajek, Cooling schedules for optimal annealing, *Math. Operat. Res.* **13** (1988) 311–329.
[49] B. Kosko, *Neural Networks and Fuzzy Systems*: *A Dynamical Systems Approach to Machine Intelligence* (Prentice Hall, 1991).