The Expectation-Maximization and Alternating Minimization Algorithms

Shane M. Haas

September 11, 2002

1 Summary

The Expectation-Maximization (EM) algorithm is a hill-climbing approach to finding a local maximum of a likelihood function [7, 8]. The EM algorithm alternates between finding a greatest lower bound to the likelihood function (the "E Step"), and then maximizing this bound (the "M Step"). The EM algorithm belongs to a broader class of alternating minimization algorithms [6], which includes the Arimoto-Blahut algorithm for calculating channel capacity and rate distortion functions [1, 3], and Cover's portfolio algorithm to maximize expected log-investment [4].

2 The Expectation-Maximization (EM) Algorithm

The primary purpose of this report is to introduce the EM algorithm and examine its relationship to other alternating minimization algorithms. For good tutorials on the EM algorithm, see [10, 2]. Roweis has a good review of linear Gaussian models, using the EM algorithm to estimate the model parameters [13]. This paper also provides pseudo-code for many popular EM applications. The book [9] is a more complete reference on the EM algorithm.

The basic problem of maximum likelihood estimation is to find the parameter θ that maximizes the likelihood function $L(\theta) \equiv f\,(\,y \mid \theta\,)$ for a given observation y. Because the logarithm is an increasing function, an equivalent problem is to find the parameter θ that maximizes the log-likelihood, i.e.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \;\; \log f\,(\,y \mid \theta\,). \tag{1}$$

If the likelihood (or log-likelihood) function is sufficiently well-behaved, then we can sometimes calculate a closed-form solution for the maximum likelihood estimator. More often than not, however, a closed form solution does not exist, and we must find the maximum using numerical optimization techniques, such as gradient-based ascent or the Newton-Raphson method. The EM algorithm

is one such hill-climbing algorithm that converges to a local maximum of the likelihood surface.

As the name suggests, the EM algorithm alternates between an expectation and a maximization step. The "E step" finds a lower bound that is equal to the log-likelihood function at the current parameter estimate θ_k. The "M step" generates the next estimate θ_{k+1} as the parameter that maximizes this greatest lower bound. This alternating process is shown pictorially in Fig. 1.

The EM algorithm, therefore, is a "divide and conquer" approach that breaks the original optimization problem into two hopefully easier problems. It then alternates between solving each easier optimization problem, using the solution of one to solve the other.

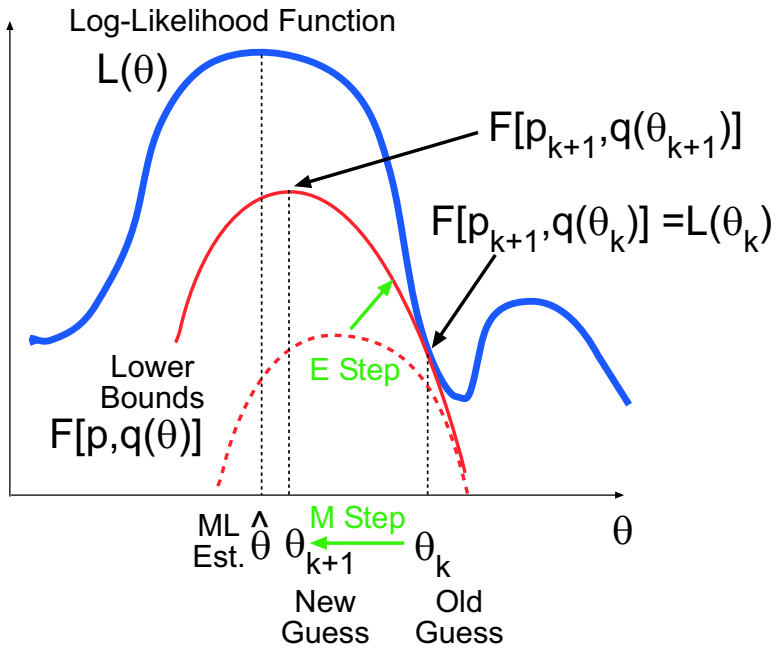

Figure 1: The EM algorithm alternates between finding a greatest lower bound ("E step"), and maximizing this bound ("M step").

The set of log-likelihood lower bounds comes from introducing a hidden or unobserved random variable x that has a joint density with the observation, $q(x, y \mid \theta)$. The likelihood function is then $f(y \mid \theta) = \int q(x, y \mid \theta)\, dx$. This hidden variable often has a natural physical meaning such as the hidden state of a linear dynamical system [14] or hidden Markov model [12], and can simplify the likelihood expression.

The log-likelihood lower bounds are parameterized by an arbitrary probability density $p(x)$ not necessarily equal to $\int q(x, y \mid \theta)\, dy$ for this hidden variable. The lower bounds come from applying Jensen's inequality:

$$
\begin{aligned}
L(\theta) &\equiv \log f\,(\,y \mid \theta\,) \\
&= \log \int q\,(\,x, y \mid \theta\,) dx \\
&= \log \int p\,(\,x\,) \frac{q\,(\,x, y \mid \theta\,)}{p\,(\,x\,)} dx \\
&\geq \int p\,(\,x\,) \log \left[\frac{q\,(\,x, y \mid \theta\,)}{p\,(\,x\,)} \right] dx \qquad (2) \\
&\equiv F\,[\,p,\, q(\theta)\,].
\end{aligned}
$$

As seen in Figure 1, the E step finds the density $p\,(\,x\,)$ that maximizes this lower bound for the current estimate θ_k. The M step then finds the value of θ that maximizes the lower bound generated by this density. This alternating process is summarized below:

$$\textbf{E Step: } p_{k+1} = \operatorname*{argmax}_{p} F\,[\,p,\, q(\theta_k)\,], \qquad (3)$$

$$\textbf{M Step: } \theta_{k+1} = \operatorname*{argmax}_{\theta} F\,[\,p_{k+1},\, q(\theta)\,]. \qquad (4)$$

2.1 The E Step

We will now focus on finding the probability density $p\,(\,x\,)$ that maximizes the lower bound $F\,[\,p,\, q(\theta_k)\,]$ while holding θ_k fixed. But first, we will examine three interpretations of what this lower bound family represents:

Free Energy: One interesting interpretation is that the lower bounds are the negative of a quantity known in statistical physics as free energy [11]. Defining energy as $-\log q\,(\,x, y \mid \theta\,)$, the free energy for a given y is the average energy with respect to $p\,(\,x\,)$ minus the entropy of $p\,(\,x\,)$, i.e.

$$-F\,[\,p,\, q(\theta)\,] = \underbrace{-E \log q\,(\,x, y \mid \theta\,)}_{\text{Avg. Energy}} - \underbrace{[-E \log p\,(\,x\,)]}_{\text{Entropy}}, \qquad (5)$$

where the expectations are with respect to $p\,(\,x\,)$. The E step, therefore, chooses $p\,(\,x\,)$ to minimize the free energy for the current parameter estimate θ_k. For a fixed $p\,(\,x\,)$, the entropy term does not depend on the parameter θ. Consequently, the M step minimizes that average energy with respect to the parameter θ, holding constant the density found in the E step.

KL Divergence: Another interpretation of this lower bound family is in terms of the Kullback-Leibler (KL) informational divergence between $p\,(\,x\,)$ and $q\,(\,x, y \mid \theta\,)$ for a given y, i.e.

$$F\,[\,p,\, q(\theta)\,] = -D[\,p \parallel q(\theta)\,], \qquad (6)$$

where

$$D[\ p\ \|\ q(\theta)\ ] \equiv \int p\,(\,x\,) \log \left[\frac{p\,(\,x\,)}{q\,(\,x, y \mid \theta\,)} \right] dx. \tag{7}$$

Notice that for a fixed y, $q\,(\,x, y \mid \theta\,)$ does not integrate to unity over x, and hence, as defined, this divergence might be negative.

We now have another interpretation of the E step as choosing a conditional density on x, $p\,(\,x\,)$, that minimizes the KL informational divergence between this density and the joint density of x and y, $q\,(\,x, y \mid \theta\,)$, for a fixed realization of y.

This interpretation will be important when relating the EM algorithm to other alternating minimization algorithms in [6]. The maximum-likelihood problem can be viewed as an alternating minimization problem, i.e.

$$\hat{\theta} = \operatorname*{argmin}_{\theta} \min_{p}\ D[\ p\ \|\ q(\theta)\ ], \tag{8}$$

where the E step performs the minimization over p for a fixed value of θ and the M step minimizes over θ for a fixed value of p.

KL Divergence: The third interpretation of the lower bound is also in terms of a divergence, and gives more insight into its relationship to the log-likelihood and to its maximization. Let $w\,(\,x \mid y, \theta\,)$ be the conditional probability function implied by $q\,(\,x, y \mid \theta\,)$, i.e.

$$w\,(\,x \mid y, \theta\,) = \frac{q\,(\,x, y \mid \theta\,)}{f\,(\,y \mid \theta\,)}, \tag{9}$$

where $f\,(\,y \mid \theta\,) = \int q\,(\,x, y \mid \theta\,)\ dx$. We can then express the lower bound as

$$\begin{aligned} F\,[\,p,\, q(\theta)\,] &= \int p\,(\,x\,) \log \left[\frac{w\,(\,x \mid y, \theta\,) f\,(\,y \mid \theta\,)}{p\,(\,x\,)} \right] dx \\ &= \log f\,(\,y \mid \theta\,) - D[\ p\ \|\ w(\theta)\ ]. \end{aligned} \tag{10}$$

Notice that $w\,(\,x \mid y, \theta\,)$ does integrate to unity over x, and hence

$$D[\ p\ \|\ w(\theta)\ ] \geq 0,$$

with equality when $p\,(\,x\,) \equiv w\,(\,x \mid y, \theta\,)$.

We now see that the lower bound is actually the log-likelihood minus the divergence between the arbitrarily chosen density, $p\,(\,x\,)$, and the actual conditional density, $w\,(\,x \mid y, \theta\,)$. Choosing $p\,(\,x\,) \equiv w\,(\,x \mid y, \theta\,)$, therefore, maximizes the lower bound for a fixed value of θ. Furthermore, this choice makes the lower bound equal to the log-likelihood at this particular value of θ as illustrated in Fig. 1.

Based on this third interpretation, the E step is

$$\textbf{E Step: } p_{k+1} = \underset{p}{\operatorname{argmax}}\ F\,[\,p,\, q(\theta_k)\,] = w\,(\,x \mid y, \theta_k\,), \tag{11}$$

or using Baye's Rule

$$\textbf{E Step: } p_{k+1} = w\,(\,x \mid y, \theta_k\,) = \frac{h(\,y \mid x, \theta_k\,)\pi(\,x \mid \theta_k\,)}{\int h(\,y \mid \chi, \theta_k\,)\pi(\,\chi \mid \theta_k\,)d\chi}, \tag{12}$$

where $h(\,y \mid x, \theta\,) = q\,(\,x, y \mid \theta\,)\pi(\,x \mid \theta\,)$ and $\pi(\,x \mid \theta\,) = \int q\,(\,x, y \mid \theta\,)\ dy$. This step of the EM algorithm is called the expectation step because the lower bound that it maximizes is expressed in terms of conditional expectations as seen in (5).

2.2 The M Step

The M step finds the value of θ that maximizes the greatest lower bound $F\,[\,p_{k+1},\, q(\theta)\,]$ produced by the E step. Evaluating the lower bound at this maximizing distribution, i.e. $p_{k+1}(x) = w\,(\,x \mid y, \theta_k\,)$, results in

$$\begin{aligned} F\,[\,p_{k+1},\, q(\theta)\,] &= \int w\,(\,x \mid y, \theta_k\,) \log \left[\frac{q\,(\,x, y \mid \theta\,)}{w\,(\,x \mid y, \theta_k\,)}\right] dx \\ &= \int w\,(\,x \mid y, \theta_k\,) \log q\,(\,x, y \mid \theta\,)\ dx \\ &\quad - \int w\,(\,x \mid y, \theta_k\,) \log w\,(\,x \mid y, \theta_k\,)\ dx. \end{aligned} \tag{13}$$

Because the second term does not depend on θ, the M step becomes

$$\textbf{M Step: } \theta_{k+1} = \underset{\theta}{\operatorname{argmax}} \int w\,(\,x \mid y, \theta_k\,) \log q\,(\,x, y \mid \theta\,)\ dx. \tag{14}$$

Notice that the M-step maximizes the conditional expectation of the log-joint observation and the hidden variable density, $q\,(\,x, y \mid \theta\,)$. The unknown parameter in the conditional density $w\,(\,x \mid y, \theta\,)$ is fixed to its previous estimate, θ_k, and does not vary in the maximization.

Unlike the E step, the M step is problem specific. In other words, maximizing $F\,[\,p_{k+1},\, q(\theta)\,]$ will depend on the structure of the problem under consideration.

3 Example: Mixture Problem

To illustrate the EM algorithm we will examine the mixture problem from the introductory chapter of [9] and the applications section of [6]. This mixture problem assumes that the observed data $y = \{y_1, \ldots, y_N\}$ is generated in the following manner. For each sample, a group g, $1 \leq g \leq G$, is randomly chosen with unknown probability π_g. The observed sample y_n is then randomly

generated according to a known probability density, $h_g(y_n)$, for group g. Furthermore, each sample is generated independently. The maximum-likelihood problem is to find the group probabilities, $\theta = \{\pi_1, \ldots, \pi_{G-1}\}$, that maximize the log-likelihood function

$$L(\theta) = \sum_{n=1}^{N} \log \left\{ \sum_{g=1}^{G} h_g(y_n)\pi_g \right\}. \tag{15}$$

The constraint $\sum_{g=1}^{G} \pi_g = 1$ is enforced by defining $\pi_G = 1 - \sum_{g=1}^{G-1} \pi_g$, and only estimating $\theta = \{\pi_1, \ldots, \pi_{G-1}\}$. For simplicity, we will not enforce the constraints that each π_g be non-negative. This assumption does not hurt us if the components of the maximizing parameter are non-negative. Denote the set of all valid parameters as Θ. Elements of Θ have $G-1$ real components whose sum is less than or equal to one.

Notice that for a given y, this log-likelihood function is concave with respect to the parameters because it is the sum of concave functions. An example of a log-likelihood realization is shown in Figure 2 for five observations ($N = 5$) and two groups ($G = 2$).

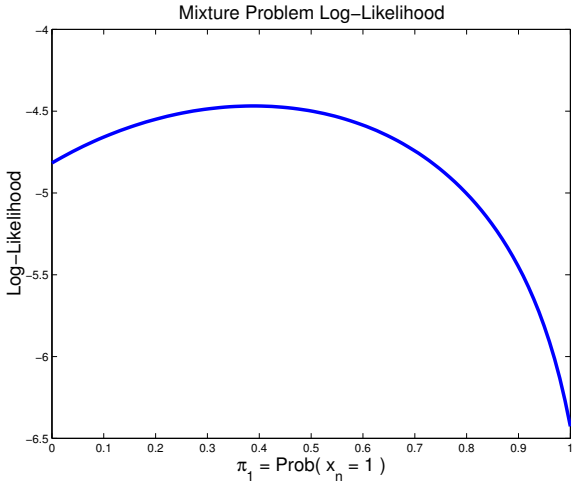

Figure 2: An example of the log-likelihood surface for a mixture of two groups ($G = 2$) with five observations ($N = 5$)

Even though the log-likelihood is concave, a closed form solution for its maximum does not exist. We can use the EM algorithm, however, to iteratively find the maximum.

To use the EM algorithm we will introduce the unobserved or hidden variable x_n that takes a value g, if group g produced the sample y_n. The joint density

of $x = \{x_1, \ldots, x_N\}$ and $y = \{y_1, \ldots, y_N\}$ for a given θ is, therefore,

$$q\,(\,x, y \mid \theta\,) = \prod_{n=1}^{N} h_{x_n}(y_n)\pi_{x_n}. \tag{16}$$

The EM algorithm proceeds as follows. The E step (11) produces the conditional density (9) based on the current parameter estimate, i.e.

$$\textbf{E Step: } p_{k+1} = w\,(\,x \mid y, \theta_k\,) = \prod_{n=1}^{N} \Pr\{\,x_n \mid y_n, \theta_k\}, \tag{17}$$

where Baye's rule gives

$$\Pr\{\,x_n = g \mid y_n, \theta_k\} = \frac{h_g(y_n)\pi_g^k}{\sum_{\gamma=1}^{G} h_\gamma(y_n)\pi_\gamma^k} \equiv m_n^k(g), \tag{18}$$

and π_g^k denotes the estimate of the group g probability from the current parameter estimate θ_k.

The M step (14) then finds the parameter θ that maximizes

$$E[\,\log q\,(\,x, y \mid \theta\,)\,] = \sum_{n=1}^{N} \left(E[\,\log h_{x_n}(y_n)] + E[\,\log \pi_{x_n}]\right), \tag{19}$$

where the expectation is with respect to $p_{k+1}(x) = w\,(\,x \mid y, \theta_k\,)$. We can ignore the first term because it does not depend on θ. The expectation in the second term is

$$\begin{aligned} E[\,\log \pi_{x_n}] &= \sum_{x} w\,(\,x \mid y, \theta_k\,) \log \pi_{x_n} \\ &= \sum_{g=1}^{G} \sum_{x: x_n = g} w\,(\,x \mid y, \theta_k\,) \log \pi_g \\ &= \sum_{g=1}^{G} \Pr\{\,x_n = g \mid y_n, \theta_k\} \log \pi_g \\ &= \sum_{g=1}^{G} m_n^k(g) \log \pi_g. \end{aligned} \tag{20}$$

This result can also be seen as an iterated expectation,

$$E[\,\log \pi_{x_n}] = E[\,E[\,\log \pi_g \mid x_n = g, y, \theta_k]\,] = E[\,\log \pi_g],$$

where this last expectation is with respect to $\Pr\{\,x_n = g \mid y_n, \theta_k\} \equiv m_n^k(g)$.

Differentiating $E[\,\log q\,(\,x, y \mid \theta\,)\,]$ with respect to π_g, gives the necessary condition for a fixed point:

$$\frac{\partial E[\,\log q\,(\,x, y \mid \theta\,)\,]}{\partial \pi_g} = \sum_{n=1}^{N} \frac{m_n^k(g)}{\pi_g^{k+1}} - \sum_{n=1}^{N} \frac{m_n^k(G)}{\pi_G^{k+1}} = 0, \qquad 1 \le g \le G-1. \tag{21}$$

This condition implies that

$$\textbf{M Step: } \pi_g^{k+1} = \frac{\sum_{n=1}^N m_n^k(g)}{\sum_{n=1}^N \sum_{g=1}^G m_n^k(g)} = \frac{1}{N}\sum_{n=1}^N m_n^k(g), \tag{22}$$

maximizes the log-likelihood conditional expectation.

The EM algorithm for this example has a very appealing interpretation. Had we observed x, then the maximum likelihood estimate of the group probability π_g would be the number of times x_n equals g divided by the total number of observations. The EM algorithm follows a very similar procedure. The next estimate of the group probability π_g^{k+1} is the *expected* number of times that x_n equals g conditioned on y and the previous estimate θ_k divided by the total number of observations, i.e.

$$\pi_g^{k+1} = (\text{ Avg. \# of times } x_n = g \text{ given } y \text{ and } \theta_k)/N. \tag{23}$$

4 Alternating Minimization

As mentioned previously, the EM algorithm belongs to class of alternating minimization procedures that have a nice geometric interpretation. We will now summarize the main results of [6], and then show their relationship to the EM algorithm.

Let P and Q be two elements from arbitrary sets $\mathcal{P}$ and $\mathcal{Q}$, respectively. Define an arbitrary "distance" function[1] $d(P,Q)$ that maps elements of $\mathcal{P}$ and $\mathcal{Q}$ to the extended real numbers.

We say that the sequences $\{P_k\}_{k=0}^\infty$ and $\{Q_k\}_{k=0}^\infty$ are obtained by alternating minimization if for $k = 0, 1, 2, \ldots$

$$P_{k+1} = \operatorname*{argmin}_{P\in\mathcal{P}} \; d(P, Q_k), \tag{24}$$

$$Q_{k+1} = \operatorname*{argmin}_{Q\in\mathcal{Q}} \; d(P_{k+1}, Q), \tag{25}$$

with the iterations starting at $Q_0 = \operatorname{argmin}_{Q\in\mathcal{Q}} \; d(P_0, Q)$, and the starting point P_0 arbitrary. We can describe an alternating minimization sequence using the notation $P_0 \to Q_0 \to P_1 \to Q_1 \to \cdots$.

The main theorem (Th. 3) of [6] proves that if $\mathcal{P}$ and $\mathcal{Q}$ are convex measures (not necessarily probability measures) and $d(P,Q) = D(P\|Q)$ is the KL informational divergence, then all alternating minimization divergences converge. Furthermore, they converge monotonically to a global minimum.

The proof of the theorems in [6] are very general, developing the geometric properties of $\mathcal{P}$,$\mathcal{Q}$, and d necessary for convergence. The KL informational divergence happens to satisfy these properties over convex sets of measures. Another example that satisfies the geometric properties necessary for convergence is that of closed, convex sets from a Hilbert space with d as the induced norm. This example of projection onto convex sets is illustrated in Fig. 3.

[1] The function d is not a true "distance" because it can be negative and asymmetric. Yet, it is intuitive to still think of it as measuring the "distance" between elements of $\mathcal{P}$ and $\mathcal{Q}$.

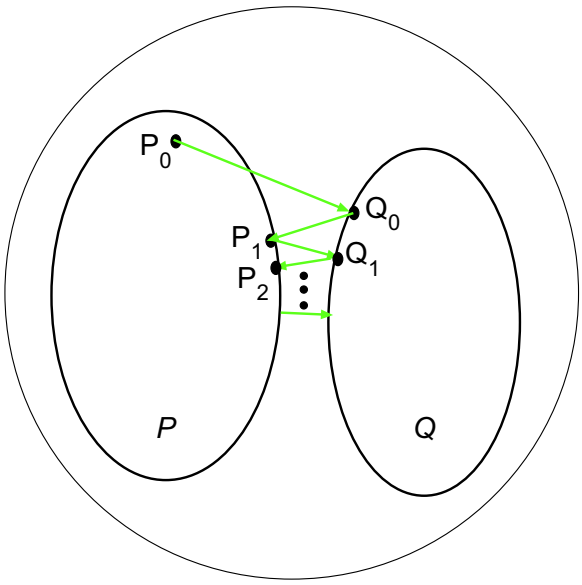

Figure 3: Alternating minimization iterates between finding the minimum of $d(P, Q)$ holding Q fixed, and the minimum holding P fixed, i.e. $P_0 \to Q_0 \to P_1 \to Q_1 \to \cdots$.

5 Relationship Between Alternating Minimization Algorithms

We will now examine the relationship between the EM, Arimoto-Blahut, and a best constant rebalanced portfolio algorithm as alternating minimization procedures. As mentioned previously, the EM is an alternating minimization algorithm that minimizes $D[\, p \,\|\, q(\theta)\,]$. *If* $q(\theta)$ forms a convex set, then the EM algorithm will converge to a global minimum.

The algorithm to find the best constant rebalanced portfolio for a sequence of stock returns is just the mixture problem with $h_g(y_n)$ replaced by the return of stock g at time n, and the group probabilities with the portfolio weights. This best constant balanced portfolio is the target portfolio in Cover's universal portfolio algorithm [5].

The Arimoto-Blahut algorithm to calculate discrete memoryless channel capacity minimizes $D[\, p(\theta) \,\|\, q(\phi)\,]$, where $p(\theta) = h(\, y \mid x\,)\theta(x)$ and $q(\phi) = h(\, y \mid x\,)\phi(\, x \mid y\,)$. Here, $\theta(x)$ is the density over the channel inputs, $h(\, y \mid x\,)$ is the channel matrix relating inputs x to outputs y, and $\phi(\, x \mid y\,)$ is an arbitrary stochastic matrix. Because $p(\theta)$ and $q(\phi)$ form convex sets parameterized by θ and ϕ, respectively, the Arimoto-Blahut algorithm converges to a global minimum.

6 Conclusions

The EM algorithm is an alternating minimization algorithm that iterates between finding a greatest lower bound to the log-likelihood function and maximizing this bound. The EM algorithm converges when the observed and hidden variable joint density form a convex set over the allowable parameters. Other alternating minimization algorithms include the Arimoto-Blahut algorithm for calculating channel capacity and rate-distortion functions, and some optimal portfolio algorithms.

References

[1] S. Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Trans. on Inform. Theory*, 18(1):14–20, January 1972.

[2] J.A. Bilmes. A gentle tutorial of the EM algorithm and its applications to parameter estimation for gaussian mixture and hidden markov models. Technical Report TR-97-021, International Computer Science Institute, 1947 Center St., Berkeley, CA 94704-1198, April 1998.

[3] R. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Trans. on Inform. Theory*, 18(4):460–473, July 1972.

[4] T.M. Cover. An algorithm for maximizing expected log investment return. *IEEE Transactions on Information Theory*, 30(2):369–373, March 1984.

[5] T.M. Cover. Universal portfolios. *Mathematical Finance*, 1(1):1–29, January 1991.

[6] I. Csiszar and G. Tusnady. Information geometry and alternating minimization procedures. *Statistics and Decisions*, 1:205–237, 1984. Supplement Issue.

[7] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximimum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39(1):1–38, 1977.

[8] H. Hartley. Maximum likelihood estimation from incomplete data. *Biometrica*, 14:174–194, 1958.

[9] G.J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley series in probability and statstics. John Wiley & Sons, 1997.

[10] T.P. Minka. Expectation-maximization as lower bound maximization. Technical report, M.I.T., November 1998. http://www.stat.cmu.edu/ minka/papers/learning.html.

[11] R.M. Neal and G.E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M.I. Jordan, editor, *Learning in Graphical Models*, pages 355–368. Kluwer, Dordrecht, MA, 1998.

[12] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257–286, February 1989.

[13] S. Roweis and Z. Ghahramani. A unifying review of linear Gaussian models. *Neural Computation*, 11:305–345, 1999.

[14] R.H. Shumway and D.S. Stoffer. An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis*, 3(4):253–264, 1982.