e-Appendix C

The E-M Algorithm

The E-M algorithm is a convenient heuristic for likelihood maximization. The E-M algorithm will never decrease the likelihood. Our discussion will focus on mixture models, the GMM being a special case, even though the E-M algorithm applies to more general settings.

Let $P_k(\mathbf{x};\theta_k)$ be a density for $k = 1,\ldots,K$, where θ_k are the parameters specifying P_k. We will refer to each P_k as a bump. In the GMM setting, all the P_k are Gaussians, and $\theta_k = \{\boldsymbol{\mu}_k, \Sigma_k\}$ (the mean vector and covariance matrix for each bump). A mixture model is a weighted sum of these K bumps,

$$P(\mathbf{x};\Theta) = \sum_{k=1}^{K} w_k P_k(\mathbf{x};\theta_k),$$

where the weights satisfy $w_k \geq 0$ and $\sum_{k=1}^{K} w_k = 1$ and we have collected all the parameters into a single grand parameter, $\Theta = \{w_1,\ldots,w_K;\theta_1,\ldots,\theta_K\}$. Intuitively, to generate a random point $\mathbf{x}$, you first pick a bump according to the probabilities $w_1,\ldots,w_K$. Suppose you pick bump k. You then generate a random point from the bump density P_k.

Given data $X = \mathbf{x}_1,\ldots,\mathbf{x}_N$ generated independently, we wish to estimate the parameters of the mixture which maximize the log-likelihood,

$$\begin{aligned}
\ln P(X|\Theta) &= \ln \prod_{n=1}^{N} P(\mathbf{x}_n|\Theta) \\
&= \ln \prod_{n=1}^{N} \left(\sum_{k=1}^{K} w_k P_k(\mathbf{x}_n;\theta_k) \right) \\
&= \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} w_k P(\mathbf{x}_n|\theta_k) \right).
\end{aligned} \tag{C.1}$$

In the first step above, $P(X|\Theta)$ is a product because the data are independent. Note that X is known and fixed. What is not known is which particular bump was used to generate data point $\mathbf{x}_n$. Denote by $j_n \in \{1, \ldots, K\}$ the bump that generated $\mathbf{x}_n$ (we say $\mathbf{x}_n$ is a 'member' of bump j_n). Collect all bump memberships into a set $J = \{j_1, \ldots, j_N\}$. If we knew which data belonged to which bump, we can estimate each bump density's parameters separately, using only the data belonging to that bump. We call (X, J) the *complete data*. If we know the complete data, we can easily optimize the log-likelihood. We call X the *incomplete data*. Though X is all we can measure, it is still called the 'incomplete' data because it does not contain enough information to easily determine the optimal parameters Θ^* which minimize $E_{\text{in}}(\Theta)$. Let's see mathematically how knowing the complete data helps us.

To get the likelihood of the complete data, we need the joint probability $\mathbb{P}[\mathbf{x}_n, j_n|\Theta]$. Using Bayes' theorem,

$$\begin{aligned}
\mathbb{P}[\mathbf{x}_n, j_n|\Theta] &= \mathbb{P}[j_n|\Theta]\,\mathbb{P}\left[\mathbf{x}_n|j_n, \Theta\right] \\
&= w_{j_n} P_{j_n}(\mathbf{x}_n; \theta_{j_n}).
\end{aligned}$$

Since the data are independent,

$$\begin{aligned}
P(X, J|\Theta) &= \prod_{n=1}^{N} \mathbb{P}[\mathbf{x}_n, j_n|\Theta] \\
&= \prod_{n=1}^{N} w_{j_n} P_{j_n}(\mathbf{x}_n; \theta_{j_n}).
\end{aligned}$$

Let N_k be the number of occurrences of bump k in J, and let X_k be those data points corresponding to the bump k, so $X_k = \{\mathbf{x}_n \in X : j_n = k\}$. We compute the log-likelihood for the complete data as follows:

$$\begin{aligned}
\ln P(X, J|\Theta) &= \sum_{n=1}^{N} \ln w_{j_n} + \sum_{n=1}^{N} \ln P_{j_n}(\mathbf{x}_n; \theta_{j_n}) \\
&= \sum_{k=1}^{K} N_k \ln w_k + \sum_{k=1}^{K} \underbrace{\sum_{\mathbf{x}_n \in X_k} \ln P_k(\mathbf{x}_n; \theta_k)}_{L_k(X_k, \theta_k)} \\
&= \sum_{k=1}^{K} N_k \ln w_k + \sum_{k=1}^{K} L_k(X_k; \theta_k).
\end{aligned} \tag{C.2}$$

There are two simplifications which occur in (C.2) from knowing the complete data (X, J). The w_k (in the first term) are separated from the θ_k (in the second term); and, the second term is the sum of K non-interacting log-likelihoods $L_k(X_k, \theta_k)$ corresponding to the data belonging to X_k and only involving bump k's parameters θ_k. Each log-likelihood L_k can be optimized independently of the others. For many choices of P_k, $L_k(X_k; \theta_k)$ can be optimized analytically, even though the log-likelihood for the incomplete data in (C.1) is intractable. The next exercise asks you to analytically maximize (C.2) for the GMM.

Exercise C.1

(a) Maximize the first term in (C.2) subject to $\sum_k w_k = 1$, and show that the optimal weights are $w_k^* = N_k/N$. *[Hint: Lagrange multipliers.]*

(b) For the GMM,

$$P_k(\mathbf{x}; \boldsymbol{\mu}_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^{\mathrm{T}}\Sigma_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right).$$

Maximize $L_k(X_k; \boldsymbol{\mu}_k, \Sigma_k)$ to obtain the optimal parameters:

$$\begin{aligned} \boldsymbol{\mu}_k^* &= \frac{1}{N_k}\sum_{\mathbf{x}_n \in X_k} \mathbf{x}_n; \\ \Sigma_k^* &= \frac{1}{N_k}\sum_{\mathbf{x}_n \in X_k} (\mathbf{x}_n-\boldsymbol{\mu}_k)(\mathbf{x}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}. \end{aligned}$$

These are exactly the parameters you would expect. $\boldsymbol{\mu}_k^*$ is the in-sample mean for the data belonging to bump k; similarly, Σ_k^* is the in-sample covariance matrix.
[Hint: Set $\mathrm{S_k} = \Sigma_k^{-1}$ and optimize with respect to $\mathrm{S_k}$. Also, from the Linear Algebra e-appendix, you may find these derivatives useful:

$$\frac{\partial}{\partial \mathrm{S}}(\mathbf{z}^{\mathrm{T}}\mathrm{S}\mathbf{z}) = \mathbf{z}\mathbf{z}^{\mathrm{T}}; \quad \textit{and} \quad \frac{\partial}{\partial \mathrm{S}} \ln|\mathrm{S}| = \mathrm{S}^{-1}.\textit{]}$$

In reality, we do not have access to J, and hence it is called a 'hidden variable'. So what can we do now? We need a heuristic to maximize the likelihood in Equation (C.1). One approach is to guess J and maximize the resulting complete likelihood in Equation (C.2). This almost works. Instead of maximizing the complete likelihood for a *single* guess, we consider an average of the complete likelihood over all possible guesses. Specifically, we treat J as an unknown random variable and maximize the expected value (with respect to J) of the complete log-likelihood in Equation (C.2). This expected value is as easy to minimize as the complete likelihood. The mathematical implementation of this idea will lead us to the E-M Algorithm, which stands for Expectation-Maximization Algorithm. Let's start with a simpler example.

Example C.1. You have two opaque bags. Bag 1 has red and green balls, with μ_1 being the fraction of red balls. Bag 2 has red and blue balls with μ_2 being the fraction of red. You pick four balls in independent trials as follows. First pick one of the bags at random, each with probability $\frac{1}{2}$; then, pick a ball at random from the bag. Here is the sample of four balls you got: , one green, one red and two blue. The task is to estimate μ_1 and μ_2. It would be much easier if we knew which bag each ball came from.

Here is one way to reason. Half the balls will come from Bag 1 and the other half from Bag 2. The blue balls come from Bag 2 (that's already Bag 2's budget of balls), so the other two should come from Bag 1: | . Using in-sample estimates, $\hat{\mu}_1 = \frac{1}{2}$ and $\hat{\mu}_2 = 0$. We can get these same estimates using a maximum likelihood argument. The log-likelihood of the data is

$$\ln(1-\mu_1) + 2\ln(1-\mu_2) + \ln(\mu_1+\mu_2) - 4\ln 2. \tag{C.3}$$

The reader should explicitly maximize the above expression with respect to $\mu_1, \mu_2 \in [0,1]$ to obtain the estimates $\hat{\mu}_1 = \frac{1}{2}, \hat{\mu}_2 = 0$. In our data set of four balls, there is a red one, so it seems a little counter-intuitive that we would estimate $\hat{\mu}_2 = 0$, for isn't there a *positive* probability that the red ball came from Bag 2? Nevertheless, $\hat{\mu}_2 = 0$ *is* the estimate that *maximizes* the probability of generating the data. You are uneasy with this, and rightly so, because we put all our eggs into this single 'point' estimate; a very unnatural thing given that any point estimate has infinitesimal probability of being correct. Nevertheless, that is the maximum likelihood method, and we are following it.

Here is another way to reason. 'Half' of each red ball came from Bag 1 and the other 'half' from Bag 2. So, $\hat{\mu}_1 = \frac{1}{2}/(1+\frac{1}{2}) = \frac{1}{3}$ because $\frac{1}{2}$ a red ball came from Bag 1 out of a total of $1+\frac{1}{2}$ balls. Similarly, $\hat{\mu}_2 = \frac{1}{2}/(2+\frac{1}{2}) = \frac{1}{5}$. This reasoning is wrong because it does not correctly use the knowledge that the ball is red. For example, as we just reasoned, $\hat{\mu}_1 = \frac{1}{3}$ and $\hat{\mu}_2 = \frac{1}{5}$. But, if these estimates are correct, and indeed $\hat{\mu}_1 > \hat{\mu}_2$, then a red ball is more likely to come from Bag 1, so more than half if it should come from Bag 1. This contradicts the original assumption that led us to these estimates.

Now, let's see how expectation-maximization solves this problem. The reasoning is similar to our false start above; it just adds iteration till consistency. We begin by considering the two cases for the red ball. Either it is from Bag 1 or Bag 2. We can compute the log-likelihood for each of these two cases:

$$\begin{aligned} &\ln(1-\mu_1) + \ln(\mu_1) + 2\ln(1-\mu_2) - 4\ln 2 && \text{(Bag 1);} \\ &\ln(1-\mu_1) + \ln(\mu_2) + 2\ln(1-\mu_2) - 4\ln 2 && \text{(Bag 2).} \end{aligned}$$

Suppose we have estimates $\hat{\mu}_1$ and $\hat{\mu}_2$. Using Bayes theorem, we can compute $p_1 = \mathbb{P}[\text{Bag 1} \mid \hat{\mu}_1, \hat{\mu}_2]$ and $p_2 = \mathbb{P}[\text{Bag 2} \mid \hat{\mu}_1, \hat{\mu}_2]$. The reader can verify that

$$p_1 = \frac{\hat{\mu}_1}{\hat{\mu}_1 + \hat{\mu}_2}, \qquad p_2 = \frac{\hat{\mu}_2}{\hat{\mu}_1 + \hat{\mu}_2}.$$

Now comes the *expectation* step. Compute the expected log-likelihood using p_1 and p_2:

$$\mathbb{E}[\textsf{log-likelihood}] = \ln(1-\mu_1) + p_1\ln(\mu_1) + p_2\ln(\mu_2) + 2\ln(1-\mu_2) - 4\ln 2. \tag{C.4}$$

Next comes the *maximization* step. Treating p_1, p_2 as constants, maximize the expected log-likelihood with respect to μ_1, μ_2 and update $\hat{\mu}_1, \hat{\mu}_2$ to these optimal values. Notice that the log-likelihood in (C.3) has an interaction term $\ln(\mu_1+\mu_2)$ which complicates the maximization. In the expected log-likelihood (C.4), μ_1 and μ_2 are decoupled, and so the maximization can be implemented *separately* for each variable. The reader can verify that maximizing the expected log-likelihood gives the updates:

$$\hat{\mu}_1 \leftarrow \frac{p_1}{1+p_1} = \frac{\hat{\mu}_1}{2\hat{\mu}_1 + \hat{\mu}_2} \qquad \text{and} \qquad \hat{\mu}_2 \leftarrow \frac{p_2}{2+p_2} = \frac{\hat{\mu}_2}{2\hat{\mu}_1 + 3\hat{\mu}_2}.$$

The full algorithm just iterates this update process with the new estimates. Let's see what happens if we start (arbitrarily) with estimates $\hat{\mu}_1 = \hat{\mu}_2 = \frac{1}{2}$:

	Iteration number									
	0	1	2	3	4	5	6	7	...	1000
$\hat{\mu}_1$	$\frac{1}{2}$	$\frac{1}{3}$	0.38	0.41	0.43	0.45	0.45	0.46	...	0.49975
$\hat{\mu}_2$	$\frac{1}{2}$	$\frac{1}{5}$	0.16	0.13	0.10	0.09	0.07	0.07	...	0.0005

We have highlighted in blue the result of the first iteration, which is exactly the estimate from our earlier faulty reasoning. When $\mu_1 = \mu_2$ our faulty reasoning matches the E-M step. If we continued this table, it is not hard to see what will happen: $\hat{\mu}_1 \to \frac{1}{2}$ and $\hat{\mu}_2 \to 0$.

Exercise C.2

When the E-M algorithm converges, it must be that

$$\hat{\mu}_1 = \frac{\hat{\mu}_1}{2\hat{\mu}_1 + \hat{\mu}_2} \qquad \text{and} \qquad \hat{\mu}_2 = \frac{\hat{\mu}_2}{2\hat{\mu}_1 + 3\hat{\mu}_2}.$$

Solve these consistency conditions, and report your estimates for $\hat{\mu}_1, \hat{\mu}_2$?

It's miraculous that by maximizing an expected log-likelihood using a *guess* for the parameters, we end up converging to the true maximum likelihood solution. Why is this useful? Because the maximizations for μ_1 and μ_2 are decoupled. We trade a maximization of a complicated likelihood of the incomplete data for a bunch of simpler maximizations that we iterate. □

C.1 Derivation of the E-M Algorithm

We now derive the E-M strategy and show that it will always improve the likelihood of the incomplete data.

> *Maximizing an expected log-likelihood of the complete data increases the likelihood of the incomplete data.*

What is surprising is that to compute the expected log-likelihood, we use a guess, since we don't know the best model. So it is really a guess for the expected log-likelihood that one maximizes and this increases the likelihood.

Let Θ' be any set of parameters, and define $P(J|X, \Theta')$ as the conditional probability distribution for J given the data and *assuming* that Θ' is the actual probability model for the data. The probability $P(J|X, \Theta')$ is well defined, even if Θ' is not the probability model which generated the data. We will see how to compute $P(J|X, \Theta')$ soon, but for the moment assume that it is always positive (which means that every possible assignment of $\mathbf{x}_1, \ldots, \mathbf{x}_N$ to bumps $j_1, \ldots, j_n$ has non-zero probability under Θ'). This will always be the case unless some of the P_k have bounded support or Θ' is some degenerate mixture. The following derivation establishes a connection between the log-likelihood for the incomplete data and the expectation (over J) of the log-likelihood for the complete data.

$$\begin{aligned}
L(\Theta) &= \ln P(X|\Theta) \\
&\overset{(a)}{=} \ln \sum_J P(X, J|\Theta) \\
&\overset{(b)}{=} \ln \sum_J \frac{P(X, J|\Theta)}{P(J|X, \Theta')} P(J|X, \Theta') \\
&\overset{(c)}{=} \ln \sum_J \frac{P(X, J|\Theta) P(X|\Theta')}{P(X, J|, \Theta')} P(J|X, \Theta') \\
&\overset{(d)}{\geq} \sum_J \ln \left(\frac{P(X, J|\Theta) P(X|\Theta')}{P(X, J|, \Theta')} \right) P(J|X, \Theta') \\
&\overset{(e)}{=} L(\Theta') + \mathbb{E}_{J|X,\Theta'}[\ln P(X, J|\Theta)] - \mathbb{E}_{J|X,\Theta'}[\ln P(X, J|\Theta')] \\
&\overset{(f)}{=} L(\Theta') + Q(\Theta|X, \Theta') - Q(\Theta'|X, \Theta') \qquad \text{(C.5)}
\end{aligned}$$

(a) follows by the law of total probability; (b) is justified because $P(J|X, \Theta')$ is positive; (c) follows from Bayes' theorem; (d) follows because the summation $\sum_J (\cdot) P(J \mid X, \Theta')$ is an expectation using the probabilities $P(J \mid X, \Theta')$ and by Jensen's inequality $\ln \mathbb{E}[\cdot] \geq \mathbb{E}[\ln(\cdot)]$ because $\ln(x)$ is concave; (e) follows because $\ln P(X|\Theta')$ is independent of J and so

$$\sum_J P(J|X, \Theta') \ln P(X|\Theta') = \ln P(X|\Theta') \sum_J P(J|X, \Theta') = \ln P(X|\Theta') \cdot 1;$$

finally, in (f), we have defined the function

$$Q(\Theta|X, \Theta') = \mathbb{E}_{J|X,\Theta'}[\ln P(X, J|\Theta)].$$

The function Q is a function of Θ, though its definition depends the distribution of J which in turn depends on the incomplete data X and the model Θ'. We have proved the following result.

Theorem C.2. If $Q(\Theta|X, \Theta') > Q(\Theta'|X, \Theta')$, then $L(\Theta) > L(\Theta')$.

In words, fix Θ' and compute the 'posterior' distribution of J conditioned on the data X with parameters Θ'. Now for the parameters Θ, compute the *expected log-likelihood* of the complete data (X, J) where the expectation is taken with respect to this posterior distribution for J that we just obtained. This distribution for J is fixed, depending on X, Θ', but *it does not depend on* Θ. Find Θ^* that maximizes this expected log-likelihood, and you are guaranteed to improve the actual likelihood. This theorem leads naturally to the E-M algorithm that follows.

E-M Algorithm

1: Initialize Θ_0 at $t = 0$.
2: At step t let the parameter estimate be Θ_t.
3: **[Expectation]** For X, Θ_t, compute the function of $Q_t(\Theta)$:

$$Q_t(\Theta) = \mathbb{E}_{J|X,\Theta_t}[\ln P(X, J|\Theta)],$$

which is the expected log-likelihood for the complete data.
4: **[Maximization]** Update Θ to maximize $Q_t(\Theta)$:

$$\Theta_{t+1} = \underset{\Theta}{\operatorname{argmax}}\, Q_t(\Theta).$$

5: Increment $t \to t + 1$ and repeat steps 3,4 till convergence.

In the algorithm, we need to compute $Q_t(\Theta)$, which amounts to computing an expectation with respect to $P(J|X, \Theta_t)$. We illustrate this process with our mixture model.

Recall that J is the vector of bump memberships. Since the data are independent, to compute $P(J|X, \Theta_t)$ we can compute this 'posterior' for each data point and then take the product. We need $\gamma_{nk} = P(j_n = k|\mathbf{x}_n, \Theta_t)$, the probability that data point $\mathbf{x}_n$ came from bump k. By Bayes' theorem,

$$\begin{aligned}\gamma_{nk} = P(j_n = k|\mathbf{x}_n, \Theta_t) &= \frac{P(\mathbf{x}_n, k|\Theta_t)}{P(\mathbf{x}_n|\Theta_t)} \\ &= \frac{\hat{w}_k P_k(\mathbf{x}_n|\hat{\theta}_k)}{\sum_{\ell=1}^{K} \hat{w}_\ell P_\ell(\mathbf{x}_n|\hat{\theta}_\ell)},\end{aligned}$$

where $\Theta_t = \{\hat{w}_1, \ldots, \hat{w}_K; \hat{\theta}_1, \ldots, \hat{\theta}_K\}$ and $P(J|X, \Theta') = \prod_{n=1}^{N} \gamma_{nj_n}$. We can now compute $Q_t(\Theta)$,

$$Q_t(\Theta) = \mathbb{E}_J[\ln P(X, J|\Theta)],$$

where the expectation is with respect to the ('fictitious') probabilities γ_{nk} that determine the distribution of the random variable J. These probabilities depend on X and Θ_t. Let $\hat{N}_k$ (a random variable) be the number of occurrences of bump k in the random variable J; similarly, let $\hat{X}_k$ be the random set containing the data points of bump k. From Equation (C.2),

$$\begin{aligned}Q_t(\Theta) &= \sum_{k=1}^{K} \mathbb{E}_J[\hat{N}_k] \ln w_k + \sum_{k=1}^{K} \mathbb{E}_J \left[\sum_{\mathbf{x}_n \in \hat{X}_k} \ln P(\mathbf{x}_n; \theta_k) \right] \\ &\overset{(a)}{=} \sum_{k=1}^{K} \mathbb{E}_J \left[\sum_{n=1}^{N} z_{nk} \right] \ln w_k + \sum_{k=1}^{K} \mathbb{E}_J \left[\sum_{n=1}^{N} z_{nk} \ln P(\mathbf{x}_n; \theta_k) \right] \\ &\overset{(b)}{=} \sum_{k=1}^{K} N_k \ln w_k + \sum_{k=1}^{K} \sum_{n=1}^{N} \gamma_{nk} \ln P(\mathbf{x}_n; \theta_k),\end{aligned}$$

where $\Theta = \{w_1, \ldots, w_K; \theta_1, \ldots, \theta_K\}$. In (a), we introduced an indicator random variable $z_{nk} = [\![\mathbf{x}_n \in \hat{X}_k]\!]$ which is 1 if $\mathbf{x}_n$ is from bump k and 0 otherwise; in (b), we defined

$$N_k = \mathbb{E}_J[\hat{N}_k] = \mathbb{E}_J\left[\sum_{n=1}^{N} z_{nk}\right] = \sum_{n=1}^{N} \gamma_{nk},$$

where we used $\mathbb{E}[z_{nk}] = \gamma_{nk}$. Now that we have an explicit functional form for $Q_t(\Theta)$, we can perform the maximization step. Observe that the bump-parameters $\theta_k \in \Theta$ are occurring in independent terms, and so can be optimized separately. As for the first term, observe that

$$\begin{aligned} \frac{1}{N}\sum_{k=1}^{K} N_k \ln w_k &= \sum_{k=1}^{K} \frac{N_k}{N} \ln \frac{w_k}{N_k/N} + \sum_{k=1}^{K} \frac{N_k}{N} \ln \frac{N_k}{N} \\ &\le \sum_{k=1}^{K} \frac{N_k}{N} \ln \frac{N_k}{N}, \end{aligned}$$

where the last inequality follows from Jensen's inequality and the concavity of the logarithm, which implies:

$$\sum_{k=1}^{K} \frac{N_k}{N} \ln \frac{w_k}{N_k/N} \le \ln\left(\sum_{k=1}^{K} \frac{N_k}{N} \cdot \frac{w_k}{N_k/N}\right) = \ln\left(\sum_{k=1}^{K} w_k\right) = \ln(1) = 0.$$

Equality holds when $w_k = N_k/N$. Maximizing $Q_t(w_1, \ldots, w_K, \theta_1, \ldots, \theta_K)$, therefore, gives the following updates:

$$w_k^* = \frac{N_k}{N} = \frac{\sum_{n=1}^{N} \gamma_{nk}}{N}; \tag{C.6}$$

$$\theta_k^* = \operatorname*{argmax}_{\theta_k} \sum_{n=1}^{N} \gamma_{nk} \ln P(\mathbf{x}_n; \theta_k). \tag{C.7}$$

The update to get w_k^* can be viewed as follows. A fraction γ_{nk} of data point $\mathbf{x}_n$ belongs to bump k. Thus, $N_k = \sum_n \gamma_{nk}$ is the total number of data points belonging to bump k. Similarly, the update to get θ_k^* can be viewed as follows. The 'likelihood' for the parameter θ_k of bump k is just the weighted sum of the likelihoods of each point, weighted by the fraction of the point that belongs to bump k. This intuitive interpretation of the update is another reason for the popularity of the E-M algorithm.

The E-M algorithm for mixture density estimation is remarkably simple once we get past the machinery used to set it up: we have an analytic update for the weights w_k and K *separate* optimizations for each bump parameter θ_k. The miracle is that these simple updates are *guaranteed* to improve the log-likelihood (from Theorem C.2). There are other ways to maximize the likelihood, for example using gradient and Hessian based iterative optimization techniques. However, the E-M algorithm is simpler and works well in practice.

Example Let's derive the E-M update for the GMM with current parameter estimate is $\Theta_t = \{\hat{w}_1, \ldots, \hat{w}_K; \hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_K; \hat{\Sigma}_1, \ldots, \hat{\Sigma}_K\}$. Let $\hat{\mathrm{S}}_k = (\hat{\Sigma}_k)^{-1}$. The posterior bump probabilities γ_{nk} for the parameters Θ_t are:

$$\gamma_{nk} = \frac{\hat{w}_k P(\mathbf{x}_n | \hat{\boldsymbol{\mu}}_k, \hat{\mathrm{S}}_k)}{\sum_{\ell=1}^{K} \hat{w}_\ell P(\mathbf{x}_n | \hat{\boldsymbol{\mu}}_\ell, \hat{\mathrm{S}}_\ell)},$$

where $P(\mathbf{x}|\boldsymbol{\mu}, \mathrm{S}) = (2\pi)^{-d/2} |\mathrm{S}|^{1/2} \exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \mathrm{S}(\mathbf{x} - \boldsymbol{\mu}))$. The w_k update is immediate from (C.6), which matches Equation (6.9). Since

$$\ln P(\mathbf{x}|\boldsymbol{\mu}, \mathrm{S}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \mathrm{S}(\mathbf{x} - \boldsymbol{\mu}) + \frac{1}{2} \ln |\mathrm{S}| - \frac{d}{2} \ln(2\pi),$$

to get the $\boldsymbol{\mu}_k$ and S_k updates using (C.7), we need to minimize

$$\sum_{n=1}^{N} \gamma_{nk} \left((\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}} \mathrm{S}_k (\mathbf{x}_n - \boldsymbol{\mu}_k) - \ln |\mathrm{S}_k| \right).$$

Setting the derivative with respect to $\boldsymbol{\mu}_k$ to $\mathbf{0}$, gives

$$2\mathrm{S_k} \sum_{n=1}^{N} \gamma_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k) = \mathbf{0}.$$

Since S_k is invertible, $\boldsymbol{\mu}_k \sum_{n=1}^{N} \gamma_{nk} = \sum_{n=1}^{N} \gamma_{nk} \mathbf{x}_n$, and since $N_k = \sum_{n=1}^{N} \gamma_{nk}$, we obtain the update in (6.9),

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \mathbf{x}_n.$$

To take the derivative with respect to S_k, we use the identities in the hint of Exercise C.1(b). Setting the derivative with respect to S_k to zero gives

$$\sum_{n=1}^{N} \gamma_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}} - \mathrm{S}_k^{-1} \sum_{n=1}^{N} \gamma_{nk} = 0,$$

or

$$\mathrm{S}_k^{-1} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}.$$

Since $\mathrm{S}_k^{-1} = \Sigma_k$, we recover the update in (6.9).

Commentary The E-M algorithm is a remarkable example of a recurring theme in learning. We want to learn a model Θ that explains the data. We start with a guess $\hat{\Theta}$ that is *wrong*. We use this wrong model to estimate some other quantities of the world (the bump memberships in our example). We now learn a new model which is better than the old model at explaining the combined data plus the inaccurate estimates of the other quantities. Miraculously, by doing this, we bootstrap ourselves up to a better model, one that is better at explaining the data. This theme reappears in Reinforcement Learning as well. If we didn't know better, it seems like a free lunch.

C.2 Problems

Problem C.1 Consider the general case of Example C.1. The sample has N balls, with N_g green, N_b blue and N_r red ($N_g + N_b + N_r = N$). Show that the log-likelihood of the incomplete data is

$$N_g \ln(1-\mu_1) + N_b \ln(1-\mu_2) + N_r \ln(\mu_1+\mu_2) - N \ln 2. \tag{C.8}$$

What are the maximum likelihood estimates for μ_1, μ_2. (Be careful with the cases $N_g > N/2$ and $N_b > N/2$.

Problem C.2 Consider the general case of Example C.1 as in Problem C.1, with N_g green, N_b blue and N_r red balls.

(a) Suppose your starting estimates are $\hat{\mu}_1, \hat{\mu}_2$. For a red ball, what are $p_1 = \mathbb{P}[\text{Bag } 1|\hat{\mu}_1, \hat{\mu}_2]$ and $p_2 = \mathbb{P}[\text{Bag } 2|\hat{\mu}_1, \hat{\mu}_2]$

(b) Let N_{r_1} and N_{r_2} (with $N_r = N_{r_1} + N_{r_2}$) be the number of red balls from Bag 1 and 2 respectively. Show that the log-likelihood of the complete data is

$$N_g \ln(1-\mu_1) + N_b \ln(1-\mu_2) + N_{r_1} \ln(\mu_1) + N_{r_2} \ln(\mu_2) - N \ln 2.$$

(c) Compute the function $Q_t(\mu_1, \mu_2)$ by taking the expectation of the log-likelihood of the complete data. Show that

$$Q_t(\mu_1,\mu_2) = N_g \ln(1-\mu_1) + N_b \ln(1-\mu_2) + p_1 N_r \ln(\mu_1) + p_2 N_r \ln(\mu_2) - N \ln 2.$$

(d) Maximize $Q_t(\mu_1, \mu_2)$ to obtain the E-M update.

(e) Show that repeated E-M iteration will ultimately converge to

$$\hat{\mu}_1 = \frac{N - 2N_g}{N} \qquad \hat{\mu}_2 = \frac{N - 2N_b}{N}.$$

Problem C.3 A sequence of N balls, $X = x_1, \ldots, x_N$ is drawn iid as follows. There are 2 bags. Bag 1 contains only red balls and bag 2 contains red and blue balls. A fraction π in this second bag are red. A bag is picked randomly with probability $\frac{1}{2}$ and one of the balls is picked randomly from that bag; $x_n = 1$ if ball n is red and 0 if it is blue. You are given N and the number of red balls $N_r = \sum_{n=1}^{N} x_n$.

(a) (i) Show that the likelihood $P[X|\pi, N]$ is

$$P[X|\pi,N] = \prod_{n=1}^{N} \left(\frac{1+\pi}{2}\right)^{x_n} \left(\frac{1-\pi}{2}\right)^{1-x_n}.$$

Maximize to obtain an estimate for π (be careful with $N_r < N/2$).

(ii) For $N_r = 600$ and $N = 1000$, what is your estimate of π.

(b) Maximizing the likelihood is tractable for this simple problem. Now develop an E-M iterative approach.

(i) What is an appropriate hidden/unmeasured variable $J = j_1, \ldots, j_N$.

(ii) Give a formula for the likelihood for the full data, $\mathbb{P}[X, J|\pi, N]$.

(iii) If at step t your estimate is π_t, for the expectation step, compute $Q_t(\pi) = \mathbb{E}_{J|X,\pi_t}[-\ln \mathbb{P}[X, J|\pi, N]]$ and show that

$$Q_t(\pi) = \frac{\pi_t}{1+\pi_t} N_r \ln \pi + (N - N_r)\ln(1-\pi),$$

and hence show that the E-M update is given by

$$\pi_{t+1} = \frac{\pi_t N_r}{\pi_t N_r + (1+\pi_t)(N-N_r)}.$$

What are the limit points when $N_r \geq N/2$ and $N_r < N/2$?

(iv) Plot π_t versus t, starting from $\pi_0 = 0.9$ and $\pi_0 = 0.2$, with $N_r = 600, N = 1000$.

(v) The values of the hidden variables can often be useful. After convergence of the E-M, how could you get estimates of the hidden variables?

Problem C.4 **[E-M for Supervised Learning]** We wish to learn a function $f(x)$ which predicts the temperature as a function of the time x of the day, $x \in [0,1]$. We believe that the temperature has a linear dependence on time, so we model f with the linear hypotheses, $h(x) = w_0 + w_1 x$.

We have N data points $y_1, \ldots, y_N$, the temperature measurements on different (independent) days, where $y_n = f(x_n) + \epsilon_n$ and $\epsilon_n \sim N(0,1)$ is iid zero mean Gaussian noise with unit variance. The problem is that we do not know the time x_n at which these measurements were made. Assume that each temperature measurement was taken at some random time in the day, chosen uniformly on $[0,1]$.

(a) Show that the log-likelihood for weights $\mathbf{w}$ is

$$\ln \mathbb{P}[\mathbf{y}|\mathbf{w}] = \sum_{n=1}^{N} \ln\left(\int_0^1 dx\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_n - w_0 - w_1 x)^2}\right).$$

(b) What is the natural hidden variable $J = j_1, \ldots, j_N$.

(c) Compute the log-likelihood for the complete data $\ln \mathbb{P}[\mathbf{y}, J|\mathbf{w}]$.

(d) Let $\gamma_n(x|\mathbf{w}) = P(x_n = x|y_n, \mathbf{w})$. Show that, for $x \in [0,1]$,

$$\gamma_n(x|\mathbf{w}) = \frac{\exp\left(-\frac{1}{2}(y_n - w_0 - w_1 x)^2\right)}{\int_0^1 dx\ \exp\left(-\frac{1}{2}(y_n - w_0 - w_1 x)^2\right)}.$$

Hence, compute $Q_t(\mathbf{w})$.

(e) Let $\alpha_n = \mathbb{E}_{\gamma_n(x|\mathbf{w}_t)}[x]$ and $\beta_n = \mathbb{E}_{\gamma_n(x|\mathbf{w}_t)}[x^2]$ (expectations taken with respect to the distribution $\gamma_n(x|\mathbf{w}_t)$). Show that the EM-updates are

$$\begin{aligned} w_1(t+1) &= \frac{\frac{1}{N}\sum_{i=1}^N (y_i - \overline{y})(\alpha_i - \overline{\alpha})}{\overline{\beta} - \overline{\alpha}^2}; \\ w_0(t+1) &= \overline{y} - w_1(t+1)\overline{\alpha}; \end{aligned}$$

where, $\overline{(\cdot)}$ denotes averaging (eg. $\overline{\alpha} = \frac{1}{N}\sum_{n=1}^N \alpha_n$) and $\mathbf{w}(t)$ are the weights at iteration t.

(f) What happens if the temperature measurement is not at a uniformly random time, but at a time distributed according to an unknown $P(x)$? You have to maintain an estimate $P_t(x)$. Now, show that

$$\gamma_n(x|\mathbf{w}) = \frac{P_t(x)\exp\left(-\frac{1}{2}(y_n - w_0 - w_1 x)^2\right)}{\int_0^1 dx\ P_t(x)\exp\left(-\frac{1}{2}(y_n - w_0 - w_1 x)^2\right)},$$

and that the updates in (e) are unchanged, except that they use this new $\gamma_n(x|\mathbf{w}_t)$. Show that the update to P_t is given by

$$P_{t+1}(x) = \frac{1}{N}\sum_{n=1}^N \gamma_n(x|\mathbf{w}_t).$$

What happens if you tried to maximize the log-likelihood for the incomplete data, instead of using the E-M approach?