Chapter 7

Expectation-Maximization *

7.1 Overview

Expectation-maximization (EM) is a method for dealing with missing data. For example, for classification, the complete data consists of the features $\boldsymbol{x}$ and labels y, as shown in the left panel of Figure 7.1. With a probabilistic model for this data, we can find the parameters for each class through maximum likelihood, where the log-likelihood function is

$$\log p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}),$$

where $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ and $\boldsymbol{\theta}$ represents the parameters of class-conditional distributions for each of the classes.

But what if the class labels are not given as in the right panel of Figure 7.1? The problem becomes more difficult, but doesn't seem hopeless as we can still distinguish two clusters and assign points to these with various degrees of confidence.

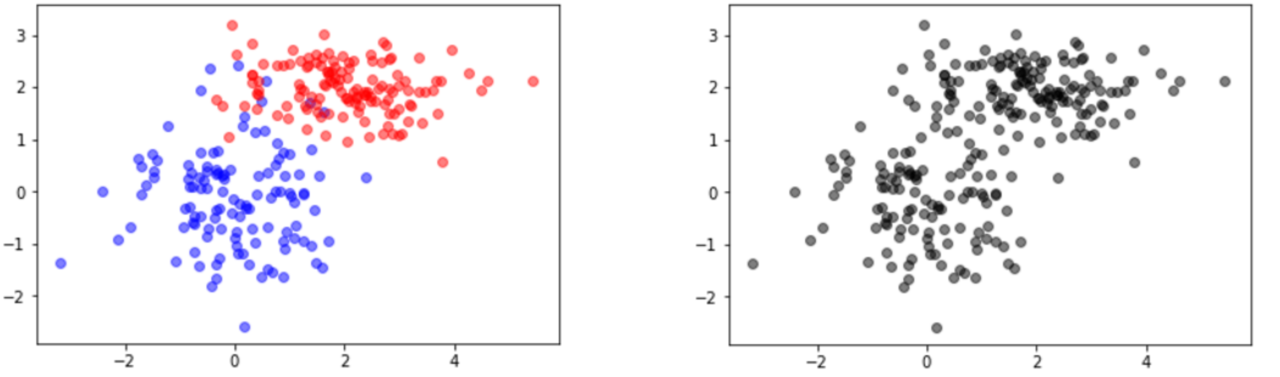


Figure 7.1: Data from two classes, with labels given as colors (left) and not given (right).

We thus formulate this problem as finding $\boldsymbol{\theta}$ that maximizes

$$\log p(\boldsymbol{x}; \boldsymbol{\theta}) = \log \sum_{\boldsymbol{y}} p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$$

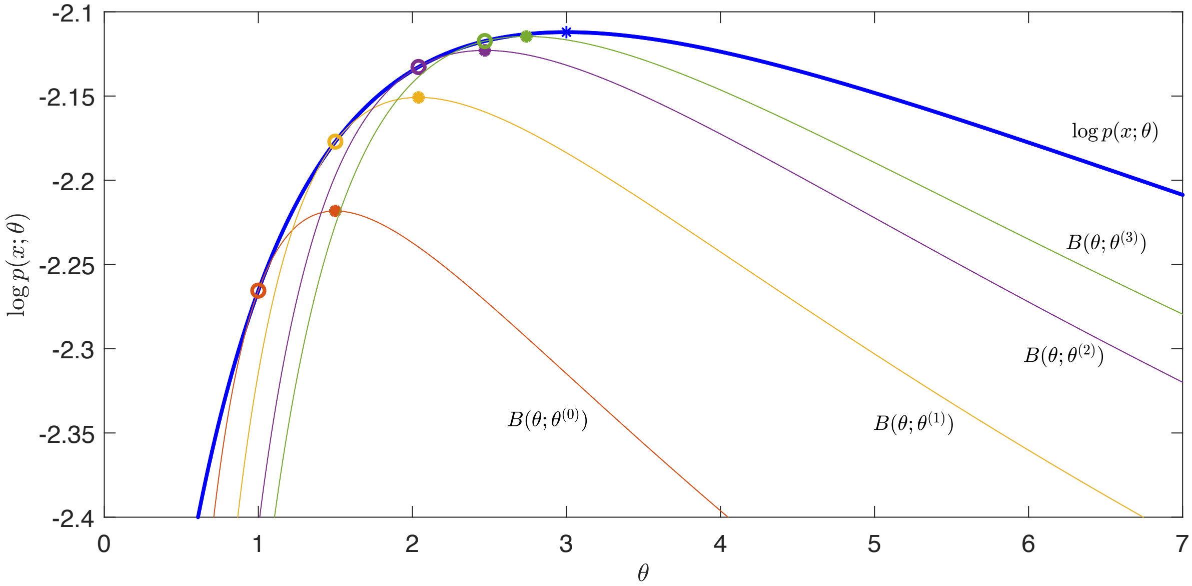


Figure 7.2: The log-likelihood of the observation and consecutive EM lower bounds and estimates. In each iteration, the current value of θ is denoted by $\circ$ and the new value by $*$. Here, $\theta^{(0)} = 1, \theta^{(1)} = 1.5, \theta^{(2)} = 2.04, \theta^{(3)} = 2.472$. Continuing in the same manner, we would obtain estimates $2.740, 2.880, 2.946, 2.976, \ldots$, where 3 is the true maximum.

In this case, $(\boldsymbol{x}, \boldsymbol{y})$ is the complete data, for which computing the likelihood is easy, but a component of this data, namely $\boldsymbol{y}$, is missing. Now computing the likelihood is difficult because of the summation, which is typically over a large number of possibilities. Expectation-maximization is a method for handling missing data.

EM is an iterative method that given the current estimate for the parameter, finds a new estimate. The idea behind EM is finding lower bounds on the log-likelihood of the observed data and maximizing these lower bounds. This is illustrated in Figure 7.2 (see Example 75). Suppose our current estimate of θ is θ'. In each iteration, we find a lower bound $B(\theta, \theta')$ on $\log p(x;\theta)$ that coincides with it at $\theta = \theta'$, i.e.,

$$\begin{aligned} \log p(x;\theta) &\geq B(\theta,\theta'), && \text{for all } \theta,\\ \log p(x;\theta) &= B(\theta,\theta'), && \text{for } \theta = \theta'. \end{aligned} \tag{7.1}$$

Now let our new estimate be

$$\theta'' = \arg\max_{\theta} B(\theta, \theta').$$

Note that we have not used $\log p(x;\theta)$ to find θ''. It follows that

$$\log p(x;\theta'') \geq \log p(x;\theta').$$

So our new estimate is at least as good as the old one, and under certain conditions, it is going to be strictly better. We then use θ'' in place of θ' and repeat. Note that if $\log p(x;\theta)$ is bounded, since the sequence $\log p(x;\theta')$ is non-decreasing, it will converge. Under appropriate conditions, this means that θ' also converges to a stationary point of $p(x;\theta)$. See [1] for details.

It remains to find a lower bound that satisfies (7.1). For the likelihood of the observation and for any y such that $p(y|x;\theta) > 0$,

$$\ell(\theta) = \ln p(x;\theta) = \ln \frac{p(x,y;\theta)}{p(y|x;\theta)}.$$

Then, for any distribution q for the missing data y,

$$\begin{aligned}\ell(\theta) &= \sum_y q(y) \ln \frac{p(x,y;\theta)}{p(y|x;\theta)} \\ &\geq \sum_y q(y) \ln \frac{p(x,y;\theta)}{p(y|x;\theta)} - D(q(y)||p(y|x;\theta)) \\ &= \sum_y q(y) \ln \frac{p(x,y;\theta)}{p(y|x;\theta)} - \sum_y q(y) \ln \frac{q(y)}{p(y|x;\theta)} \\ &= \sum_y q(y) \ln p(x,y;\theta) - \sum_y q(y) \ln q(y),\end{aligned}$$

where for two distribution p_1 and p_2, $D(p_1(z)\|p_2(z))$ is the *relative entropy* (also called the Kullback–Leibler divergence or KL divergence) between p_1 and p_2 defined as

$$\sum_z p_1(z) \log \frac{p_1(z)}{p_2(z)}.$$

Relative entropy is a measure of dissimilarity between distributions and can be shown to be non-negative and is equal to 0 if and only if $p_1 = p_2$.

Thus for any distribution q, we have a lower bound on $\ell(\theta)$. Suppose our current guess for θ is $\theta^{(t)}$. We would like this lower bound to be equal to $\ell(\theta)$ at $\theta = \theta^{(t)}$. For this to occur, we need

$$D(q(y)||p(y|x;\theta^{(t)})) = 0 \iff q(y) = p(y|x;\theta^{(t)}),$$

resulting in

$$\ell(\theta) \geq \sum_y p(y|x;\theta^{(t)}) \ln p(x,y;\theta) - \sum_y p(y|x;\theta^{(t)}) \ln p(y|x;\theta^{(t)}) = B(\theta,\theta^{(t)}).$$

Now instead of maximizing ℓ, we can maximize B. We note however that the second term in B is not a function of θ. So we instead define the following expectation

$$Q(\theta,\theta^{(t)}) = \sum_y p(y|x;\theta^{(t)}) \ln p(x,y;\theta),$$

and find

$$\theta^{(t+1)} = \arg\max_\theta Q(\theta,\theta^{(t)}).$$

For simplicity of notation, I often use θ' to denote $\theta^{(t)}$ and θ'' to denote $\theta^{(t+1)}$. Also, let $\mathbb{E}'$ be expected value assuming the value of θ'. We can then describe the EM algorithm as

- The E-step:
$$Q(\theta;\theta') = \sum_y p(y|x;\theta') \ln p(x,y;\theta) = \mathbb{E}'[\ln p(x,y;\theta)|x]$$

- The M-step:
$$\theta'' = \arg\max_\theta Q(\theta;\theta').$$

Update $\theta' \leftarrow \theta''$ and repeat.

Roughly speaking, EM can be viewed as alternatively finding an estimate for the missing data through expectation by assuming a value for the parameters (the E-step) and finding a new estimate for the parameter based on the estimate of the data.

7.2 Clustering with EM

For classification the complete data is $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. When the labels y_i are missing, the problem becomes *clustering.*

We assume Gaussian class-conditionals:

$$\begin{aligned} p(y_i = 1) &= \pi, & \boldsymbol{x}_i|y_i = 1 &\sim \mathcal{N}(\mu_1, K_1) \\ p(y_i = 0) &= 1 - \pi, & \boldsymbol{x}_i|y_i = 0 &\sim \mathcal{N}(\mu_0, K_0) \end{aligned}$$

Let $\boldsymbol{\theta} = (\pi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, K_0, K_1)$. Ideally, we would want to maximize the likelihood for the observed data $\{(\boldsymbol{x}_i)\}_{i=1}^n$,

$$\ell(\boldsymbol{\theta}) = \ln p(\boldsymbol{x}_1^n|\boldsymbol{\theta}) = \ln \sum_{y_1^n} p(\boldsymbol{x}_1^n, y_1^n|\boldsymbol{\theta}).$$

But this is difficult to do because of a lack of an analytical solution due to the summation. Instead, we can use a computational method such as EM.

We will proceed as follows:

- **Set-up:** It is helpful to start with the log-likelihood of the complete data and simplify it before proceeding to the EM algorithm. We have

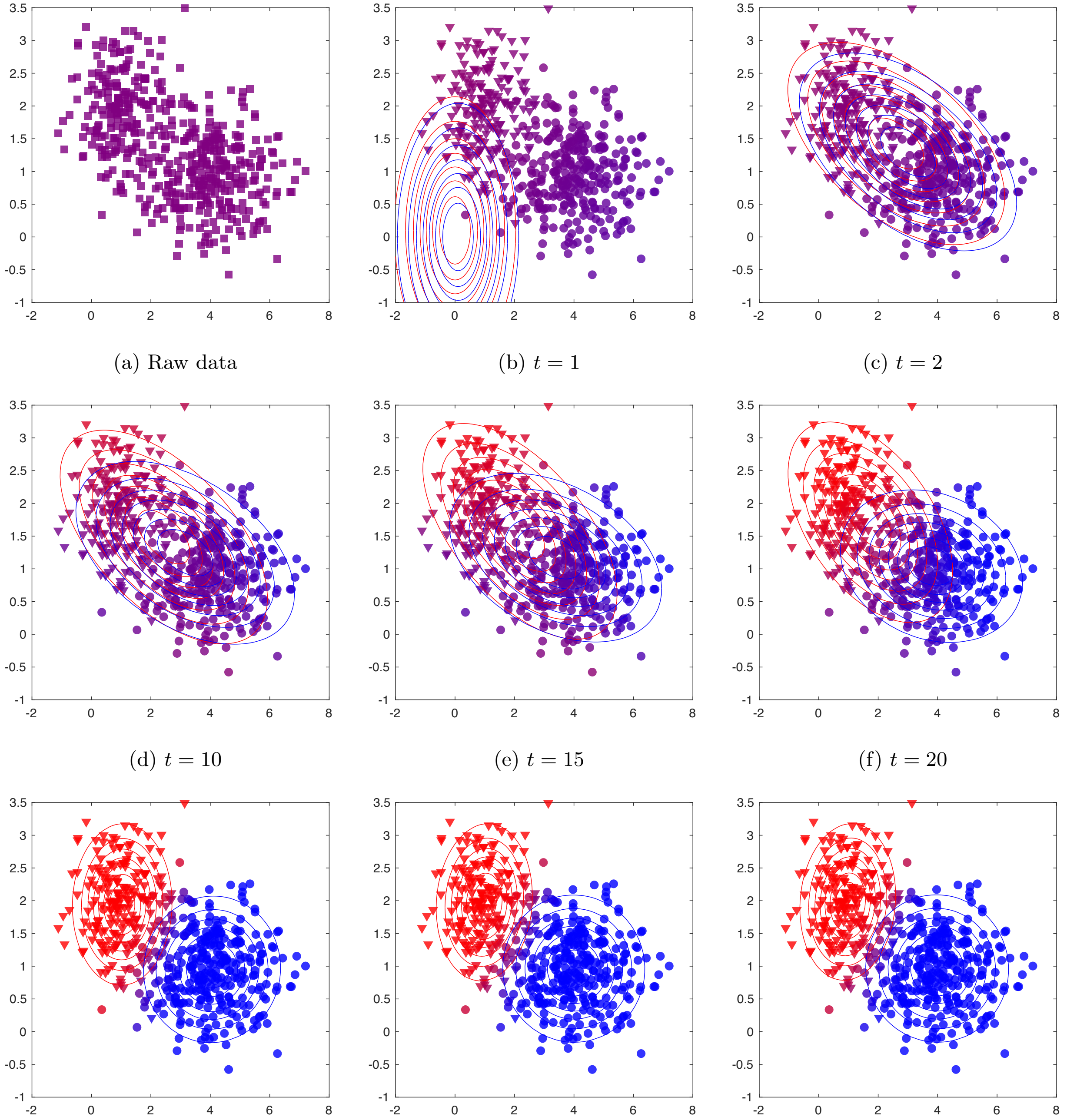
$$\ln p(\boldsymbol{x}_1^n, y_1^n; \theta) = \sum_{i=1}^n \ln p(\boldsymbol{x}_i, y_i; \theta),$$

  and for each term in this sum,

$$\begin{aligned} \ln p(\boldsymbol{x}_i, y_i; \theta) &= \ln\Big((\pi p(\boldsymbol{x}_i|y_i = 1; \theta))^{y_i} ((1-\pi) p(\boldsymbol{x}_i|y_i = 0; \theta))^{1-y_i} \Big) \\ &= y_i \ln(\pi p(\boldsymbol{x}_i|y_i = 1; \theta)) + (1 - y_i) \ln((1-\pi) p(\boldsymbol{x}_i|y_i = 0; \theta)). \end{aligned}$$

- **The E-step:** Let θ' be the current estimate for θ and let $\mathbb{E}'$ denote expected value operator with respect to the distribution $p(y|x; \theta')$. We have

$$\begin{aligned} Q(\theta; \theta') &= \mathbb{E}'[\ln p(\boldsymbol{x}_1^n, y_1^n; \theta)|\boldsymbol{x}_1^n] \\ &= \mathbb{E}'\left[\sum_{i=1}^n \ln p(\boldsymbol{x}_i, y_i; \theta)|\boldsymbol{x}_1^n\right] \\ &= \sum_{i=1}^n \mathbb{E}'[\ln p(\boldsymbol{x}_i, y_i; \theta)|\boldsymbol{x}_i] \end{aligned}$$



(a) Raw data

(b) $t = 1$

(c) $t = 2$

(d) $t = 10$

(e) $t = 15$

(f) $t = 20$

(g) $t = 30$

(h) $t = 40$

(i) $t = 50$

Figure 7.3: EM clustering of a mixture of two Gaussian datasets. In (a) the raw data is shown and in (b-i), steps of the EM algorithm are shown. To compare with the underlying distributions and clusters, the points from each of the Gaussian distributions are shown with triangles and circles. However, the EM algorithm does not have access to this data. The contour plots represent the current estimate for the parameters of each of the Gaussian distributions and the color of each data point represents the estimate of the EM algorithm for the probability that the point belongs to the clusters ($\gamma_i' = p(y_i = 1|\boldsymbol{x}_i; \theta')$). A video of the clustering can be found here.

And for each term in the sum,

$$\begin{aligned}\mathbb{E}'[\ln p(\boldsymbol{x}_i, y_i; \theta)|\boldsymbol{x}_i] &= \mathbb{E}'[y_i \ln(\pi p(\boldsymbol{x}_i|y_i = 1; \theta)) + (1 - y_i)\ln((1 - \pi)p(\boldsymbol{x}_i|y_i = 0; \theta))|\boldsymbol{x}_i]\\ &= \mathbb{E}'[y_i|\boldsymbol{x}_i]\ln(\pi p(\boldsymbol{x}_i|y_i = 1; \theta)) + \mathbb{E}'[1 - y_i|\boldsymbol{x}_i]\ln((1 - \pi)p(\boldsymbol{x}_i|y_i = 0; \theta))\\ &= \gamma_i' \ln \pi + (1 - \gamma_i')\ln(1 - \pi) + \gamma_i' \ln p(\boldsymbol{x}_i|y_i = 1; \theta) + (1 - \gamma_i')\ln p(\boldsymbol{x}_i|y_i = 0; \theta),\end{aligned}$$

where

$$\begin{aligned}\gamma_i' &= \mathbb{E}'[y_i|\boldsymbol{x}_i]\\ &= p(y_i = 1|\boldsymbol{x}_i; \theta')\\ &= \frac{p(x_i, y_i = 1; \theta')}{p(x_i, y_i = 1; \theta') + p(x_i, y_i = 0; \theta')}\\ &= \frac{\pi'\mathcal{N}(x_i; \mu_1', K_1')}{\pi'\mathcal{N}(x_i; \mu_1', K_1') + (1 - \pi')\mathcal{N}(x_i; \mu_0', K_0')}.\end{aligned}$$

Here, γ_i' has a significant meaning. It represents the probability that a given point $\boldsymbol{x}_i$ belongs to class 1 given the current estimate of the parameters. Instead of computing the likelihood based on a known value for y_i, in the E-step, we compute the likelihood by partially assigning $\boldsymbol{x}_i$ to class 1 and to class 0.

- **The M-step:** To find π'':

$$\frac{\partial Q}{\partial \pi} = \sum_{i=1}^{n}\left(\frac{\gamma_i'}{\pi} - \frac{1 - \gamma_i'}{1 - \pi}\right) = 0 \Rightarrow \pi'' = \frac{\sum_{i=1}^n \gamma_i'}{n}.$$

To find μ_1'' :

$$\begin{aligned}\frac{\partial Q}{\partial \mu_1} &= \frac{\partial}{\partial \mu_1}\sum_{i=1}^{n}\gamma_i' \ln p(\boldsymbol{x}_i|y_i = 1; \theta)\\ &= \frac{\partial}{\partial \mu_1}\sum_{i=1}^{n}\gamma_i'\left(-\frac{1}{2}(\boldsymbol{x}_i - \mu_1)^T K_1^{-1}(\boldsymbol{x}_i - \mu_1)\right)\\ &= \sum_{i=1}^{n}\gamma_i' K_1^{-1}(\boldsymbol{x}_i - \mu_1) = 0 \Rightarrow \mu_1'' = \frac{\sum_{i=1}^n \gamma_i'\boldsymbol{x}_i}{\sum_{i=1}^n \gamma_i'}.\end{aligned}$$

To find K_1'':

$$\begin{aligned}\frac{\partial Q}{\partial K_1^{-1}} &= \frac{\partial}{\partial K_1^{-1}}\sum_{i=1}^{n}\gamma_i'\left(\frac{1}{2}\ln\left|K_1^{-1}\right| - \frac{1}{2}(\boldsymbol{x}_i - \mu_1)^T K_1^{-1}(\boldsymbol{x}_i - \mu_1)\right)\\ &= \frac{1}{2}K_1\sum_{i=1}^{n}\gamma_i' - \frac{1}{2}\sum_{i=1}^{n}\gamma_i'(\boldsymbol{x}_i - \mu_1)(\boldsymbol{x}_i - \mu_1)^T = 0 \Rightarrow K_1'' = \frac{\sum_{i=1}^n \gamma_i'(\boldsymbol{x}_i - \mu_1)(\boldsymbol{x}_i - \mu_1)^T}{\sum_{i=1}^n \gamma_i'}.\end{aligned}$$

Several steps of an EM clustering of a dataset are shown in Figure 7.3. In essence, the EM algorithm uses the current estimates of posterior class probabilities of a point as labels and updates the distributions. Having updated the distributions, it updates the posterior class probabilities and repeats.

7.3 EM with general missing data **

So far, we have considered problems in which data can be divided into an observed component x and a hidden component y, with the expectation given by

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}') = \sum_y p(y|x;\boldsymbol{\theta}') \ln p(x,y;\boldsymbol{\theta})$$

But we can use EM to solve a more general class of problems, where this division may not be possible. Specifically, we assume that the complete data is given by z and the observed data is given by x, where x is a function of z. In this case, the expectation is given by

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}') = \sum_z p(z|x;\boldsymbol{\theta}') \ln p(z;\boldsymbol{\theta})$$

Example 75 ([1])**.** Let

$$\begin{aligned} &x = s + \epsilon, \\ &s \sim \mathcal{N}(0,\theta), \qquad \theta \ge 0 \\ &\epsilon \sim \mathcal{N}(0,\sigma^2) \qquad \sigma^2 > 0, \end{aligned}$$

where s and ϵ are independent, σ is known, and θ is unknown. Our goal is to estimate θ. In this case, the complete data is $\boldsymbol{z} = (s,\epsilon)$ and observed data is $x = s+\epsilon$.

We can solve this problem directly by noting that

$$x \sim \mathcal{N}(0, \theta+\sigma^2),$$

where we have used

$$\operatorname{Var}(x) = \operatorname{Cov}(s+\epsilon, s+\epsilon) = \sigma^2 + \theta.$$

The maximum likelihood estimate for the variance of x is then

$$\hat{\theta}_{ML} = \begin{cases} x^2 - \sigma^2 & \text{if } x^2 \ge \sigma^2, \\ 0 & \text{if } x^2 < \sigma^2. \end{cases}$$

With EM:

- The E-step:

$$\begin{aligned} Q(\theta;\theta') &= \mathbb{E}'[\ln p(\boldsymbol{z};\theta)|x] \\ &= \mathbb{E}'[\ln p(s;\theta) + \ln p(\epsilon;\theta)|x] \\ &\doteq \mathbb{E}'[\ln p(s;\theta)|x] \\ &\doteq \mathbb{E}'\left[-\frac{\ln\theta}{2} - \frac{s^2}{2\theta}|x\right] \\ &= -\frac{\ln\theta}{2} - \frac{\mathbb{E}'\left[s^2|x\right]}{2\theta} \end{aligned}$$

- The M-step:

$$\frac{\partial Q}{\partial\theta} = -\frac{1}{2\theta} + \frac{\mathbb{E}'\left[s^2|x\right]}{2\theta^2} = 0 \Rightarrow \theta'' = \mathbb{E}'[s^2|x].$$

This is a very intuitive result.

With some manipulation (HW), this results in

$$\theta'' = \left(\frac{\theta'}{\theta' + \sigma^2}\right)^2 x^2 + \frac{\theta' \sigma^2}{\theta' + \sigma^2}.$$

The plot for the log-likelihood and the EM estimates, starting from $\theta^{(0)} = 1$, is given in Figure 7.2, where $\sigma^2 =$ and $x = 2$ and thus $\hat{\theta}_{ML} = 3$.

Bibliography

[1] B. Hajek, *Random Processes for Engineers*. 2014.

[2] T. T. Nguyen and S. Sanner, “Algorithms for direct 0-1 loss optimization in binary classification,” in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML’13, (Atlanta, GA, USA), pp. III–1085–III–1093, JMLR.org, June 2013.