The EM algorithm and Minimum Divergence

Niko Brümmer

October 14, 2009

In Patrick Kenny's recipe[1] for maximum-likelihood training of the parameters of the generative JFA speaker recognition system, he uses an EM-algorithm, which uses two different kinds of M-steps, the standard one and an additional special one denoted *minimum-divergence*. The latter is useful for escaping saddle points in the optimization landscape. This note is intended to help to understand this minimum-divergence step. It is written in general terms and requires no special knowledge of the JFA model.

1 The EM-algorithm

Let the observed *training data* be denoted as x and the generative model that is supposed to have generated x be parametrized by $\boldsymbol{\lambda}$. The EM-algorithm maximizes the log-likelihood, $\log p(x|\boldsymbol{\lambda})$, w.r.t. $\boldsymbol{\lambda}$. Let the generative model include a number of *hidden variables*, collectively denoted by y, where $y \in \mathcal{Y}$. We can decompose the log-likelihood as:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\lambda}) &= \log p(x|\boldsymbol{\lambda}) && (1.1)\\
&= \log \int_{\mathcal{Y}} p(x,y|\boldsymbol{\lambda})\, dy && (1.2)\\
&= \int_{\mathcal{Y}} q(y) \log \frac{p(x,y|\boldsymbol{\lambda})}{q(y)}\, dy + \int_{\mathcal{Y}} q(y) \log \frac{q(y)}{p(y|x,\boldsymbol{\lambda})}\, dy && (1.3)\\
&= \int_{\mathcal{Y}} q(y) \log \frac{p(x,y|\boldsymbol{\lambda})}{q(y)}\, dy + D\big(q(y)\|p(y|x,\boldsymbol{\lambda})\big), && (1.4)
\end{aligned}$$

where $q(y)$ is an arbitrary density and where $D\big(\cdot\|\cdot\big)$ denotes KL-divergence. If we perform the *E-step* of the EM algorithm by assigning:

$$q(y) = p(y|x,\boldsymbol{\lambda}_0), \tag{1.5}$$

[1] See several papers here: `http://www.crim.ca/perso/patrick.kenny/`

where $\boldsymbol{\lambda}_0$ is an initial choice of model parameters, then the first term of the RHS of (1.4) is the *EM-auxiliary*, which we express as:

$$\mathcal{Q}(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}) = \int_{\mathcal{Y}} p(y|x, \boldsymbol{\lambda}_0) \log \frac{p(x, y|\boldsymbol{\lambda})}{p(y|x, \boldsymbol{\lambda}_0)}\, dy. \tag{1.6}$$

If we use the short-hand $D\big(\boldsymbol{\lambda}_0 \| \boldsymbol{\lambda}\big)$ for the second term of (1.4), then:

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathcal{Q}(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}) + D\big(\boldsymbol{\lambda}_0 \| \boldsymbol{\lambda}\big). \tag{1.7}$$

Now since $D\big(\boldsymbol{\lambda}_0 \| \boldsymbol{\lambda}\big) \geq D\big(\boldsymbol{\lambda} \| \boldsymbol{\lambda}\big) = 0$, we find:

$$\mathcal{Q}(\boldsymbol{\lambda}, \boldsymbol{\lambda}) = \mathcal{L}(\boldsymbol{\lambda}) \geq \mathcal{L}(\boldsymbol{\lambda}) - D\big(\boldsymbol{\lambda}_0 \| \boldsymbol{\lambda}\big) = \mathcal{Q}(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}). \tag{1.8}$$

The EM-algorithm makes progress from $\boldsymbol{\lambda}_0$ to $\boldsymbol{\lambda}$, if (during the *M-step*), it can choose $\boldsymbol{\lambda}$ such that $\mathcal{Q}(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}) > \mathcal{Q}(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_0)$, because then:

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathcal{Q}(\boldsymbol{\lambda}, \boldsymbol{\lambda}) \geq \mathcal{Q}(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}) > \mathcal{Q}(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_0) = \mathcal{L}(\boldsymbol{\lambda}_0). \tag{1.9}$$

2 The M-step

Here we are interested in the *M-step*, where given $\boldsymbol{\lambda}_0$, we need to find some $\boldsymbol{\lambda}$, such that $\mathcal{Q}(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}) > \mathcal{Q}(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_0)$. Now, suppose that we can decompose the model parameters $\boldsymbol{\lambda}$ into two components: $\boldsymbol{\lambda} = (\mathbf{V}, \boldsymbol{\Pi})$, so that

$$p(x, y|\boldsymbol{\lambda}) = p(x|y, \mathbf{V}) p(y|\boldsymbol{\Pi}), \tag{2.1}$$

which in turn allows decomposition of $\mathcal{Q}$ as follows:

$$\begin{aligned} \mathcal{Q}(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}) &= \int_{\mathcal{Y}} p(y|x, \boldsymbol{\lambda}_0) \log \frac{p(x|y, \mathbf{V}) p(y|\boldsymbol{\Pi})}{p(y|x, \boldsymbol{\lambda}_0)}\, dy & (2.2) \\ &= \tilde{\mathcal{Q}}(\boldsymbol{\lambda}_0, \mathbf{V}) - D\big(\boldsymbol{\lambda}_0 \| \boldsymbol{\Pi}\big), & (2.3) \end{aligned}$$

where

$$\begin{aligned} \tilde{\mathcal{Q}}(\boldsymbol{\lambda}_0, \mathbf{V}) &= \int_{\mathcal{Y}} p(y|x, \boldsymbol{\lambda}_0) \log p(x|y, \mathbf{V})\, dy, & (2.4) \\ D\big(\boldsymbol{\lambda}_0 \| \boldsymbol{\Pi}\big) &= D\big(p(y|x, \boldsymbol{\lambda}_0) \| p(y|\boldsymbol{\Pi})\big). & (2.5) \end{aligned}$$

Now we can *independently* optimize the two terms in (2.3): $\tilde{\mathcal{Q}}$ can be maximized to find $\mathbf{V}$ and the divergence can be minimized to find $\boldsymbol{\Pi}$.

3 Overparametrized models

Here we continue to work with a model of the form $\boldsymbol{\lambda} = (\mathbf{V}, \boldsymbol{\Pi})$, but we additionally assume that this parametrization is redundant, in the sense that different values of the parameters can be *equivalent*: By $(\mathbf{V}_1, \boldsymbol{\Pi}_1) \equiv (\mathbf{V}_2, \boldsymbol{\Pi}_2)$, we mean that

$$\begin{aligned} p(x|\mathbf{V}_2, \boldsymbol{\Pi}_2) &= \int_{\mathcal{Y}} p(x|y, \mathbf{V}_2) p(y|\boldsymbol{\Pi}_2)\, dy \\ &= \int_{\mathcal{Y}} p(x|y, \mathbf{V}_1) p(y|\boldsymbol{\Pi}_1)\, dy = p(x|\mathbf{V}_1, \boldsymbol{\Pi}_1), \end{aligned} \tag{3.1}$$

for any x. If we use the notation $\mathcal{L}(\mathbf{V}, \boldsymbol{\Pi}) = \log p(x|\mathbf{V}, \boldsymbol{\Pi})$, then (3.1) also implies $\mathcal{L}(\mathbf{V}_2, \boldsymbol{\Pi}_2) = \mathcal{L}(\mathbf{V}_1, \boldsymbol{\Pi}_1)$. We define a model to be *overparametrized*, if for any given $\boldsymbol{\lambda} = (\mathbf{V}, \boldsymbol{\Pi})$ and any given $\boldsymbol{\Pi}'$, there exists a $\mathbf{V}'$, so that $(\mathbf{V}', \boldsymbol{\Pi}') \equiv (\mathbf{V}, \boldsymbol{\Pi})$. For overparametrized models, we consider three different strategies for the EM-algorithm:

3.1 EM Strategy 1

Choose a standard parameter $\bar{\boldsymbol{\Pi}}$ for the hidden variable prior $p(y|\bar{\boldsymbol{\Pi}})$. This parameter is kept constant during the whole EM-algorithm and only the parameter $\mathbf{V}$ is updated. The M-step can be performed by using $\tilde{\mathcal{Q}}$ as auxiliary, rather than using the full auxiliary $\mathcal{Q}$. (In Patrick Kenny's JFA EM-algorithm, he found this strategy is vulnerable to getting stuck in a saddle-point.)

3.2 EM Strategy 2

Here, we have two different M-steps. Both find a $\boldsymbol{\lambda}$ such that $\mathcal{Q}(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}) > \mathcal{Q}(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_0)$:

- In the standard M-step, proceed as above: fix $\bar{\boldsymbol{\Pi}}$ and update only $\mathbf{V}$ by maximizing $\tilde{\mathcal{Q}}$, so that $\boldsymbol{\lambda} = (\mathbf{V}, \bar{\boldsymbol{\Pi}})$.
- In the special M-step: Start with $\boldsymbol{\lambda}_0 = (\mathbf{V}_0, \bar{\boldsymbol{\Pi}})$, then: (i) Temporarily update only $\boldsymbol{\Pi}$ by minimizing the divergence in (2.3). (ii) Normalize $\boldsymbol{\lambda}$ by finding $\mathbf{V}$, such that $\boldsymbol{\lambda} = (\mathbf{V}, \bar{\boldsymbol{\Pi}}) \equiv (\mathbf{V}_0, \boldsymbol{\Pi})$.

The two different M-steps may be interleaved according to a convenient schedule. Patrick Kenny finds[2] a single application of the special M-step

[2] "My experience with min div is that it moves you like a magic carpet from the vicinity of one local optimum to a slightly better one in a completely different part of the parameter space."

(which he denotes the *minimum-divergence step*) is enough to escape the saddle point.

3.3 EM Strategy 3

Just update both $\mathbf{V}$ and $\mathbf{\Pi}$ *simultaneously* in a single M-step, by maximizing $\mathcal{Q}$, as the original EM-algorithm (1.9) suggests. The 'trick' involved here is to use an overparametrized model, rather than one with only the necessary parameters. At any stage during or after the algorithm, equivalence may be used to normalize the model so that $\mathbf{\Pi} = \bar{\mathbf{\Pi}}$.

These three strategies are compared in figure 1, where we exercised a 'JFA-style' EM algorithm on synthetic data. Strategies 1 and 3 in the graph are as described in sections 3.1 and 3.3, while 2a and 2b are as described in section 3.2, with different interleaving schedules: 2a does the special 'min-div' step on every second iteration, while 2b does it only on the second iteration. Figure 2 shows the same strategies, but for 400 iterations, to show that the methods that do use minimum-divergence converge much faster than the one that does not.

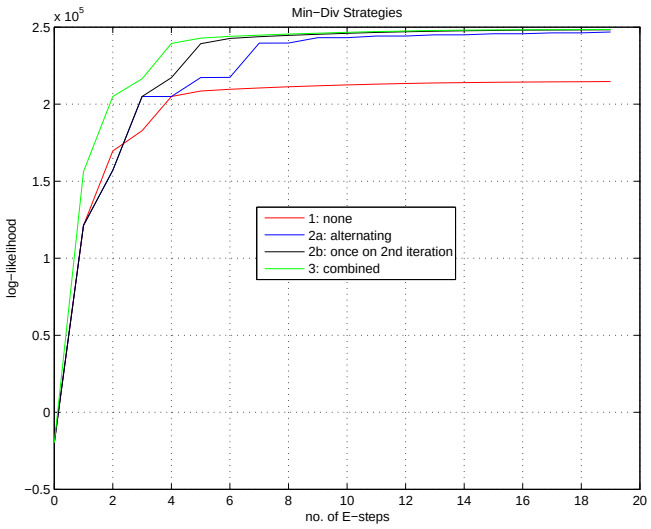

Figure 1: EM strategy comparison.

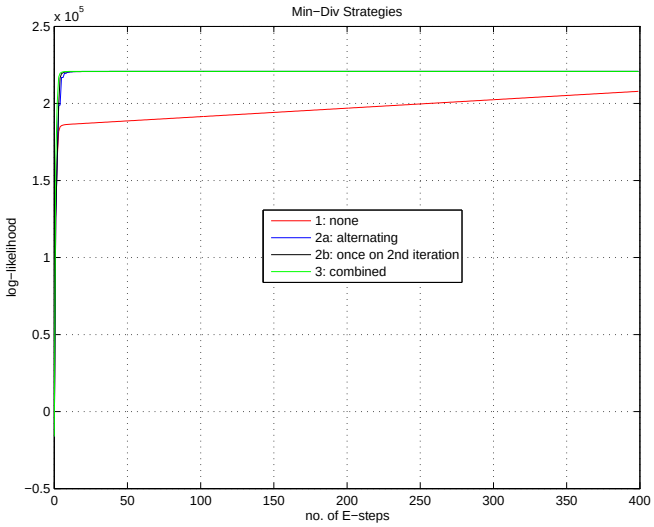

Figure 2: EM strategy comparison (with more iterations).

4 Equivalence via hidden-variable transformation

Here we specialize the concept of equivalence, to find a more concrete way to apply the equivalence transformations. Let $\phi : \mathcal{Y} \mapsto \mathcal{Y}$ be a bijection, so that $\tilde{y} = \phi(y)$ and $y = \phi^{-1}(\tilde{y})$ for any $y, \tilde{y} \in \mathcal{Y}$. If we start with some probability density distribution for y, say $p(y|Z)$, where Z is any conditioning, then the corresponding conditional density for $\tilde{y}$ is:

$$\tilde{p}(\tilde{y}|Z) = \frac{1}{|\det(\mathbf{J})|} p(\phi^{-1}(\tilde{y})|Z) \tag{4.1}$$

where $|\det(\mathbf{J})|$ is the absolute value of the determinant of the Jacobian of the transformation ϕ. In general, $\mathbf{J}$ is dependent on y, or equivalently on $\tilde{y}$, but we don't show this to avoid clutter.

For example, let $y \in \mathbb{R}^n$ have a standard normal distribution, $y \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

and we apply an invertible linear transform $\tilde{y} = \mathbf{J}y$ and $y = \mathbf{J}^{-1}\tilde{y}$, then:

$$p(y) = \frac{1}{\sqrt{\det(2\pi\mathbf{I})}} \exp(-\tfrac{1}{2}y'y), \tag{4.2}$$

$$\tilde{p}(\tilde{y}) = \frac{1}{|\det(\mathbf{J})|} \frac{1}{\sqrt{\det(2\pi\mathbf{I})}} \exp(-\tfrac{1}{2}\tilde{y}'(\mathbf{J}^{-1})'\mathbf{J}^{-1}\tilde{y}), \tag{4.3}$$

$$= \frac{1}{\sqrt{\det(2\pi\mathbf{J}\mathbf{J}')}} \exp(-\tfrac{1}{2}\tilde{y}'(\mathbf{J}\mathbf{J}')^{-1}\tilde{y}), \tag{4.4}$$

$$= \mathcal{N}(\tilde{y}, \mathbf{0}, \mathbf{J}\mathbf{J}') \tag{4.5}$$

Returning to the general bijection $\phi : \mathcal{Y} \mapsto \mathcal{Y}$, let us transform the second integral of (3.1), via the substitutions $y = \phi^{-1}(\tilde{y})$ and $dy = \frac{1}{|\det(\mathbf{J})|} d\tilde{y}$:

$$\int_{\mathcal{Y}} p(x|y, \mathbf{V}_1) p(y|\mathbf{\Pi}_1)\, dy \tag{4.6}$$

$$= \int_{\mathcal{Y}} p(x|\phi^{-1}(\tilde{y}), \mathbf{V}_1) \frac{p(\phi^{-1}(\tilde{y})|\mathbf{\Pi}_1)}{|\det(\mathbf{J})|}\, d\tilde{y} \tag{4.7}$$

$$= \int_{\mathcal{Y}} p(x|\phi^{-1}(\tilde{y}), \mathbf{V}_1) \tilde{p}(\tilde{y}|\mathbf{\Pi}_1)\, d\tilde{y} \tag{4.8}$$

If our parametric model allows us to find $(\mathbf{V}_2, \mathbf{\Pi}_2)$, such that

$$p(\tilde{y}|\mathbf{\Pi}_2) = \tilde{p}(\tilde{y}|\mathbf{\Pi}_1), \tag{4.9}$$

$$p(x|\tilde{y}, \mathbf{V}_2) = p(x|\phi^{-1}(\tilde{y}), \mathbf{V}_1), \tag{4.10}$$

for every x and $\tilde{y}$, then (3.1) is satisfied.

In summary, the bijective hidden-variable transformation $\phi : \mathcal{Y} \mapsto \mathcal{Y}$, with Jacobian $\mathbf{J}$, gives us new conditions which are sufficient for equivalence. That is, $(\mathbf{V}_1, \mathbf{\Pi}_1) \equiv (\mathbf{V}_2, \mathbf{\Pi}_2)$ if, for every x, y:

$$p(y|\mathbf{\Pi}_2) = \frac{1}{|\det(\mathbf{J})|} p(\phi^{-1}(y)|\mathbf{\Pi}_1), \tag{4.11}$$

$$p(x|y, \mathbf{V}_2) = p(x|\phi^{-1}(y), \mathbf{V}_1). \tag{4.12}$$

4.1 The minimum-divergence step

Here we use hidden-variable transformation to effect a minimum-divergence step, which first finds a non-standard prior which minimizes (2.5) and then adapts to equivalent parameters with standard prior. The recipe starts with the hidden-variable posterior $p(y|x, \boldsymbol{\lambda}_0)$ and some parameter[3] $\boldsymbol{\lambda}_1 = (\mathbf{V}_1, \bar{\mathbf{\Pi}})$

[3]This notation accommodates both strategy 3.2, for which $\mathbf{V}_1 = \mathbf{V}_0$, and strategy 3.3, for which $\mathbf{V}_1 = \arg\max \tilde{\mathcal{Q}}(\boldsymbol{\lambda}_0, \mathbf{V})$.

such that $\mathcal{Q}(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1) \geq \mathcal{Q}(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_0)$. Then the following steps are performed in order:

1. Find $\boldsymbol{\Pi}_1$, where
$$\boldsymbol{\Pi}_1 = \arg\min_{\boldsymbol{\Pi}} D\big(\boldsymbol{\lambda}_0 \| \boldsymbol{\Pi}\big), \tag{4.13}$$
2. Next, find a bijection $\phi : \mathcal{Y} \mapsto \mathcal{Y}$, with Jacobian $\mathbf{J}$, such that for every y,
$$p(y|\bar{\boldsymbol{\Pi}}) = \frac{1}{|\det(\mathbf{J})|} p(\phi^{-1}(y)|\boldsymbol{\Pi}_1) \tag{4.14}$$
3. and then find $\mathbf{V}_2$, such that for every x, y,
$$p(x|y, \mathbf{V}_2) = p(x|\phi^{-1}(y), \mathbf{V}_1). \tag{4.15}$$
4. Choose as the result of the M-step $\boldsymbol{\lambda} = (\mathbf{V}_2, \bar{\boldsymbol{\Pi}})$.

The result of the minimum-divergence step can be analyzed as follows. Step 1 ensures that $D\big(\boldsymbol{\lambda}_0 \| \boldsymbol{\Pi}_1\big) \leq D\big(\boldsymbol{\lambda}_0 \| \bar{\boldsymbol{\Pi}}\big)$, so that if we choose $\boldsymbol{\lambda}_2 = (\mathbf{V}_1, \boldsymbol{\Pi}_1)$ and apply (1.9) and (2.3), then we have:

$$\mathcal{L}(\boldsymbol{\lambda}_2) \geq \mathcal{Q}(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_2) \geq \mathcal{Q}(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1) \geq \mathcal{Q}(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_0) = \mathcal{L}(\boldsymbol{\lambda}_0). \tag{4.16}$$

Finally, steps 2 and 3 ensure $\mathcal{L}(\boldsymbol{\lambda}) = \mathcal{L}(\boldsymbol{\lambda}_2)$, so that progress is made relative to the initial parameter λ_0:

$$\mathcal{L}(\boldsymbol{\lambda}) \geq \mathcal{L}(\boldsymbol{\lambda}_0). \tag{4.17}$$

This progress will be strict if $D\big(\boldsymbol{\lambda}_0 \| \boldsymbol{\Pi}_1\big) < D\big(\boldsymbol{\lambda}_0 \| \bar{\boldsymbol{\Pi}}\big)$.

4.1.1 Analysis of strategy 3

If we are doing strategy 2, then $\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_0$ and there is nothing further to analyze. If however, we are using strategy 3, then $\boldsymbol{\lambda}_1$ is the result of the standard M-step, where the auxiliary has already been improved. In this case, the above analysis does *not* yet prove that the combined M-steps are better, in other words that $\mathcal{L}(\boldsymbol{\lambda}) \geq \mathcal{L}(\boldsymbol{\lambda}_1)$.

However, if $\boldsymbol{\lambda}_0$ is in a saddle point, where all first order derivatives of $\mathcal{L}(\mathbf{V}_0, \bar{\boldsymbol{\Pi}})$ and therefore also of $\tilde{\mathcal{Q}}(\lambda_0, \mathbf{V})$, w.r.t. $\mathbf{V}$ are zero, then the standard M-step, which typically uses only first-order derivatives, will make no progress, so that $\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_0$. In this case, if the derivatives w.r.t. the prior parameters are not also zero, then the minimum-divergence step can still make progress: $\mathcal{L}(\boldsymbol{\lambda}) \geq \mathcal{L}(\boldsymbol{\lambda}_1) = \mathcal{L}(\boldsymbol{\lambda}_0)$.