Quick Refresher on EM Algorithm

Shoichiro Yamanishi

March 31, 2020

Abstract

This is a personal notes as my own memory aid on EM for my future-self to avoid going through the books trying to find the relevant parts and re-read them, which is time consuming. This is also a summplemental material to explain why the maximization step works for MLE with i.i.d. samples, which was not obvious to me in the text books.

PRML[1] Chap 9.4. explains the EM algorithm with the standard graph plot of lower bounds and θ along the horizontal axis. Computer Vision by JD Prince[2] Chap 7.3., also has a better and more succinct explanation of EM.

The purpose of this document is to explain the EM algorithm without significant gap throught the course of deductions at the cost of lengthiness, and to explain why the M-Step works for MLE, which is omitted in the text books.

1 EM-Algorithm

EM-Algorithm is applicable for the following case.

- you have samples $\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_N}$, that are i.i.d.
- you want to model a probability distribution over $\mathbf{x}$ with latent variables $\mathbf{z}$ and parameters like $\int p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})d\mathbf{z}$.
- you want MLE. i.e. $\underset{\boldsymbol{\theta}}{\text{argmax}} \left(\Sigma_{i=1}^{N} \ln \int p(\mathbf{x}_i, \mathbf{z}|\boldsymbol{\theta})d\mathbf{z}\right)$
- $\int p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})d\mathbf{z}$ is intractable.
- $\nabla_{\boldsymbol{\theta}} \left(\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})]\right)$ is tractable for some probability distribution $q(\mathbf{z})$.

Usually $\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})]$ is represented with some sufficient statistics of $\mathbf{z}$ such as moments $\mathbb{E}[\mathbf{z}]$ and $\mathbb{E}[\mathbf{z}\mathbf{z}^T]$. Preferably, $\mathbb{E}[p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})]$ is convex w.r.t. $\boldsymbol{\theta}$, and $\nabla_{\boldsymbol{\theta}}(\mathbb{E}[p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})])$ is analytically in closed form so that $\nabla_{\boldsymbol{\theta}}(\mathbb{E}[p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})]) = 0$ can be used.

It is difficult to directly optimize $\boldsymbol{\theta}$ but we can introduce a distribution $q(\mathbf{z})$ over the latent variable $\mathbf{z}$ to guide $\boldsymbol{\theta}$ toward optimization. This is basically a **coordinate descent algorithm** that alternates between the variational space of

$q(\mathbf{z})$ and the parameter space of $\boldsymbol{\theta}$, which are mutually independent conditioned on $\mathbf{x}$, since $q(\mathbf{z})$ is arbitrary.

1.1 E-Step

This step is basically a coordinate descent in the variational space of $q(\mathbf{z})$. In the variational space, the optimum condition is derived by the following argument. Consider the following.

$$\ln p(\mathbf{x}|\boldsymbol{\theta}) = \ln(\int p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) d\mathbf{z})$$

We arbitrarily introduce a probability density function $0 < q(\mathbf{z}) < \infty$.

$$\ln p(\mathbf{x}|\boldsymbol{\theta}) = \ln(\int q(\mathbf{z}) \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} d\mathbf{z})$$

By Jensen's inequality, $\mathbb{E}[f(x)] \leq f(\mathbb{E}[x])$ for any concave function $f(x)$, and $\ln(x)$ is a concave function. Therefore, $\ln(\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[g(\mathbf{z}])) \geq \mathbb{E}[ln(g(\mathbf{z}))]$ and then $\ln(\int q(\mathbf{z}) g(\mathbf{z}) d\mathbf{z}) \geq \int q(\mathbf{z}) \ln g(\mathbf{z}) d\mathbf{z}$.

$$\begin{aligned}
\ln p(\mathbf{x}|\boldsymbol{\theta}) &\geq \int q(\mathbf{z}) \ln \left(\frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} \right) d\mathbf{z} = \mathcal{L}(q(\mathbf{z}), \boldsymbol{\theta}) \\
&= \int q(\mathbf{z}) \ln \left(\frac{p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta})}{q(\mathbf{z})} \right) d\mathbf{z} \qquad (1) \\
&= \int q(\mathbf{z}) \ln p(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{z} + \int q(\mathbf{z}) \ln \left(\frac{p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})}{q(\mathbf{z})} \right) d\mathbf{z} \\
&= \ln p(\mathbf{x}|\boldsymbol{\theta}) - D_{KL}(q(\mathbf{z}) || p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}))
\end{aligned}$$

$\mathcal{L}(q(\mathbf{z}), \boldsymbol{\theta})$ is ELBO (Evidence Lower Bound). If $(q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}))$ then the KL divergence above vanishes, and hence the ELBO is maximal. Therfore for a fixed $\mathbf{x}_i$ and $\boldsymbol{\theta}^*$,

$$\begin{aligned}
\underset{q(\mathbf{z})}{\operatorname{argmax}}(ELBO) &= \ln p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta}^*) \\
\max_{q(\mathbf{z})}(ELBO) &= \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta})}[\ln p(\mathbf{x}_i, \mathbf{z}|\boldsymbol{\theta}) - \ln p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta})] \\
&= \ln p(\mathbf{x}_i|\boldsymbol{\theta}^*)
\end{aligned} \qquad (2)$$

1.2 M-Step

Now, assume $q(\mathbf{z})$ is fixed at $p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta}^*)$ in $\mathcal{L}(q(\mathbf{z}), \boldsymbol{\theta})$. Can we further raise the evidence lower bound by changing $\boldsymbol{\theta}$? I.e., is there $\boldsymbol{\theta}$ such that:

$$\Sigma_{i=1}^{N}\left(\ln\left(\int p(\mathbf{x}_i, \mathbf{z}|\boldsymbol{\theta})d\mathbf{z}\right)\right) > \Sigma_{i=1}^{N}\left(\ln\left(\int p(\mathbf{x}_i, \mathbf{z}|\boldsymbol{\theta}^*)d\mathbf{z}\right)\right)$$

The following argument says yes. From equation 1, for each $\mathbf{x}_i$,

$$\mathcal{L}(q(\mathbf{z}), \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}\left[\ln p(\mathbf{x}_i, \mathbf{z}|\boldsymbol{\theta}) - \ln p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta})\right] \tag{3}$$

We fix $q(\mathbf{z})$ at $p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta}^*)$.

$$\mathcal{L}(p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta}^*), \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta}^*)}\left[\ln p(\mathbf{x}_i, \mathbf{z}|\boldsymbol{\theta}) - \ln p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta})\right] \tag{4}$$

If we also fix $\boldsymbol{\theta}$ at $\boldsymbol{\theta}^*$, then

$$\mathcal{L}(p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta}^*), \boldsymbol{\theta}^*) = \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta}^*)}\left[\ln p(\mathbf{x}_i, \mathbf{z}|\boldsymbol{\theta}^*) - \ln p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta}^*)\right] \tag{5}$$

Subtract equation 5 from 4, then

$$\begin{aligned}
&\mathcal{L}(p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta}^*), \boldsymbol{\theta}) - \mathcal{L}(p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta}^*), \boldsymbol{\theta}^*) \\
&= \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta}^*)}\left[\ln p(\mathbf{x}_i, \mathbf{z}|\boldsymbol{\theta}) - \ln p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta})\right] - \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta}^*)}\left[\ln p(\mathbf{x}_i, \mathbf{z}|\boldsymbol{\theta}^*) - \ln p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta}^*)\right] \\
&= \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta}^*)}\left[\ln p(\mathbf{x}_i, \mathbf{z}|\boldsymbol{\theta}) - \ln p(\mathbf{x}_i, \mathbf{z}|\boldsymbol{\theta}^*)\right] + \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta}^*)}\left[\ln p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta}^*) - \ln p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta})\right] \\
&= \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta}^*)}\left[\ln p(\mathbf{x}_i, \mathbf{z}|\boldsymbol{\theta}) - \ln p(\mathbf{x}_i, \mathbf{z}|\boldsymbol{\theta}^*)\right] + D_{KL}\left(p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta}^*)||p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta})\right)
\end{aligned} \tag{6}$$

This implies

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta}^*)}\left[\ln p(\mathbf{x}_i, \mathbf{z}|\boldsymbol{\theta})\right] \geq \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta}^*)}\left[\ln p(\mathbf{x}_i, \mathbf{z}|\boldsymbol{\theta}^*)\right] \\
\Rightarrow\ &\mathcal{L}(p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta}^*), \boldsymbol{\theta}) \geq \mathcal{L}(p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta}^*), \boldsymbol{\theta}^*)
\end{aligned} \tag{7}$$

and this holds for each $\mathbf{x}_i$. Therefore, in the MLE setting,

$$\begin{aligned}
&\Sigma_{i=1}^{N}\left(\mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta}^*)}\left[\ln p(\mathbf{x}_i, \mathbf{z}|\boldsymbol{\theta})\right] - \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta}^*)}\left[\ln p(\mathbf{x}_i, \mathbf{z}|\boldsymbol{\theta}^*)\right]\right) \leq 0 \\
\Rightarrow\ &\ln p(\mathbf{x}|\boldsymbol{\theta}) \geq \ln p(\mathbf{x}|\boldsymbol{\theta}^*)
\end{aligned} \tag{8}$$

Taking the argmax on the RHS of equation 8 corresponds to the coordinate descent in the parameter space $\boldsymbol{\theta}$.

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007.

[2] Simon J. D. Prince. *Computer Vision: Models, Learning, and Inference*. Cambridge University Press, 2012.