A minimalist's exposition of EM

Karl Stratos

1 What EM optimizes

Let O, H be a random variables representing the space of samples. Let θ be the parameter of a generative model with an associated probability function $P(O, H|\theta)$. In many cases, it is easy to find the maximum-likelihood estimate (MLE) solution if both O, H are observed:

$$\theta^* = \arg\max_\theta \quad \log P(O, H|\theta) \tag{1}$$

This is largely due to the log operation directly on $P(O, H|\theta)$. For instance, if $P(O, H|\theta)$ is a member of the exponential family, the log application results in a substantially simpler expression (e.g., is smooth and concave in θ).

However, suppose we are only given *partial* samples O. The MLE solution is now

$$\theta^* = \arg\max_\theta \quad \log \sum_H P(O, H|\theta) \tag{2}$$

A critical change is that log is no longer directly on the distribution. As a result, this objective is often difficult to optimize. By convention, we will write $\log P(O|\theta)$ to refer to $\log \sum_H P(O, H|\theta)$.

The Expectation-Maximization (EM) algorithm allows us to optimize a significantly simpler objective which nevertheless is guaranteed to improve $\log P(O|\theta)$:

```
EM
Input: model $P(O, H|\theta)$, partial samples $O$, number of iterations $T$
Output: (not necessarily exact in the limit) estimation of (2)

  • Initialize $\theta^0$ (e.g., randomly).
  • For $t = 1 \ldots T$:
      – E-step: Calculate $P(H|O, \theta^{t-1})$.
      – M-step: $\theta^t \leftarrow \arg\max_\theta Q(\theta, \theta^{t-1})$ where
            $Q(\theta, \theta^{t-1}) := \sum_H P(H|O, \theta^{t-1}) \log P(O, H|\theta)$
  • Return $\theta^T$.
```

Optimizing $Q(\theta, \theta')$ over θ is typically much easier since we have log back on $P(O, H|\theta)$. Thus we have reduced a difficult problem into a series of easy problems.

1.1 Proof that EM converges

At a first glance, EM does not seem to be optimizing $\log P(O|\theta)$. But we can show that it always increases this value unless it has reached a stationary point. To show this, let q be any distribution over H and define the following quantity:

$$L(q, \theta) := \sum_H q(H) \log \frac{P(O, H|\theta)}{q(H)}$$

First, we show that the target objective $\log P(O|\theta)$ is lower bounded by $L(q, \theta)$ for any choice of q.

Lemma 1.1. $\log P(O|\theta) = L(q, \theta) + KL\{q||P(H|O, \theta)\}$ *for any* q*, where*

$$KL\{q||P(H|O, \theta)\} := \sum_H q(H) \log \frac{q(H)}{P(H|O, \theta)} \geq 0$$

is the Kullback-Leibler (KL) divergence between $q(H)$ *and* $P(H|O, \theta)$*.*

Proof. It can be easily verified by using $P(O, H|\theta) = P(H|O, \theta)P(O|\theta)$ in $L(q, \theta)$. □

Next, we claim that for a fixed θ, the lower bound $L(q, \theta)$ can be tightened to $\log P(O|\theta)$ by choosing $q(H) = P(H|O, \theta)$.

Lemma 1.2.

$$\max_q L(q, \theta) = \log P(O|\theta)$$
$$\arg\max_q L(q, \theta) = P(H|O, \theta)$$

Proof. By Lemma 1.1, we have $L(q, \theta) + KL\{q||P(H|O, \theta)\} = C(\theta)$ where $C(\theta)$ is constant in q. This means the maximizing q for $L(q, \theta)$ is one that has zero KL divergence with $P(H|O, \theta)$. This gives the desired result. □

The final lemma shows that when $q(H) = P(H|O, \bar{\theta})$ is fixed, maximizing the lower bound $L(q, \theta)$ over θ is equivalent to maximizing the $Q(\theta, \bar{\theta})$ over θ.

Lemma 1.3. *When* $q(H)$ *is fixed as* $P(H|O, \bar{\theta})$*,*

$$\arg\max_\theta L(q, \theta) = \arg\max_\theta Q(\theta, \bar{\theta})$$

Proof. Plugging in $q(H) = P(H|O, \bar{\theta})$ inside $L(q, \theta)$, we see the desired result since

$$L(q, \theta) = \underbrace{\sum_H P\left(H|O, \bar{\theta}\right) \log P(H|O, \theta)}_{Q(\theta, \bar{\theta})} - \underbrace{\sum_H P\left(H|O, \bar{\theta}\right) \log P\left(H|O, \bar{\theta}\right)}_{\text{independent of } \theta}$$

□

These lemmas together yield the following statement.

Theorem 1.4. *In **EM**, we have* $\log P(O|\theta^t) \geq \log P(O|\theta^{t-1})$ *with equality iff* θ^{t-1} *is a stationary point in* $\log P(O|\theta)$*.*

Proof. In the t-th iteration, we fix $P(H|O, \theta^{t-1})$ to compute $\theta^t = \arg\max_\theta Q(\theta, \theta^{t-1})$. Then $\theta^t = \arg\max_\theta L(q', \theta)$ where $q'(H) = P(H|O, \theta^{t-1})$ by Lemma 1.3. Regarding $L(q', \theta^t)$ and $L(q', \theta^{t-1})$, we have:

1. $L(q', \theta^t) \leq \log P(O|\theta^t)$ by Lemma 1.1.
2. $L(q', \theta^t) \geq L(q', \theta)$ for any θ since $\theta^t = \arg\max_\theta L(q', \theta)$.
3. $L(q', \theta^{t-1}) = \log P(O|\theta^{t-1})$ by Lemma 1.2 since $q'(H) = P(H|O, \theta^{t-1})$.

These observations give us:

$$\log P(O|\theta^t) \geq L(q', \theta^t) \geq L(q', \theta^{t-1}) = \log P(O|\theta^{t-1})$$

To see why $L(q', \theta^t) > L(q', \theta^{t-1})$ unless θ^{t-1} is a stationary point in $\log P(O|\theta)$, suppose θ^{t-1} is not a stationary point. Then the gradient of $\log P(O|\theta)$ at θ^{t-1} is nonzero. By Lemma 1.1 and 1.2, we see that

$$\frac{\partial}{\partial \theta^{t-1}} L(q', \theta) + \frac{\partial}{\partial \theta^{t-1}} KL\{q' || P(H|O, \theta)\} = \frac{\partial}{\partial \theta^{t-1}} L(q', \theta) \neq 0$$

Thus the gradient of $L(q', \theta)$ at θ^{t-1} is nonzero either, so $\theta^t = \arg\max_\theta L(q', \theta)$ will satisfy $L(q', \theta^t) > L(q', \theta^{t-1})$ and therefore $\log P(O|\theta^t) > \log P(O|\theta^{t-1})$. □

EM can be viewed as an (unconventional) alternating optimization algorithm that iteratively improves the *lower bound* of the objective instead of the objective itself. It repeats the following two steps:

- **E-step**: $q' \leftarrow \arg\max_q L(q, \theta')$
- **M-step**: $\theta' \leftarrow \arg\max_\theta L(q', \theta)$

The E-step defines a new (tight) lower bound function that is tangent to $\log P(O|\theta)$ at θ'. The M-step modifies θ' to optimize this lower bound, in the process making the lower bound loose. The argument about strict improvement can be visualized as follows: if θ' is not a stationary point in $\log P(O|\theta)$, then it cannot be a stationary point in $L(q, \theta)$ which is tangent to $\log P(O|\theta)$ and shares the gradient at θ'.

2 Examples of EM

2.1 Gaussian mixture model (GMM)

Consider a mixture of m univariate spherical Gaussians $\theta = \{(\gamma_j, \mu_j)\}_{j=1}^m$: the j-th Gaussian has mean μ_j and standard deviation 1, and is associated with a prior probability γ_j of being selected.

Assume we have n iid. samples from this model. Complete samples would have the form $X = \{(o^{(i)}, h^{(i)})\}_{i=1}^n$ where $o^{(i)}$ is generated by the $h^{(i)}$-th Gaussian: in that case, solving (1) is trivial. However, assume instead we have partial samples $O = \{o^{(i)}\}_{i=1}^n$;

we do not observe the corresponding $H = \{h^{(i)}\}_{i=1}^n$. Then we must maximize $P(O|\theta)$ over θ:

$$\sum_{i=1}^n \log \sum_{j=1}^m \frac{\gamma_j}{\sqrt{2\pi}} \exp\left(-\frac{(o^{(i)} - \mu_j)^2}{2}\right)$$

in which complicated interactions between parameters are unresolved by log. Thus we turn to EM: we optimize instead

$$Q(\theta, \theta^{t-1}) = \sum_{i=1}^n \sum_{j=1}^m p(h^{(i)} = j|o^{(i)}, \theta^{t-1}) \left(\log \gamma_j - \log 2\pi - \frac{(o^{(i)} - \mu_j)^2}{2}\right)$$

over θ. This leads to a simple formula for setting θ^t using θ^{t-1}.

2.2 Hidden Markov model (HMM)

Consider an HMM with a discrete observation space $\mathcal{X}$ and a discrete state space $\mathcal{Y}$. It is parameterized by π, t, and o and defines the probability distribution

$$p(x_1 \ldots x_N, y_1 \ldots y_N) = \pi(y_1) \times \prod_{j=1}^N o(x_j|y_j) \times \prod_{j=2}^N t(y_j|y_{j-1})$$

for a sequence pair $x_1 \ldots x_N \in \mathcal{X}^N$ and $y_1 \ldots y_N \in \mathcal{Y}^N$.

Assume we have n iid. sample sequences of length N from this model. Complete samples would have the form $X = \{(x^{(i)}, y^{(i)})\}_{i=1}^n$ where $x^{(i)} \in \mathcal{X}^N$ is an observation sequence and $y^{(i)} \in \mathcal{Y}^N$ is the corresponding state sequence. Again, solving (1) is trivial if we are given X. Assume instead we have partial samples $O = \{x^{(i)}\}_{i=1}^n$. It is unclear how to maximize $P(O|\theta)$ over θ:

$$\sum_{i=1}^n \log \sum_{y \in \mathcal{Y}^N} \pi(y_1) \times \prod_{j=1}^N o(x_j^{(i)}|y_j) \times \prod_{j=2}^N t(y_j|y_{j-1})$$

In comparison, we can optimize a substantially simpler objective using EM:

$$Q(\theta, \theta^{t-1}) = \sum_{i=1}^n \sum_{y \in \mathcal{Y}^N} p(y|o^{(i)}, \theta^{t-1}) \left(\log \pi(y_1) + \sum_{j=1}^N \log o(x_j^{(i)}|y_j) + \sum_{j=2}^N \log t(y_j|y_{j-1})\right)$$

The sum over the elements of $\mathcal{Y}^N$ can be achieved with dynamic programming.

3 Discussion

Despite the widespread use of EM today, it is often heuristically understood. Especially, it is often mistaken as plugging in conditional expectations in place of unobserved values in (1). This is in general incorrect (Flury and Zoppe, 2000), even though it is the case for a wide class of practical models (Appendix). It is important to understand EM from an optimization point of view to avoid this pitfall.

Reference: Bishop, Christopher M. Pattern recognition and machine learning (2006).

4 Appendix: trick for categorical distributions

Whenever $P(O, H|\theta)$ is a *categorical distribution*, i.e., it defines a distribution over counts of events as a product of parameters associated with the events, we can actually plug in conditional expectations in place of unobserved values in (1) and solve that problem. This makes the derivation of EM almost trivial. Since many useful models in the world are categorical (e.g., HMMs), this trick can be often handy.

To make this concrete, suppose we have a categorical distribution over $n + m$ events in which n events are observed and m events are unobserved. Let o_i be the count of the i-th observed event and h_j the count of the j-th unobserved event. The model defines a probability distribution as:

$$p(o_1 \dots o_n, h_1 \dots h_m|\alpha, \beta) = \prod_{i=1}^{n} \alpha_i^{o_i} \times \prod_{j=1}^{m} \beta_j^{h_j}$$

where $\theta = (\alpha, \beta)$ is the model parameter. We will now examine the form of $Q(\theta, \bar{\theta})$ ($\bar{\theta}$ is fixed):

$$\begin{aligned}
Q(\theta, \bar{\theta}) &:= \sum_{h_1 \dots h_m} p(h_1 \dots h_m|o_1 \dots o_n, \bar{\theta}) \left(\sum_{i=1}^{n} o_i \log \alpha_i + \sum_{j=1}^{m} h_j \log \beta_j \right) \\
&= \sum_{i=1}^{n} o_i \log \alpha_i + \sum_{h_1 \dots h_m} p(h_1 \dots h_m|o_1 \dots o_n, \bar{\theta}) \sum_{j=1}^{m} h_j \log \beta_j \\
&= \sum_{i=1}^{n} o_i \log \alpha_i + \sum_{j=1}^{m} \sum_{h_j} p(h_j|o_1 \dots o_n, \bar{\theta}) h_j \log \beta_j \\
&= \sum_{i=1}^{n} o_i \log \alpha_i + \sum_{j=1}^{m} \hat{h}_j(\bar{\theta}) \log \beta_j
\end{aligned}$$

where $\hat{h}_j(\bar{\theta}) = \sum_{h_j} p(h_j|o_1 \dots o_n, \bar{\theta}) h_j$ is the expected count of the j-th unobserved event under $\bar{\theta}$. Compare this to the setting in which $h_1 \dots h_m$ are fully observed and we maximize:

$$\log p(o_1 \dots o_n, h_1 \dots h_m|\alpha, \beta) = \sum_{i=1}^{n} o_i \log \alpha_i + \sum_{j=1}^{m} h_j \log \beta_j$$

We see that this is exactly the same as $Q(\theta, \bar{\theta})$ except that $\hat{h}_j(\bar{\theta})$ is switched with h_j. Hence in this particular setup, each iteration of EM amounts to solving the fully observed MLE estimation in (1) but with unobserved values replaced by their corresponding conditional expections.