Expectation Maximization Algorithm

CSE, UNSW

April 14, 2021

Motivation

- Missing data
- Latent variable
- Easier optimization

Convex Function

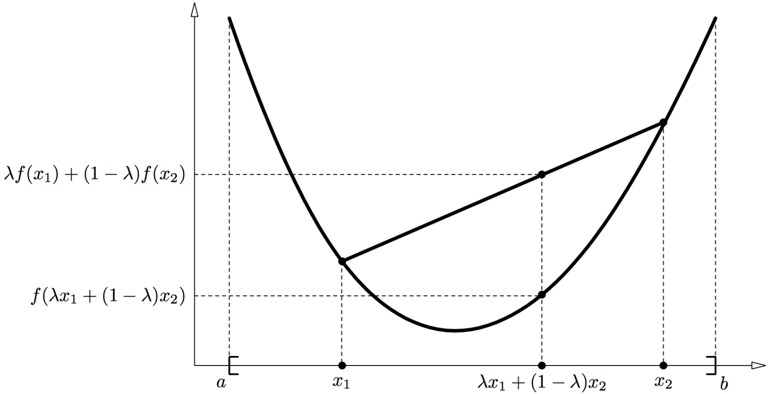


Figure 1: f is *convex* on $[a,b]$ if $f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$ $\forall x_1, x_2 \in [a,b], \quad \lambda \in [0,1]$.

- e.g., $-\log(x)$
- An important concept in optimization / machine learning.

Jensen's Inequality

- Let f be a convex function defined on an interval I. If $\{x_i\}_{i=1}^n \in I$ and $\{\lambda_i\}_{i=1}^n \geq 0$ with $\sum_i \lambda_i = 1$, then

$$f\left(\sum_{i=1}^{n} \lambda_i x_i\right) \leq \sum_{i=1}^{n} \lambda_i f(x_i)$$

  The equality holds iff $x_1 = x_2 = \ldots = x_n$ or f is linear.

  - Corollary: Since $\ln(x)$ is a concave function (i.e., $-\ln(x)$ is a convex function), then

$$\ln\left(\sum_{i=1}^{n} \lambda_i f(x_i)\right) \geq \sum_{i=1}^{n} \lambda_i \ln\left(f(x_i)\right)$$

    In addition, the equality holds iff $f(x_i)$ is a constant.

Log Likelihood

- Define log likelihood function $L(\theta) = \ln \Pr\{\mathrm{x} \mid \theta\}$. For i.i.d. examples, $L(\theta) = \sum_i L^{(i)}(\theta) = \sum_i \ln \Pr\left\{x^{(i)} \mid \theta\right\}$.
  - Goal: find θ^* that maximizes the log likelihood.
- What if the model contains latent variable $\mathrm{z} = [\mathrm{z}^{(i)}]_i$ (whose value is unknown)?

$$
\begin{aligned}
L^{(i)}(\theta) &\stackrel{\text{def}}{=} \ln \Pr\left\{x^{(i)} \mid \theta\right\} = \ln \sum\nolimits_{z^{(i)}} \Pr\left\{x^{(i)}, z^{(i)} \mid \theta\right\} \\
&= \ln \sum\nolimits_{z^{(i)}} q(z^{(i)}) \cdot \frac{\Pr\{x^{(i)}, z^{(i)} \mid \theta\}}{q(z^{(i)})} \qquad (\dagger) \\
&\geq \sum\nolimits_{z^{(i)}} q(z^{(i)}) \cdot \ln \frac{\Pr\{x^{(i)}, z^{(i)} \mid \theta\}}{q(z^{(i)})}
\end{aligned}
$$

  - If $q(z^{(i)}) = \Pr\left\{z^{(i)} \mid x^{(i)}, \theta\right\}$, then the equality holds

- Given the current parameter $\theta_{\text{(old)}}$, and let

$$q_{\text{(old)}}(z^{(i)}) \stackrel{\text{def}}{=} \Pr\left\{z^{(i)} \mid x^{(i)}, \theta_{\text{(old)}}\right\}$$

$$L^{(i)}(\theta) = \ln\left(\sum_{z^{(i)}} q_{\text{(old)}}(z^{(i)}) \frac{\Pr\left\{x^{(i)}, z^{(i)} \mid \theta\right\}}{q_{\text{(old)}}(z^{(i)})}\right) \tag{$\dagger$}$$

$$\geq \sum_{z^{(i)}} q_{\text{(old)}}(z^{(i)}) \ln\left(\frac{\Pr\left\{x^{(i)}, z^{(i)} \mid \theta\right\}}{q_{\text{(old)}}(z^{(i)})}\right)$$

$$= \sum_{z^{(i)}} \Pr\left\{z^{(i)} \mid x^{(i)}, \theta_{\text{(old)}}\right\} \ln\left(\Pr\left\{x^{(i)}, z^{(i)} \mid \theta\right\}\right)$$

$$- \sum_{z^{(i)}} \Pr\left\{z^{(i)} \mid x^{(i)}, \theta_{\text{(old)}}\right\} \ln\left(\Pr\left\{z^{(i)} \mid x^{(i)}, \theta_{\text{(old)}}\right\}\right)$$

$$= \underbrace{\sum_{z^{(i)}} \Pr\left\{z^{(i)} \mid x^{(i)}, \theta_{\text{(old)}}\right\} \ln\left(\Pr\left\{x^{(i)}, z^{(i)} \mid \theta\right\}\right)}_{\stackrel{\text{def}}{=} Q^{(i)}(\theta, \theta_{\text{(old)}})} + \underbrace{C}_{constant, entropy(q)}$$

Hence, the EM algorithm iterates the following two steps:

- **[E-step]**: Compute the $q_{(\mathrm{old})}(z^{(i)}) = \Pr\{z^{(i)} \mid x^{(i)}, \theta_{(\mathrm{old})}\}$
- **[M-step]**: Find θ that maximizes the function $Q(\theta, \theta_{(\mathrm{old})})$ (see above (just sum over i).

Alternative interpretation:

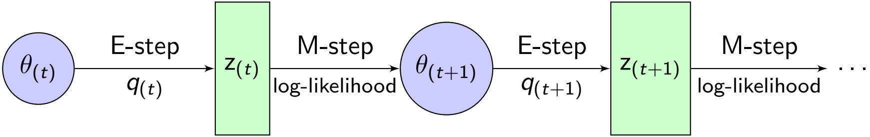
$$Q^{(i)}(\theta, \theta_{(\mathrm{old})}) \stackrel{\mathrm{def}}{=} \sum_{z^{(i)}} \Pr\left\{z^{(i)} \mid x^{(i)}, \theta_{(\mathrm{old})}\right\} \ln\left(\Pr\left\{x^{(i)}, z^{(i)} \mid \theta\right\}\right)$$

$$= \mathrm{E}_{\mathrm{z^{(i)}} \sim \mathbf{q}_{(\mathrm{old})}(\mathrm{z^{(i)}})}\left[\!\left[\ln \Pr\left\{x^{(i)}, z^{(i)} \mid \theta\right\}\right]\!\right]$$

i.e., the expected complete log-likelihood (function)

- Sample z from the *proposal distribution* q
- Then it is easy to compute the complete log-likelihood
- Do this for every possible z

Illustration



How EM converges

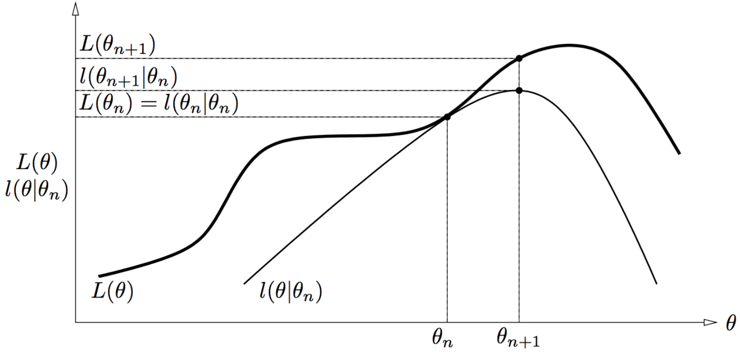


Figure 2: Graphical interpretation of a single iteration of the EM algorithm: The function $L(\theta|\theta_n)$ is upper-bounded by the likelihood function $L(\theta)$. The functions are equal at $\theta = \theta_n$. The EM algorithm chooses θ_{n+1} as the value of θ for which $l(\theta|\theta_n)$ is a maximum. Since $L(\theta) \geq l(\theta|\theta_n)$ increasing $l(\theta|\theta_n)$ ensures that the value of the likelihood function $L(\theta)$ is increased at each step.

Example 1: Three Coins

- Given three coins: z, a, and b, with head probabilities π, α, and β, respectively.
- Generative process: if toss(z) $==$ head, return(toss(a)); else return(toss(b)).
- Observed data $\mathrm{x} = [1, 1, 0, 1, 0, 0, 1, 0, 1, 1]$.
- Goal: estimate the parameters
- The usual assumption: all tosses are i.i.d.

If we know $\{ z^{(i)} \}_{i=1}^{10}$

Observed data:

$z^{(i)}$	1	0	1	1	1	0	0	1	0	0
coin $\to x^{(i)}$	a	b	a	a	a	b	b	a	b	b
$x^{(i)}$	1	1	0	1	0	0	1	0	1	1

$\pi_{\text{MLE}} =$ $\alpha_{\text{MLE}} =$ $\beta_{\text{MLE}} =$

Solution /1

- Problem setup!:
  - $\theta = ?$
  - Missing data (i.e., z) = ?
    - **Complete** likelihood (for a single item): $\Pr\{x_i, z_i \mid \theta\}$ (change of notation henceforth)
- The E-step: Given current θ_t, we can determine the distribution q

$$\mu_{i,t} \stackrel{\text{def}}{=} \Pr\{z_i = 1 \mid x_i, \theta_t\} = \frac{\Pr\{z_i = 1, x_i \mid \theta_t\}}{\Pr\{x_i \mid \theta_t\}}$$

$$= \frac{\pi_t \alpha_t^{x_i}(1-\alpha_t)^{1-x_i}}{\pi_t \alpha_t^{x_i}(1-\alpha_t)^{1-x_i} + (1-\pi_t)\beta_t^{x_i}(1-\beta_t)^{1-x_i}}$$

- Numerator:
  - $= \Pr\{z_i = 1 \mid \theta_t\} \cdot \Pr\{x_i \mid z_i = 1, \theta_t\}$
  - typical trick to write the piece-wise function for the likelihood.
- Denominator: sum over $z_i = 1$ and $z_i = 0$.

Compute $Q(\theta, \theta_{(\text{old})})$

- First

$$\ln(\Pr\{x_i, z_i \mid \theta\}) = \ln\left(\pi[\alpha^{x_i}(1-\alpha)^{1-x_i}]^{z_i} \cdot [(1-\pi)\beta^{x_i}(1-\beta)^{1-x_i}]^{1-z_i}\right)$$
$$= \ln\pi + z_i \cdot (x_i \ln\alpha + (1-x_i)\ln(1-\alpha)) +$$
$$(1-z_i)\cdot(x_i \ln\beta + (1-x_i)\ln(1-\beta))$$

- Then:

$$Q = \sum_i \sum_{z_i} q(z_i) \ln(\Pr\{x_i, z_i \mid \theta_t\})$$
$$= \sum_i \left(\mu_{i,t} \ln(\Pr\{x_i, z_i = 1 \mid \theta_t\}) + (1-\mu_{i,t}) \ln(\Pr\{x_i, z_i = 0 \mid \theta_t\})\right)$$

- The M-step:

$$\frac{\partial Q(\theta \mid \theta_t)}{\partial \pi} = 0 \Longrightarrow \pi_{t+1} = \frac{1}{n}\sum_i \mu_{i,t}$$
$$\frac{\partial Q(\theta \mid \theta_t)}{\partial \alpha} = 0 \Longrightarrow \alpha_{t+1} = \frac{\sum_i \mu_{i,t} x_i}{\sum_i \mu_{i,t}}$$
$$\frac{\partial Q(\theta \mid \theta_t)}{\partial \beta} = 0 \Longrightarrow \beta_{t+1} = \frac{\sum_i (1-\mu_{i,t}) x_i}{\sum_i (1-\mu_{i,t})}$$

Understanding the Equations

$$\frac{\partial Q(\theta \mid \theta_t)}{\partial \pi} = 0 \Longrightarrow \pi_{t+1} = \frac{1}{n} \sum_i \mu_{i,t}$$

$$\frac{\partial Q(\theta \mid \theta_t)}{\partial \alpha} = 0 \Longrightarrow \alpha_{t+1} = \frac{\sum_i \mu_{i,t} x_i}{\sum_i \mu_{i,t}}$$

$$\frac{\partial Q(\theta \mid \theta_t)}{\partial \beta} = 0 \Longrightarrow \beta_{t+1} = \frac{\sum_i (1 - \mu_{i,t}) x_i}{\sum_i (1 - \mu_{i,t})}$$

Consider the example on the question page. In that example, we can deem that $\mu_{i,t}$ is a binary variable, i.e., $\mu_{i,t} = 1$ iff coin $z^{(i)} =$ head, or equivalent, coin a is chosen to determine $x^{(i)}$. Then one can easily verify that the MLE estimation is the same as the update rules in EM. Therefore, these rules can be deemed as a "soft" version of MLE: informally, each $x^{(i)}$ has $\mu_{i,t}$ contribution to the parameter estimation of coin a, and $(1 - \mu_{i,t})$ contribution to the parameter estimation of coin b.

Concrete Example

$$\mu_{i,t} = p(z_i = 1 \mid x_i = 1, \underbrace{\theta_t}_{\pi=0.6,\alpha=0.1,\beta=0.8})$$

$$= \frac{p(z_i = 1, x_i = 1 \mid \theta_t)}{p(x_i = 1 \mid \theta_t)}$$

$$= \frac{p(z_i = 1, x_i = 1 \mid \theta_t)}{p(z_i = 1, x_i = 1 \mid \theta_t) + p(z_i = 0, x_i = 1 \mid \theta_t)}$$

$$= \frac{0.6 \cdot 0.1}{0.6 \cdot 0.1 + 0.4 \cdot 0.8} = 0.16$$

Similarly,

$$p(z_i = 1 \mid x_i = 0, \theta_t) = \frac{0.6 \cdot 0.9}{0.6 \cdot 0.9 + 0.6 \cdot 0.2} = 0.82$$

Concrete Example /2

- How many different scenarios?

z_i	x_i	$p(z_i \mid x_i, \theta_t)$
0	0	0.18
0	1	0.84
1	0	**0.82**
1	1	**0.16**

- Observations: 6 1's and 4 0's.

$$\pi_{t+1} = \frac{1}{n} \sum_i \mu_{i,t} = \frac{0.16 \cdot 6 + 0.82 \cdot 4}{10} = 0.424$$

$$\alpha_{t+1} = \frac{\sum_i \mu_{i,t} x_i}{\sum_i \mu_{i,t}} = \frac{0.16 \cdot 6}{4.24} = 0.226$$

$$\beta_{t+1} = \frac{\sum_i (1 - \mu_{i,t}) x_i}{\sum_i (1 - \mu_{i,t})} = \frac{0.84 \cdot 6}{0.84 \cdot 6 + 0.18 \cdot 4} = 0.875$$