Principles of Parameter Estimation

Predrag Radivojac
School of Informatics, Computing, and Engineering
Bloomington, Indiana

INDIANA UNIVERSITY BLOOMINGTON

PRELIMINARIES

Given: a set of observations $\mathcal{D} = \{x_i\}_{i=1}^{n}$, $x_i \in \mathcal{X}$

Objective: find a model $\hat{f} \in \mathcal{F}$ that models the phenomenon well

Requirements:

(i) the ability to generalize well

(ii) the ability to incorporate prior knowledge and assumptions

(iii) scalability

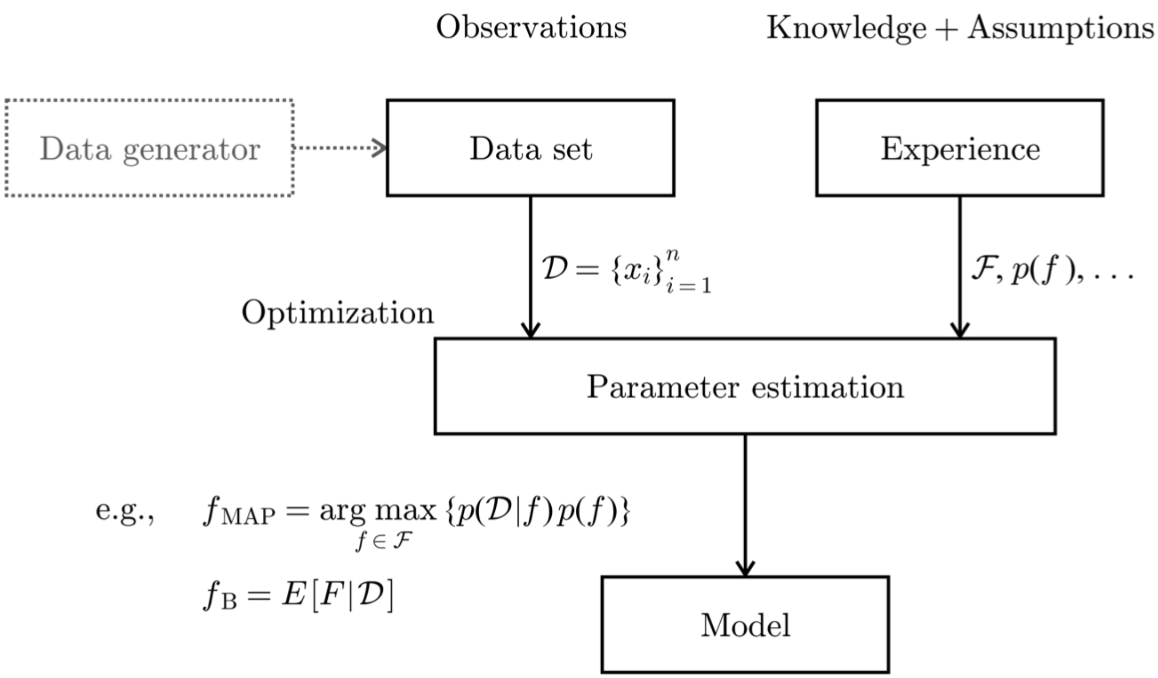
Terminology through an example: $\mathcal{D} = \{3.1, 2.4, -1.1, 0.1\}$

What is the data generator?

$\mathcal{F}$ = Gaussian(μ, σ^2), $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$

Parameter estimation

STATISTICAL FRAMEWORK



$$\textit{Model inference}: \text{Observations} + \begin{matrix}\text{Knowledge}\\ \text{and}\\ \text{Assumptions}\end{matrix} + \text{Optimization}$$

Maximum A Posteriori Inference

Idea:

$$f_{\text{MAP}} = \underset{f \in \mathcal{F}}{\arg\max} \left\{ p(f|\mathcal{D}) \right\},$$

where $p(f|\mathcal{D})$ is called the posterior distribution.

How do we calculate it?

$$p(f|\mathcal{D}) = \frac{p(\mathcal{D}|f) \cdot p(f)}{p(\mathcal{D})}$$

where $p(\mathcal{D}|f)$ = likelihood, $p(f)$ = prior, and $p(\mathcal{D})$ = data distribution.

Maximum A Posteriori Inference

Finding the data distribution:

$$p(\mathcal{D}) = \begin{cases} \sum_{f \in \mathcal{F}} p(\mathcal{D}|f)p(f) & f : \text{discrete} \\ \\ \int_{\mathcal{F}} p(\mathcal{D}|f)p(f)df & f : \text{continuous} \end{cases}$$

We can now simplify the process if we observe that

$$\begin{aligned} p(f|\mathcal{D}) &= \frac{p(\mathcal{D}|f) \cdot p(f)}{p(\mathcal{D})} \\ &\propto p(\mathcal{D}|f) \cdot p(f) \end{aligned}$$

Maximum Likelihood Inference

Express the posterior distribution as

$$p(f|\mathcal{D}) = \frac{p(\mathcal{D}|f) \cdot p(f)}{p(\mathcal{D})}$$
$$\propto p(\mathcal{D}|f) \cdot p(f)$$

Now, ignore $p(f)$ to get

$$f_{\mathrm{ML}} = \arg\max_{f \in \mathcal{F}} \{p(\mathcal{D}|f)\}$$

There are technical problems with this approach.

Examples of Maximum Likelihood Inference

Example: $\mathcal{D} = \{2, 5, 9, 5, 4, 8\}$ is an i.i.d. sample from Poisson(λ), $\lambda \in \mathbb{R}^+$

Find λ

Solution: Poisson probability mass function is $p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$

$$\lambda_{\mathrm{ML}} = \arg\max_{\lambda \in (0,\infty)} \{p(\mathcal{D}|\lambda)\}.$$

Likelihood:
$$\begin{aligned} p(\mathcal{D}|\lambda) &= p(\{x_i\}_{i=1}^n|\lambda) \\ &= \prod_{i=1}^n p(x_i|\lambda) \\ &= \frac{\lambda^{\sum_{i=1}^n x_i} \cdot e^{-n\lambda}}{\prod_{i=1}^n x_i!}. \end{aligned}$$

Examples of Maximum Likelihood Inference

Likelihood: $$p(\mathcal{D}|\lambda) = \frac{\lambda^{\sum_{i=1}^{n} x_i} \cdot e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}$$

Log-likelihood: $$ll(\mathcal{D}, \lambda) = \ln\lambda \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \ln(x_i!)$$

Optimization:

$$\frac{\partial ll(\mathcal{D}, \lambda)}{\partial \lambda} = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n$$
$$= 0$$

Solution:

$$\lambda_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$= 5.5$$

MAP and ML estimates are called the point estimates.

Examples of MAP Inference

Example: $\mathcal{D} = \{2, 5, 9, 5, 4, 8\}$ is i.i.d. sample from Poisson(λ), $\lambda \in \mathbb{R}^+$

Assume λ is taken from $\Gamma(x|k, \theta)$ with parameters $k = 3$ and $\theta = 1$

Find λ

Solution: Poisson: $p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$

Gamma: $\Gamma(x|k, \theta) = \frac{x^{k-1} e^{-\frac{x}{\theta}}}{\theta^k \Gamma(k)}$, where $x > 0$, $k > 0$, and $\theta > 0$.

Likelihood:
$$p(\mathcal{D}|\lambda) = \frac{\lambda^{\sum_{i=1}^n x_i} \cdot e^{-n\lambda}}{\prod_{i=1}^n x_i!}$$

Prior:
$$p(\lambda) = \frac{\lambda^{k-1} e^{-\frac{\lambda}{\theta}}}{\theta^k \Gamma(k)}.$$

Examples of MAP Inference

Log-likelihood:

$$\ln p(\lambda|\mathcal{D}) \propto \ln p(\mathcal{D}|\lambda) + \ln p(\lambda)$$
$$= \ln \lambda(k - 1 + \sum_{i=1}^{n} x_i) - \lambda(n + \frac{1}{\theta}) - \sum_{i=1}^{n} \ln x_i! - k \ln \theta - \ln \Gamma(k)$$

We now obtain

$$\lambda_{\mathrm{MAP}} = \frac{k - 1 + \sum_{i=1}^{n} x_i}{n + \frac{1}{\theta}}$$
$$= 5$$

Another Example

Example: $\mathcal{D} = \{x_i\}_{i=1}^n$ is i.i.d. sample from Gaussian(μ, σ^2)

Find μ and σ

Solution: Gaussian: $p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$$\mu_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \sigma^2_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu_{\mathrm{ML}}\right)^2 .$$

Relationship to Kullback-Leibler (KL) Divergence

The KL divergence between two probability distributions $p(x)$ and $q(x)$ is

$$D_{\mathrm{KL}}(p||q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx$$

Assume now the data is generated according to some $p(x|\theta_t)$. We estimated it as $p(x|\theta)$.

Let's look at the KL divergence

$$D_{\mathrm{KL}}(p(x|\theta_t)||p(x|\theta)) = \int_{-\infty}^{\infty} p(x|\theta_t) \log \frac{p(x|\theta_t)}{p(x|\theta)} dx$$

$-E\left[\log p(x|\theta)\right]$

$$= \int_{-\infty}^{\infty} p(x|\theta_t) \log \frac{1}{p(x|\theta)} dx - \int_{-\infty}^{\infty} p(x|\theta_t) \log \frac{1}{p(x|\theta_t)} dx.$$

Relationship to Kullback-Leibler (KL) Divergence

$$\frac{1}{n}\sum_{i=1}^{n}\log p(x_i|\theta) \quad \overset{a.s.}{\to} \quad E[\log p(x|\theta)]$$

when $n \to \infty$.

Conclusion:

When $n \to \infty$, ML estimation implies $p(x|\theta_{\mathrm{ML}}) = p(x|\theta_t)$

This usually implies $\theta_{\mathrm{ML}} = \theta_t$

Parameter Estimation for Mixtures of Distributions

Given: a set of observations $\mathcal{D} = \{x_i\}_{i=1}^{n}, \; x_i \in \mathcal{X}$

$$p(x|\theta) = \sum_{j=1}^{m} w_j p(x|\theta_j).$$

$$w_j \geq 0, \quad \sum_{j=1}^{m} w_j = 1.$$

where $\theta = (w_1, w_2, \ldots, w_m, \theta_1, \theta_2, \ldots, \theta_m)$

Example: Consider a mixture of $m = 2$ exponential distributions.

$p(x|\theta_j) = \lambda_j e^{-\lambda_j x}$, where $\lambda_j > 0$

$$p(x|\lambda_1, \lambda_2, w_1, w_2) = w_1 \cdot \lambda_1 e^{-\lambda_1 x} + w_2 \cdot \lambda_2 e^{-\lambda_2 x}$$

where $\lambda_1, \lambda_2 > 0$, $w_1, w_2 \geq 0$, and $w_1 = 1 - w_2$

Parameter Estimation for Mixtures of Distributions

Likelihood:

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$$
$$= \prod_{i=1}^{n} \left(\sum_{j=1}^{m} w_j p(x_i|\theta_j) \right)$$

$p(\mathcal{D}|\theta)$ has $O(m^n)$ terms. It can be calculated in $O(mn)$ time as a log-likelihood.

How can we find θ? Is there a closed-form solution?

IDEA #1

Suppose we know what data point is generated by what mixing component.

That is, $\mathcal{D}_{xy} = \{(x_i, y_i)\}_{i=1}^{n}$ is an i.i.d. sample from some distribution $p(x, y)$, where $y \in \mathcal{Y} = \{1, 2, \ldots, m\}$ specifies the mixing component.

$$
\begin{aligned}
p(\mathcal{D}_{xy}|\theta) &= \prod_{i=1}^{n} p(x_i, y_i|\theta) \\
&= \prod_{i=1}^{n} p(x_i|y_i, \theta)p(y_i|\theta) \\
&= \prod_{i=1}^{n} w_{y_i} p(x_i|\theta_{y_i}),
\end{aligned}
$$

where $w_j = P(y = j)$.

Idea #1

Log-likelihood:

$$\log p(\mathcal{D}_{xy}|\theta) = \sum_{i=1}^{n} (\log w_{y_i} + \log p(x_i|\theta_{y_i}))$$
$$= \sum_{j=1}^{m} n_j \log w_j + \sum_{i=1}^{n} \log p(x_i|\theta_{y_i}),$$

where n_j is the number of data points in $\mathcal{D}_{xy}$ generated by the j-th mixing component.

Constrained optimization: Let's first find $\mathbf{w}$

$$L(\mathbf{w}, \alpha) = \sum_{j=1}^{m} n_j \log w_j + \alpha \left(\sum_{j=1}^{m} w_j - 1 \right)$$

where α is the Lagrange multiplier.

IDEA #1

Set $\frac{\partial}{\partial w_k} L(\mathbf{w}, \alpha) = 0$ for every $k \in \mathcal{Y}$ and $\frac{\partial}{\partial \alpha} L(\mathbf{w}, \alpha) = 0$. Solve it.

It follows that $w_k = -\frac{n_k}{\alpha}$ and $\alpha = -n$.

$$w_k = \frac{1}{n} \sum_{i=1}^{n} I(y_i = k),$$

where $I(\cdot)$ is the indicator function.

To find all θ_j, we need to get concrete.

$$\frac{\partial}{\partial \lambda_k} \sum_{i=1}^{n} \log p(x_i | \lambda_{y_i}) = 0,$$

for each $k \in \mathcal{Y}$.

IDEA #1

We obtain that

$$\lambda_k = \frac{n_k}{\sum_{i=1}^{n} I(y_i = k) \cdot x_i},$$

for each $k \in \mathcal{Y}$.

Recall that

$$w_k = \frac{1}{n} \sum_{i=1}^{n} I(y_i = k)$$

If the mixing component designations $\mathbf{y}$ are known, the parameter estimation is greatly simplified.

IDEA #2

Suppose we know the θ but not the mixing componen designations..

Looks like clustering, right? Let's see. Express

$$p(\mathbf{y}|\mathcal{D},\theta) = \prod_{i=1}^{n} p(y_i|x_i,\theta)$$
$$= \prod_{i=1}^{n} \frac{w_{y_i} p(x_i|\theta_{y_i})}{\sum_{j=1}^{m} w_j p(x_i|\theta_j)}$$

and subsequently find the best configuration out of m^n possibilities.

Data is i.i.d. so y_i can be estimated separately. The MAP estimate for y_i

$$\hat{y}_i = \arg\max_{k \in \mathcal{Y}} \{p(y_i|x_i,\theta)\} \qquad p(y_i|x_i,\theta) = \frac{w_{y_i} p(x_i|\theta_{y_i})}{\sum_{j=1}^{m} w_j p(x_i|\theta_j)}$$

Classification Expectation Maximization (CEM)

1. Initialize $\lambda_k^{(0)}$ and $w_k^{(0)}$ for $\forall k \in \mathcal{Y}$
2. Set $t = 0$
3. Repeat until convergence
   (a) $\hat{y}_i^{(t)} = \arg\max_{k \in \mathcal{Y}} \{p(y_i = k | x_i, \boldsymbol{w}^{(t)}, \boldsymbol{\lambda}^{(t)})\}$ for $\forall i \in \{1, 2, \dots, n\}$
   (b) $w_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} I(\hat{y}_i^{(t)} = k)$
   (c) $\lambda_k^{(t+1)} = \frac{\sum_{i=1}^{n} I(\hat{y}_i^{(t)} = k)}{\sum_{i=1}^{n} I(\hat{y}_i^{(t)} = k) \cdot x_i}$
   (d) $t = t + 1$
4. Report $\lambda_k^{(t)}$ and $w_k^{(t)}$ for $\forall k \in \mathcal{Y}$

Contains all w_k Contains all λ_k

$$p(y_i = k | x_i, \boldsymbol{w}, \boldsymbol{\lambda}) = \frac{p(y_i = k | w_k) \cdot p(x_i | \lambda_k)}{\sum_{j=1}^{m} p(y_i = j | w_j) \cdot p(x_i | \lambda_j)} = \frac{w_k p(x_i | \lambda_k)}{\sum_{j=1}^{m} w_j p(x_i | \lambda_j)}$$

Expectation Maximization

EM

1. Initialize $\lambda_k^{(0)}$ and $w_k^{(0)}$ for $\forall k \in \mathcal{Y}$
2. Set $t = 0$
3. Repeat until convergence
   (a) $r_{ik}^{(t)} = p\big(y_i = k | x_i, \boldsymbol{w}^{(t)}, \boldsymbol{\lambda}^{(t)}\big)$
   for $\forall i \in \{1, 2, \ldots, n\}$ and $\forall k \in \mathcal{Y}$
   (b) $w_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} r_{ik}^{(t)}$
   (c) $\lambda_k^{(t+1)} = \frac{\sum_{i=1}^{n} r_{ik}^{(t)}}{\sum_{i=1}^{n} r_{ik}^{(t)} \cdot x_i}$
   (d) $t = t + 1$
4. Report $\lambda_k^{(t)}$ and $w_k^{(t)}$ for $\forall k \in \mathcal{Y}$

CEM

1. Initialize $\lambda_k^{(0)}$ and $w_k^{(0)}$ for $\forall k \in \mathcal{Y}$
2. Set $t = 0$
3. Repeat until convergence
   (a) $\hat{y}_i^{(t)} = \arg\max_{k \in \mathcal{Y}} \{p\big(y_i = k | x_i, \boldsymbol{w}^{(t)}, \boldsymbol{\lambda}^{(t)}\big)\}$
   for $\forall i \in \{1, 2, \ldots, n\}$
   (b) $w_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} I(\hat{y}_i^{(t)} = k)$
   (c) $\lambda_k^{(t+1)} = \frac{\sum_{i=1}^{n} I(\hat{y}_i^{(t)} = k)}{\sum_{i=1}^{n} I(\hat{y}_i^{(t)} = k) \cdot x_i}$
   (d) $t = t + 1$
4. Report $\lambda_k^{(t)}$ and $w_k^{(t)}$ for $\forall k \in \mathcal{Y}$

$$p(y_i = k | x_i, \boldsymbol{w}, \boldsymbol{\lambda}) = \frac{w_k p(x_i | \lambda_k)}{\sum_{j=1}^{m} w_j p(x_i | \lambda_j)}$$

EM Approach

Observable data

$\mathcal{D}_x = \{x_i\}_{i=1}^n$.

Goal: maximize $ll(\mathcal{D}_x, \theta)$

Partially observable data

$\mathcal{D}_{xy} = \{(x_i, y_i)\}_{i=1}^n$

Consider complete data log-likelihood $ll(\mathcal{D}_{xy}, \theta)$

Current Estimate

$\theta^{(t)}$

E-Step

$$Q\big(\theta|\theta^{(t)}\big) = \mathbb{E}_{\boldsymbol{y}}\big[ll(D_{xy}, \theta)|\mathcal{D}_x, \theta^{(t)}\big]$$

Expectation is taken over the distribution of random vector $\boldsymbol{y}$ (containing y_i as entries) given the observable data $\mathcal{D}_x$ and the current estimate $\theta^{(t)}$.

M-Step

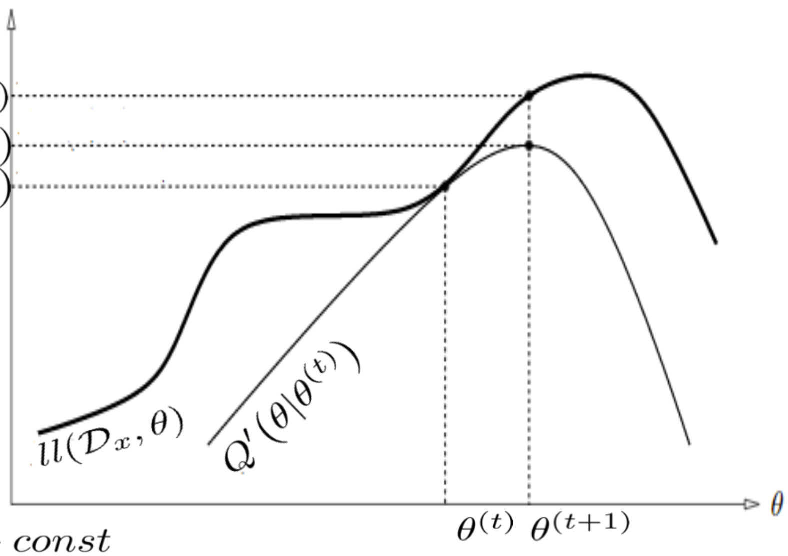
$$\theta^{(t+1)} = \arg\max_{\theta} Q\big(\theta|\theta^{(t)}\big)$$

EM framework is much more general and has many applications beyond mixture models. Hidden markov models, handling missing data, correcting sample selection bias to name a few.

EM CONVERGENCE

Converegce follows because the inequality below is true for all θ. Finding a θ that increases the Q function from $Q(\theta^{(t)}|\theta^{(t)})$ makes the RHS positive which ensures that the LHS is also positive and consequently that θ also increases the log-likehood from $ll(\mathcal{D}_x, \theta^{(t)})$.

$$ll(\mathcal{D}_x, \theta) - ll(\mathcal{D}_x, \theta^{(t)}) \geq Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)})$$



Maximizing $Q'(\theta|\theta^{(t)})$ is same as maximizing $Q(\theta|\theta^{(t)})$

$Q'(\theta|\theta^{(t)}) = Q(\theta|\theta^{(t)}) + const$

Graphical interpretation of a single iteration of the EM: The function $Q'(\theta|\theta^{(t)})$ is upper-bounded by $ll(\mathcal{D}_x, \theta)$. The functions are equal at $\theta = \theta^{(t)}$. The EM algorithm chooses $\theta^{(t+1)}$ as the value of θ for which $Q'(\theta|\theta^{(t)})$ is maximum. Since $ll(\mathcal{D}_{xy}, \theta)$ $Q'(\theta|\theta^{(t)})$ increasing $Q'(\theta|\theta^{(t)})$ ensures that the value of the likelihood function $ll(\mathcal{D}_{xy}, \theta)$ is increased at each step.

EM Convergence (proof)

$$
\begin{aligned}
ll(\mathcal{D}_x, \theta) &= \log p(\mathcal{D}_x|\theta) \\
&= \log \frac{p(\mathcal{D}_x, \boldsymbol{y}|\theta)}{p(\boldsymbol{y}|\mathcal{D}_x, \theta)} \\
&= \log p(\mathcal{D}_{xy}|\theta) - \log p(\boldsymbol{y}|\mathcal{D}_x, \theta) \\
&= ll(\mathcal{D}_{xy}, \theta) - \log p(\boldsymbol{y}|\mathcal{D}_x, \theta) \\
&= \mathbb{E}_{\boldsymbol{y}}\left[ll(\mathcal{D}_{xy}, \theta)\middle|\mathcal{D}_x, \theta^{(t)}\right] - \mathbb{E}_{\boldsymbol{y}}\left[\log p(\boldsymbol{y}|\mathcal{D}_x, \theta)\middle|\mathcal{D}_x, \theta^{(t)}\right] \\
&= \underbrace{Q\left(\theta|\theta^{(t)}\right) + H\left(\theta^{(t)}\right)}_{Q'\left(\theta|\theta^{(t)}\right)} + KL\left(\theta^{(t)}||\theta\right) \qquad (1)
\end{aligned}
$$

Constant w.r.t. θ

$H\left(\theta^{(t)}\right)$ is short for the entropy of $p(\boldsymbol{y}|\mathcal{D}_x, \theta^{(t)})$ given by $\mathbb{E}_{\boldsymbol{y}}\left[-\log p(\boldsymbol{y}|\mathcal{D}_x, \theta^{(t)})\middle|\mathcal{D}_x, \theta^{(t)}\right]$

$KL\left(\theta^{(t)}||\theta\right)$ is short for the KL divergence between $p(\boldsymbol{y}|\mathcal{D}_x, \theta^{(t)})$ and $p(\boldsymbol{y}|\mathcal{D}_x, \theta)$

Note $KL\left(\theta^{(t)}||\theta^{(t)}\right) = 0$

Thus, from (1)

$$
\begin{aligned}
ll(\mathcal{D}_x, \theta) - ll(\mathcal{D}_x, \theta^{(t)}) &= Q\left(\theta|\theta^{(t)}\right) - Q\left(\theta^{(t)}|\theta^{(t)}\right) \\
&\quad + KL\left(\theta^{(t)}||\theta\right) \\
&\geq Q\left(\theta|\theta^{(t)}\right) - Q\left(\theta^{(t)}|\theta^{(t)}\right).
\end{aligned}
$$

since $KL\left(\theta^{(t)}||\theta\right) \geq 0$

Also from (1)

$ll(\mathcal{D}_x, \theta) \geq Q'\left(\theta|\theta^{(t)}\right)$

since $KL\left(\theta^{(t)}||\theta\right) \geq 0$

$ll(\mathcal{D}_x, \theta^{(t)}) = Q'\left(\theta^{(t)}|\theta^{(t)}\right)$

since $KL\left(\theta^{(t)}||\theta^{(t)}\right) = 0$

EM FOR EXPONENTIAL MIXTURE

$$
\begin{aligned}
ll(\mathcal{D}_{xy}, \theta) &= \sum_{i=1}^{n} \log\{w_{y_i} p(x_i|\lambda_{y_i})\} \\
&= \sum_{i=1}^{n} \log\left\{ \prod_{j=1}^{m} (w_j p(x_i|\lambda_j))^{I(y_i=j)} \right\} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} I(y_i = j)(\log w_j + \log p(x_i|\lambda_j)) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} I(y_i = j)(\log w_j + \log \lambda_j - \lambda_j x_i)
\end{aligned}
$$

For exponential distribution

$$p(x|\lambda_j) = \lambda_j e^{\lambda_j x}$$

EM FOR EXPONENTIAL MIXTURE: E-STEP

$$Q\left(\theta|\theta^{(t)}\right) = \mathbb{E}_{\boldsymbol{y}}\left[ll(D_{xy},\theta)|\mathcal{D}_x,\theta^{(t)}\right]$$

$$= \mathbb{E}_{\boldsymbol{y}}\left[\sum_{i=1}^{n}\sum_{j=1}^{m} I(y_i = j)(\log w_j + \log\lambda_j - \lambda_j x_i)\middle|\mathcal{D}_x,\theta^{(t)}\right]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{E}_{\boldsymbol{y}}\left[I(y_i = j)\middle|\mathcal{D}_x,\theta^{(t)}\right](\log w_j + \log\lambda_j - \lambda_j x_i)$$

because of linearity of expectation

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{E}_{y_i}\left[I(y_i = j)\middle|x_i,\theta^{(t)}\right](\log w_j + \log\lambda_j - \lambda_j x_i)$$

Since the distribution y_i is only dependent on x_i and independent of y_j and x_j for $i \neq j$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} p\left(y_i = j\middle|x_i,\theta^{(t)}\right)(\log w_j + \log\lambda_j - \lambda_j x_i)$$

Expectation of the indicator function of an event gives probability of that event.

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij}^{(t)}(\log w_j + \log\lambda_j - \lambda_j x_i)$$

EM FOR EXPONENTIAL MIXTURE: M-STEP

$$\text{Find } \theta^{(t+1)} = \arg\max_{\theta} Q\left(\theta|\theta^{(t)}\right) \text{ by}$$

$$\frac{\partial}{\partial w_j} Q\left(\theta|\theta^{(t)}\right) = 0 \Rightarrow w_j = \frac{1}{n}\sum_{i=1}^{n} r_{ij}^{(t)}$$

$$\frac{\partial}{\partial \lambda_j} Q\left(\theta|\theta^{(t)}\right) = 0 \Rightarrow \lambda_j = \frac{\sum_{i=1}^{n} r_{ij}^{(t)}}{\sum_{i=1}^{n} r_{ij}^{(t)} \cdot x_i}$$