Expectation Maximization (EM) Algorithm

lecturer: J. Matas, matas@cmp.felk.cvut.cz

authors: O. Drbohlav, J. Matas

Czech Technical University, Faculty of Electrical Engineering
Department of Cybernetics, Center for Machine Perception
121 35 Praha 2, Karlovo nám. 13, Czech Republic

http://cmp.felk.cvut.cz

4/Dec/2020

LECTURE PLAN

- Motivation: Observations with missing values
- Sketch of the algorithm, relation to K-means
- EM algorithm derivation and properties

EM Algorithm

- Used to find maximum likelihood parameters of a statistical model when the equations cannot be directly solved.

- Two typical cases of use:

  - **Missing data**: Some observations are incomplete. E.g. features are vectors in 5-dimensional space $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5) \in \mathbb{R}^D$ but observations have a component missing, e.g.: $(2, 5, \bullet, 1, 2)$ or $(\bullet, \bullet, 1, 4, 2)$, where '$\bullet$' are the unobserved components.

  - **Latent variables**: Observations are complete but the model can be formulated and solved more simply if further variables are introduced to it. A typical example are *mixture models* where for each observed point it is advantageous to introduce a random variable which specifies which component of the mixture generated that point.

EM for Maximum Likelihood Estimation, Example (1)

Consider multivariate normal distribution in 2D. For simplicity, let us consider the isotropic case for which the covariance matrix $\mathbf{\Sigma}$ is diagonal and parametrized by a single parameter σ^2, $\mathbf{\Sigma} = \text{diag}(\sigma^2, \sigma^2)$. The normal distribution $\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \sigma^2\right)$ for this case is then

$$\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \sigma^2\right) = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}\frac{\|\mathbf{x}-\boldsymbol{\mu}^2\|}{\sigma^2}}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^2$ is the random variable and $\boldsymbol{\mu} \in \mathbb{R}^2$ is the mean.

Having the data $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, the MLE for the parameters $\boldsymbol{\mu}$ and σ^2 are computed as:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \tag{2}$$

$$\hat{\sigma^2} = \frac{1}{2N} \sum_{i=1}^{N} \|\mathbf{x}_i - \hat{\boldsymbol{\mu}}\|^2 \tag{3}$$

($2N$ in the denominator of Eq. (3) is not a mistake. It follows from the parametrization of Σ and the dimensionality of the considered space.)

EM for Maximum Likelihood Estimation, Example (2)

Now consider the case that the data are the result of random sampling from a mixture of two such distributions (denoted A and B):

$$p(\mathbf{x}|\pi_A, \pi_B, \boldsymbol{\mu}_A, \boldsymbol{\mu}_B, \sigma_A^2, \sigma_B^2) = \pi_A \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_A, \sigma_A^2\right) + \pi_B \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_B, \sigma_B^2\right) \ , \tag{4}$$

where π_A and π_B imply the frequency with which a sample is realized from the respective distribution $(\pi_A + \pi_B = 1)$ and other parameters have obvious meaning.

Analytical derivation of MLE in this case will involve logarithm of the **sum** of two exp terms. This is not as easily solvable.

This is where the EM algorithm comes in.

EM for Maximum Likelihood Estimation, Example (3)

1. Initialize $\hat{\pi}_A, \hat{\pi}_B, \hat{\boldsymbol{\mu}}_A, \hat{\boldsymbol{\mu}}_B, \hat{\sigma}^2_A, \hat{\sigma}^2_B$
2. For each of the data $\mathbf{x}_k$, compute

$$v_k^A = \hat{\pi}_A \mathcal{N}\left(\mathbf{x}_k | \hat{\boldsymbol{\mu}}_A, \hat{\sigma}^2_A\right), \quad v_k^B = \hat{\pi}_B \mathcal{N}\left(\mathbf{x}_k | \hat{\boldsymbol{\mu}}_B, \hat{\sigma}^2_B\right) \tag{5}$$

$$q_k^A = \frac{v_k^A}{v_k^A + v_k^B}, \quad q_k^B = \frac{v_k^B}{v_k^A + v_k^B} \tag{6}$$

3. Use q_k^A and q_k^B as weights. That is, if, say, $(q_k^A, q_k^B) = (0.2, 0.8)$, act as if 20% of point $\mathbf{x}_k$ were from distribution A and 80% of that point were from distribution B. Update the estimates for the respective distributions as follows:

$$\hat{\boldsymbol{\mu}}_A = \frac{1}{\sum_{i=1}^N q_k^A} \sum_{i=1}^N q_k^A \mathbf{x}_k \tag{7}$$

$$\hat{\sigma^2_A} = \frac{1}{2\sum_{i=1}^N q_k^A} \sum_{i=1}^N q_k^A \|\mathbf{x}_k - \hat{\boldsymbol{\mu}}_A\|^2 \tag{8}$$

$$\hat{\pi^A} = \frac{1}{N} \sum_{i=1}^N q_k^A \tag{9}$$

4. (and analogously for B). Iterate.

Example: Mixture of Gaussians (general non-isotropic case)

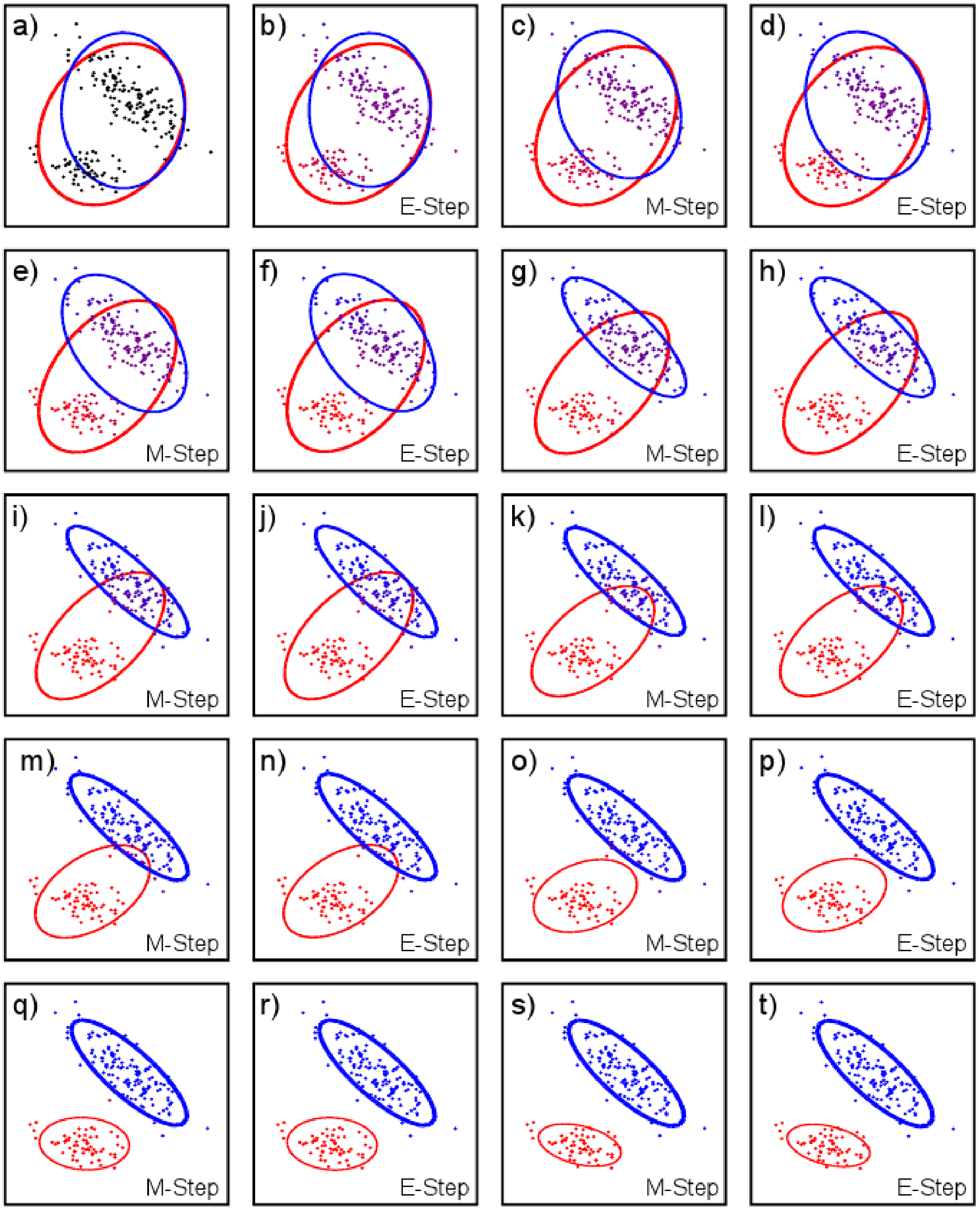


Figure 7.10 a) Initial model. b) E-step. For each data point the posterior probability that is was generated from each Gaussian is calculated (indicated by color of point). c) M-step. The mean, variance and weight of each Gaussian is updated based on these posterior probabilities. Ellipse shows Mahalanobis distance of two. Weight (thickness) of ellipse indicates weight of Gaussian. d-t) Further E-step and M-step iterations.

Image courtesy of Simon Prince. Computer Vision: Models, Learning and Inference, 2012

Toy Example 1: Estimating Means of Two Normal Distributions

We measure lengths of vehicles. The observation space has two dimensions, with $x \in \{\text{car}, \text{truck}\}$ capturing vehicle type and $y \in \mathbb{R}$ capturing length.

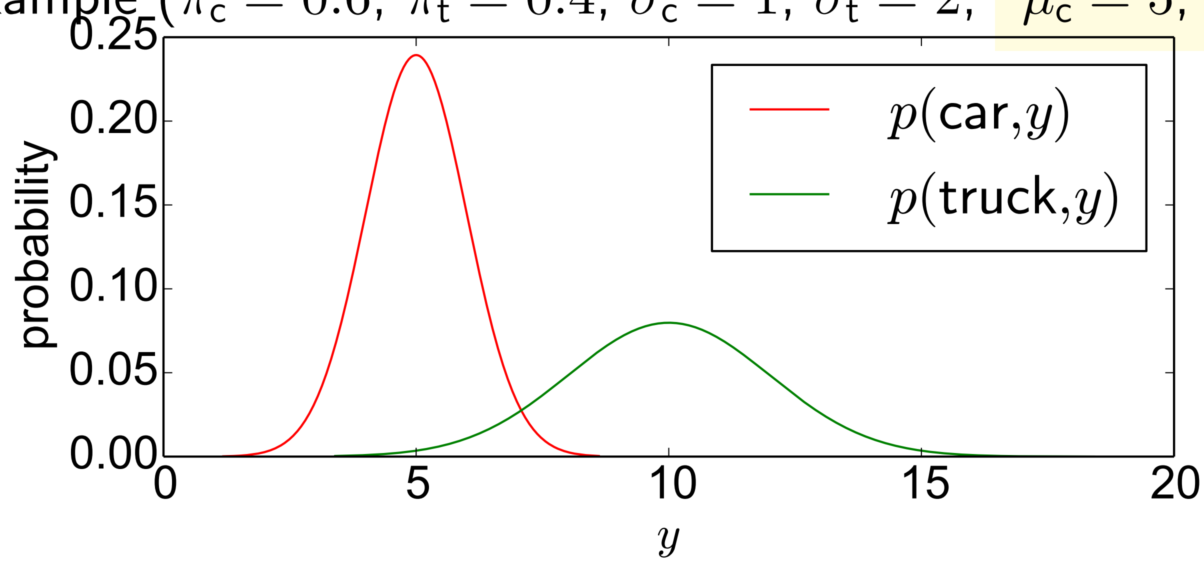
$$p(x, y) : \text{distribution}\,, \qquad x \in \{\text{car}, \text{truck}\}\,, \quad y \in \mathbb{R} \tag{10}$$

$$p(\text{car}, y) = \pi_\text{c} \mathcal{N}\left(y | \mu_\text{c}, \sigma_\text{c} = 1\right) = \kappa_\text{c} \exp\left\{-\frac{1}{2}\left(y - \mu_\text{c}\right)^2\right\}\,, \; (\kappa_\text{c} = \frac{\pi_\text{c}}{\sqrt{2\pi}}) \tag{11}$$

$$p(\text{truck}, y) = \pi_\text{t} \mathcal{N}\left(y | \mu_\text{t}, \sigma_\text{t} = 2\right) = \kappa_\text{t} \exp\left\{-\frac{1}{8}\left(y - \mu_\text{t}\right)^2\right\}\,, \; (\kappa_\text{t} = \frac{\pi_\text{c}}{\sqrt{8\pi}}) \tag{12}$$

Suppose κ_c, $\kappa_\text{t}, \sigma_\text{c}, \sigma_\text{t}$ are known. The **only unknowns** are μ_c and μ_t. We want to recover μ_c and μ_t using Maximum Likelihood.

Example $(\pi_\text{c} = 0.6,\ \pi_\text{t} = 0.4,\ \sigma_\text{c} = 1,\ \sigma_\text{t} = 2,\ \ \mu_\text{c} = 5,\ \mu_\text{t} = 10\ )$



Toy Example 1, Complete Data → Easy

The observations are:

$$\mathcal{T} = \{(x_1, y_1), (x_2, y_2), , ..., (x_N, y_N)\} \tag{13}$$

$$= \{\underbrace{(\text{car}, y_1^{(\text{c})}), (\text{car}, y_2^{(\text{c})}), ..., (\text{car}, y_C^{(\text{c})})}_{C \text{ car observations}}, \underbrace{(\text{truck}, y_1^{(\text{t})}), (\text{truck}, y_2^{(\text{t})}), ..., (\text{truck}, y_T^{(\text{t})})}_{T \text{ truck observations}}\} \tag{14}$$

Log-likelihood $\ell(\mathcal{T}) = \ln p(\mathcal{T} | \mu_\text{c}, \mu_\text{t})$:

$$\ell(\mathcal{T}) = \sum_{i=1}^{N} \ln p(x_i, y_i | \mu_\text{c}, \mu_\text{t}) = C \ln \kappa_\text{c} - \frac{1}{2} \sum_{i=1}^{C} (y_i^{(\text{c})} - \mu_\text{c})^2 + T \ln \kappa_\text{t} - \frac{1}{8} \sum_{i=1}^{T} (y_i^{(\text{t})} - \mu_\text{t})^2 \tag{15}$$

Estimation of μ_1, μ_2 using ML is easy:

$$\frac{\partial \ell(\mathcal{T})}{\partial \mu_\text{c}} = \sum_{i=1}^{C} (y_i^{(\text{c})} - \mu_\text{c}) = 0 \qquad \Rightarrow \qquad \mu_\text{c} = \frac{1}{C} \sum_{i=1}^{C} y_i^{(\text{c})} \tag{16}$$

$$\frac{\partial \ell(\mathcal{T})}{\partial \mu_\text{t}} = \frac{1}{4} \sum_{i=1}^{T} (y_i^{(\text{t})} - \mu_\text{t}) = 0 \qquad \Rightarrow \qquad \mu_\text{t} = \frac{1}{T} \sum_{i=1}^{T} y_i^{(\text{t})} \tag{17}$$

Toy Example 1, Incomplete Data → Difficult (1)

Consider some observations to have the first coordinate **missing** ($\bullet$):

$$\mathcal{T} = \{(\text{car}, y_1^{(\text{c})}), \ldots, (\text{car}, y_C^{(\text{c})}), (\text{truck}, y_1^{(\text{t})}), \ldots, (\text{truck}, y_T^{(\text{t})}), \underbrace{(\bullet, y_1^{\bullet}), \ldots, (\bullet, y_M^{\bullet})}_{\text{data with uknown vehicle type}}\} \tag{18}$$

Probability $p(y^{\bullet})$ of observing $y^{\bullet}$:

$$p(y^{\bullet}) = p(\text{car}, y^{\bullet}) + p(\text{truck}, y^{\bullet})$$

Log-likelihood:

$$\ell(\mathcal{T}) = \sum_{i=1}^{N} \ln p(x_i, y_i | \mu_\text{c}, \mu_\text{t}) = \overbrace{C \ln \kappa_\text{c} - \frac{1}{2}\sum_{i=1}^{C}(y_i^{(\text{c})} - \mu_\text{c})^2 + T \ln \kappa_\text{t} - \frac{1}{8}\sum_{i=1}^{T}(y_i^{(\text{t})} - \mu_\text{t})^2}^{\text{same term as before}} \tag{19}$$

$$+ \sum_{i=1}^{M} \ln \left(\kappa_\text{c} \exp\left\{-\frac{1}{2}(y_i^{\bullet} - \mu_\text{c})^2\right\} + \kappa_\text{t} \exp\left\{-\frac{1}{8}(y_i^{\bullet} - \mu_\text{t})^2\right\} \right) \tag{20}$$

Toy Example 1, Incomplete Data $\rightarrow$ Difficult (2)

Log-likelihood:

$$\ell(\mathcal{T}) = C \ln \kappa_{\mathsf{c}} - \frac{1}{2}\sum_{i=1}^{C}(y_i^{(\mathsf{c})} - \mu_{\mathsf{c}})^2 + T \ln \kappa_{\mathsf{t}} - \frac{1}{8}\sum_{i=1}^{T}(y_i^{(\mathsf{t})} - \mu_{\mathsf{t}})^2 \tag{21}$$

$$+ \sum_{i=1}^{M} \ln \left(\kappa_{\mathsf{c}} \exp\left\{ -\frac{1}{2}\left(y_i^{\bullet} - \mu_{\mathsf{c}}\right)^2 \right\} + \kappa_{\mathsf{t}} \exp\left\{ -\frac{1}{8}\left(y_i^{\bullet} - \mu_{\mathsf{t}}\right)^2 \right\} \right) \tag{22}$$

Optimality condition (shown for μ_{c} only):

$$0 = \frac{\partial \ell(\mathcal{T})}{\partial \mu_{\mathsf{c}}} = \sum_{i=1}^{C}(y_i^{(\mathsf{c})} - \mu_{\mathsf{c}}) \quad + \tag{23}$$

$$+ \quad \sum_{i=1}^{M} \frac{\kappa_{\mathsf{c}} \exp\left\{ -\frac{1}{2}\left(y_i^{\bullet} - \mu_{\mathsf{c}}\right)^2 \right\}}{\kappa_{\mathsf{c}} \exp\left\{ -\frac{1}{2}\left(y_i^{\bullet} - \mu_{\mathsf{c}}\right)^2 \right\} + \kappa_{\mathsf{t}} \exp\left\{ -\frac{1}{8}\left(y_i^{\bullet} - \mu_{\mathsf{t}}\right)^2 \right\}} (y_i^{\bullet} - \mu_{\mathsf{c}}) \tag{24}$$

Missing Values, Optimality Condition

Log-likelihood:

$$\ell(\mathcal{T}) = C \ln \kappa_{\mathsf{c}} - \frac{1}{2}\sum_{i=1}^{C}(y_i^{(\mathsf{c})} - \mu_{\mathsf{c}})^2 + T \ln \kappa_{\mathsf{t}} - \frac{1}{8}\sum_{i=1}^{T}(y_i^{(\mathsf{t})} - \mu_{\mathsf{t}})^2 \tag{25}$$

$$+ \sum_{i=1}^{M} \ln \left(\kappa_{\mathsf{c}} \exp\left\{ -\frac{1}{2}\left(y_i^{\bullet} - \mu_{\mathsf{c}}\right)^2 \right\} + \kappa_{\mathsf{t}} \exp\left\{ -\frac{1}{8}\left(y_i^{\bullet} - \mu_{\mathsf{t}}\right)^2 \right\} \right) \tag{26}$$

Optimality condition (shown for μ_{c} only):

$$0 = \frac{\partial \ell(\mathcal{T})}{\partial \mu_{\mathsf{c}}} = \sum_{i=1}^{C}(y_i^{(\mathsf{c})} - \mu_{\mathsf{c}}) \quad + \tag{27}$$

$$+ \quad \sum_{i=1}^{M} \frac{\overbrace{\kappa_{\mathsf{c}} \exp\left\{ -\frac{1}{2}\left(y_i^{\bullet} - \mu_{\mathsf{c}}\right)^2 \right\}}^{p(\mathsf{car},\, y_i^{\bullet} \,|\, \mu_{\mathsf{c}}, \mu_{\mathsf{t}})}}{\underbrace{\kappa_{\mathsf{c}} \exp\left\{ -\frac{1}{2}\left(y_i^{\bullet} - \mu_{\mathsf{c}}\right)^2 \right\}}_{p(\mathsf{car},\, y_i^{\bullet} \,|\, \mu_{\mathsf{c}}, \mu_{\mathsf{t}})} + \underbrace{\kappa_{\mathsf{t}} \exp\left\{ -\frac{1}{8}\left(y_i^{\bullet} - \mu_{\mathsf{t}}\right)^2 \right\}}_{p(\mathsf{truck},\, y_i^{\bullet} \,|\, \mu_{\mathsf{c}}, \mu_{\mathsf{t}})}} \left(y_i^{\bullet} - \mu_{\mathsf{c}}\right) \tag{28}$$

Missing Values, Optimality Condition

Log-likelihood:

$$\ell(\mathcal{T}) = C \ln \kappa_{\mathsf{c}} - \frac{1}{2}\sum_{i=1}^{C}(y_i^{(\mathsf{c})} - \mu_{\mathsf{c}})^2 + T \ln \kappa_{\mathsf{t}} - \frac{1}{8}\sum_{i=1}^{T}(y_i^{(\mathsf{t})} - \mu_{\mathsf{t}})^2 \tag{29}$$

$$+ \sum_{i=1}^{M} \ln \left(\kappa_{\mathsf{c}} \exp\left\{ -\frac{1}{2}\left(y_i^{\bullet} - \mu_{\mathsf{c}}\right)^2 \right\} + \kappa_{\mathsf{t}} \exp\left\{ -\frac{1}{8}\left(y_i^{\bullet} - \mu_{\mathsf{t}}\right)^2 \right\} \right) \tag{30}$$

Optimality condition (shown for μ_{c} only):

$$0 = \frac{\partial \ell(\mathcal{T})}{\partial \mu_{\mathsf{c}}} = \sum_{i=1}^{C}(y_i^{(\mathsf{c})} - \mu_{\mathsf{c}}) \quad + \tag{31}$$

$$+ \quad \sum_{i=1}^{M} \overbrace{\boxed{\frac{\kappa_{\mathsf{c}} \exp\left\{-\frac{1}{2}\left(y_i^{\bullet} - \mu_{\mathsf{c}}\right)^2\right\}}{\kappa_{\mathsf{c}} \exp\left\{-\frac{1}{2}\left(y_i^{\bullet} - \mu_{\mathsf{c}}\right)^2\right\} + \kappa_{\mathsf{t}} \exp\left\{-\frac{1}{8}\left(y_i^{\bullet} - \mu_{\mathsf{t}}\right)^2\right\}}}}^{p(\mathsf{car}|y_i^{\bullet}, \mu_{\mathsf{c}}, \mu_{\mathsf{t}})} (y_i^{\bullet} - \mu_{\mathsf{c}}) \tag{32}$$

Missing Values, Optimality Conditions

Optimality conditions (shown for both μ_{c} and μ_{t}):

$$0 = \frac{\partial \ell(\mathcal{T})}{\partial \mu_{\mathsf{c}}} = \sum_{i=1}^{C} (y_i^{(\mathsf{c})} - \mu_{\mathsf{c}}) \quad + \tag{33}$$

$$+ \quad \sum_{i=1}^{M} \overbrace{\boxed{\frac{\kappa_{\mathsf{c}} \exp\left\{-\frac{1}{2}(y_i^{\bullet} - \mu_{\mathsf{c}})^2\right\}}{\kappa_{\mathsf{c}} \exp\left\{-\frac{1}{2}(y_i^{\bullet} - \mu_{\mathsf{c}})^2\right\} + \kappa_{\mathsf{t}} \exp\left\{-\frac{1}{8}(y_i^{\bullet} - \mu_{\mathsf{t}})^2\right\}}}}^{p(\mathsf{car}|y_i^{\bullet}, \mu_{\mathsf{c}}, \mu_{\mathsf{t}})} (y_i^{\bullet} - \mu_{\mathsf{c}}) \tag{34}$$

$$0 = 4\frac{\partial \ell(\mathcal{T})}{\partial \mu_{\mathsf{t}}} = \sum_{i=1}^{T} (y_i^{(\mathsf{t})} - \mu_{\mathsf{t}}) + \sum_{i=1}^{M} p(\mathsf{truck}|y_i^{\bullet}, \mu_{\mathsf{c}}, \mu_{\mathsf{t}}) \; (y_i^{\bullet} - \mu_{\mathsf{t}}) \tag{35}$$

Note:

- Complicated equations for the uknowns μ_{c}, μ_{t}
- Both equations contain μ_{c} and μ_{t} (cf. case with no missing variables)

Missing Values, EM Approach

Optimality conditions (shown for both μ_{c} and μ_{t}):

$$\sum_{i=1}^{C}(y_i^{(\text{c})} - \mu_{\text{c}}) + \sum_{i=1}^{M} p(\text{car}|y_i^{\bullet}, \mu_{\text{c}}, \mu_{\text{t}})\ (y_i^{\bullet} - \mu_{\text{c}}) = 0 \tag{36}$$

$$\sum_{i=1}^{T}(y_i^{(\text{t})} - \mu_{\text{t}}) + \sum_{i=1}^{M} p(\text{truck}|y_i^{\bullet}, \mu_{\text{c}}, \mu_{\text{t}})\ (y_i^{\bullet} - \mu_{\text{t}}) = 0 \tag{37}$$

If $p(\text{car}|y_i^{\bullet}, \mu_{\text{c}}, \mu_{\text{t}})$ and $p(\text{truck}|y_i^{\bullet}, \mu_{\text{c}}, \mu_{\text{t}})$ were known, the estimation would've been easy:

- Let z_i $(i = 1, 2, ..., M)$, $z_i \in \{\text{car}, \text{truck}\}$ denote the missing values. Define $q(z_i) = p(z_i|y_i^{\bullet}, \mu_{\text{c}}, \mu_{\text{t}})$
- The equations lead to

$$\sum_{i=1}^{C}(y_i^{(\text{c})} - \mu_{\text{c}}) + \sum_{i=1}^{M} q(z_i = \text{car})\ (y_i^{\bullet} - \mu_{\text{c}}) = 0 \tag{38}$$

$$\Rightarrow \qquad \mu_{\text{c}} = \frac{\sum_{i=1}^{C} y_i^{(\text{c})} + \sum_{i=1}^{M} q(z_i = \text{car})\, y_i^{\bullet}}{C + \sum_{i=1}^{M} q(z_i = \text{car})} \tag{39}$$

and similarly,

$$\mu_{\text{t}} = \frac{\sum_{i=1}^{T} y_i^{(\text{t})} + \sum_{i=1}^{M} q(z_i = \text{truck})\, y_i^{\bullet}}{T + \sum_{i=1}^{M} q(z_i = \text{truck})} \tag{40}$$

Missing Values, EM Approach

$$\mu_{\text{c}} = \frac{\sum_{i=1}^{C} y_i^{(\text{c})} + \sum_{i=1}^{M} q(z_i = \text{car})\, y_i^{\bullet}}{C + \sum_{i=1}^{M} q(z_i = \text{car})} \tag{41}$$

$$\mu_{\text{t}} = \frac{\sum_{i=1}^{T} y_i^{(\text{t})} + \sum_{i=1}^{M} q(z_i = \text{truck})\, y_i^{\bullet}}{T + \sum_{i=1}^{M} q(z_i = \text{truck})} \tag{42}$$

- These expressions are weighted averages of the observed y's. Data with non-missing x have weight 1, the data with missing x have weight $q(z_i)$. How about trying the following procedure for finding the ML estimate of μ_{c} and μ_{t}:

  1. Initialize μ_{c}, μ_{t}
  2. Compute $q(z_i) = p(z_i | y_i^{\bullet}, \mu_{\text{c}}, \mu_{\text{t}})$ for all $i = 1, 2, ..., M$
  3. Recompute μ_{c}, μ_{t} according to Eqs.(41, 42)
  4. If termination condition is met, finish. Otherwise goto 2.

- This is the essence of the **EM algorithm**, with Step 2 called the **Expectation** (E) step and Step 3 called the **Maximization** (M) step.

Clustering, Soft Assignment, Relation to K-means (1)

An extreme of the previous example is that **no** data have the x-coordinate value (car/truck vehicle type). Everything works just as well:

$$\mu_{\text{c}} = \frac{\sum_{i=1}^{M} q(z_i = \text{car})\, y_i^{\bullet}}{\sum_{i=1}^{M} q(z_i = \text{car})} \tag{43}$$

$$\mu_{\text{t}} = \frac{\sum_{i=1}^{M} q(z_i = \text{truck})\, y_i^{\bullet}}{\sum_{i=1}^{M} q(z_i = \text{truck})} \tag{44}$$

1. Initialize μ_{c}, μ_{t}
2. Compute $q(z_i) = p(z_i | y_i^{\bullet}, \mu_{\text{c}}, \mu_{\text{t}})$ for all $i = 1, 2, \dots, M$
3. Recompute μ_{c}, μ_{t} according to Eqs.(45, 46)
4. If termination condition is met, finish. Otherwise goto 2.

Note: Can you imagine this algorithm to end up at a local maximum?

Clustering, Soft Assignment, Relation to K-means (2)

An extreme of the previous example is that **no** data have the x-coordinate (car/truck).

$$\mu_{\mathsf{c}} = \frac{\sum_{i=1}^{M} q(z_i = \mathsf{car})\, y_i}{\sum_{i=1}^{M} q(z_i = \mathsf{car})} \tag{45}$$

$$\mu_{\mathsf{t}} = \frac{\sum_{i=1}^{M} q(z_i = \mathsf{truck})\, y_i}{\sum_{i=1}^{M} q(z_i = \mathsf{truck})} \tag{46}$$

EM algorithm:

1. Initialize μ_{c}, μ_{t}
2. Compute $q(z_i) = p(z_i | y_i, \mu_{\mathsf{c}}, \mu_{\mathsf{t}})$ for all $i = 1, 2, \ldots, M$
3. Recompute μ_{c}, μ_{t} according to Eqs.(45, 46)
4. If termination condition is met, finish. Otherwise goto 2.

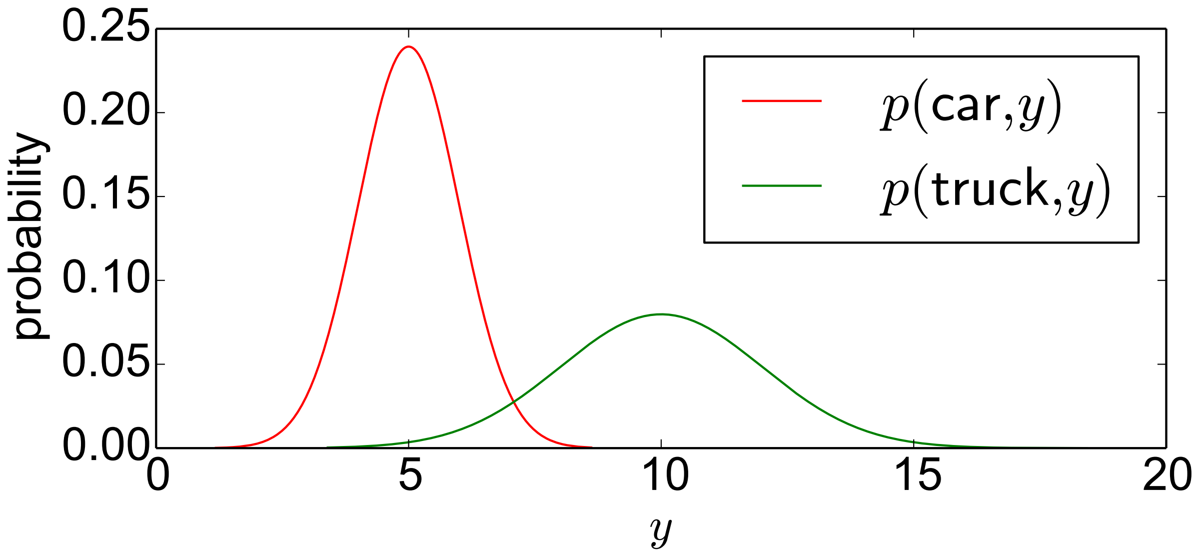
K-means:

1. ditto
2. $q(z_i = \mathsf{car}) = [\![|y_i - \mu_{\mathsf{c}}| < |y_i - \mu_{\mathsf{t}}|]\!]$
   $q(z_i = \mathsf{truck}) = [\![|y_i - \mu_{\mathsf{t}}| \le |y_i - \mu_{\mathsf{c}}|]\!]$
   for all $i = 1, 2, \ldots, M$
3. ditto
4. ditto

EM-based clustering uses soft assignment. K-means can be interpreted as an EM-based clustering with hard assignment.

Example 1 - Setting

$\pi_\mathrm{c} = 0.6,\ \pi_\mathrm{t} = 0.4,\ \sigma_\mathrm{c} = 1,\ \sigma_\mathrm{t} = 2,\quad \mu_\mathrm{c} = 5,\ \mu_\mathrm{t} = 10$



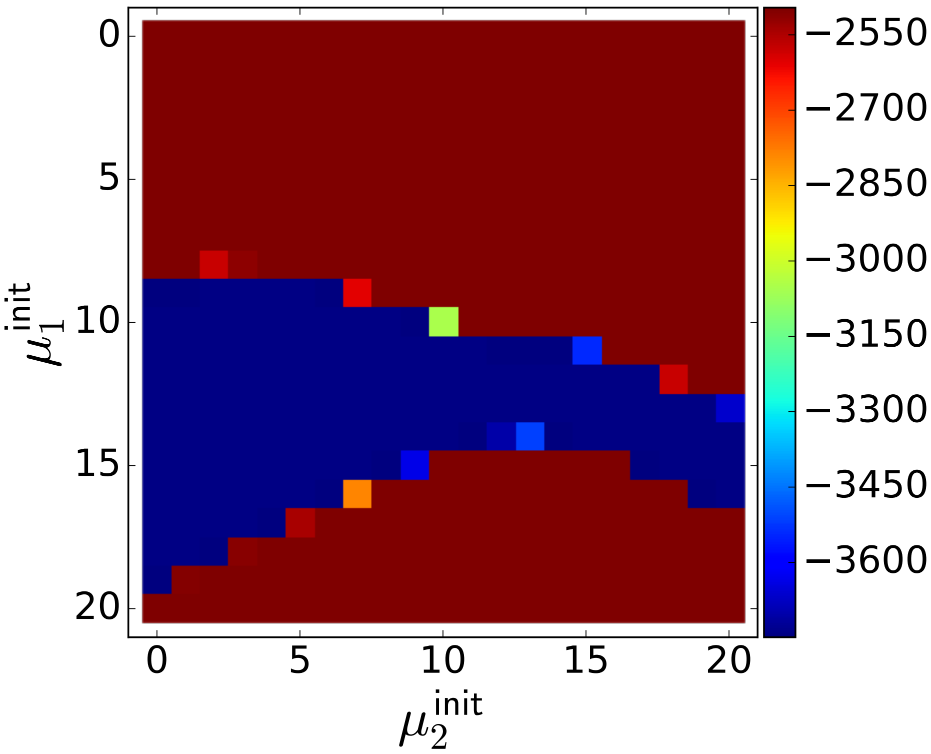
Data:

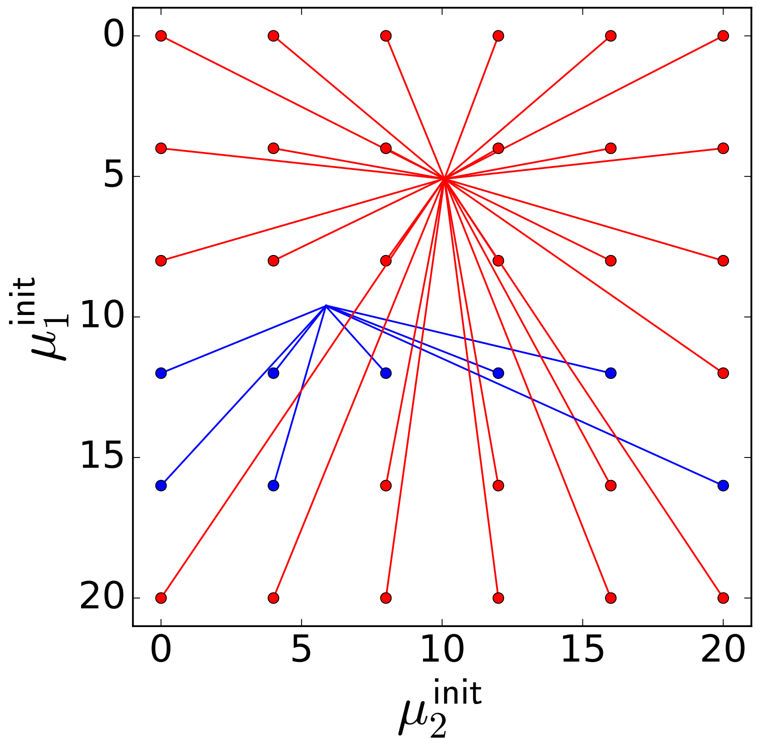
- 50 points from car distribution, 50 points from truck d., 1000 points from mixed distribution (car/truck coordinate unknown)

Experiment:

Employ EM algorithm for estimating μ_1, μ_2. Use different initializations.

Example 1 - Result





Log-likelihood ℓ after 10 iterations of EM, depending on initialization $(\mu_1^{\text{init}}, \mu_2^{\text{init}})$.

Convergence in this case is quite fast (3 iterations are enough for most of the initialization values.)

Value of (μ_1, μ_2) after 10 iterations, depending on initialization $(\mu_1^{\text{init}}, \mu_2^{\text{init}})$. The first point of convergence corresponds to the ground truth values $(\mu_1, \mu_2) = (5, 10)$. The second point is a only a local maximum of log-likelihood. It corresponds to car distribution approximating truck sample points, and vice versa.

Mixture Models

Generalization of the Motivation example with missing values.

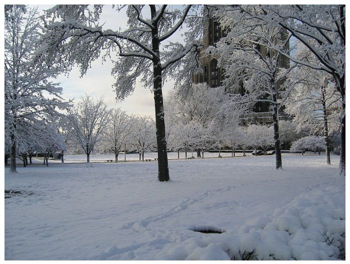
$$\mu_{\mathsf{c}} = \frac{\sum_{i=1}^{M} q(z_i = \mathsf{car})\, y_i^{\bullet}}{\sum_{i=1}^{M} q(z_i = \mathsf{car})} \tag{47}$$

$$\sigma_{\mathsf{c}}^2 = \frac{\sum_{i=1}^{M} q(z_i = \mathsf{car})\, (y_i^{\bullet} - \mu_{\mathsf{c}})^2}{\sum_{i=1}^{M} q(z_i = \mathsf{car})} \tag{48}$$

$$\pi_{\mathsf{c}} = \frac{\sum_{i=1}^{M} q(z_i = \mathsf{car})}{M} \tag{49}$$

Toy Example 2: (Temperature × Snow) Model Estimation

You are measuring temperature and amount of snow in the mountains in the month of January. Both the temperature t and the snow s observations are binary:



$$t \in \{t_0 = \text{low temperature}, t_1 = \text{high temperature}\} \quad (50)$$

$$s \in \{s_0 = \text{little snow}, s_1 = \text{lot of snow}\} \quad (51)$$

Your own long-term research suggests that the model for the joint probability $p(t, s)$ can be parametrized by two scalars a and b and written as

$p(t, s|a, b)$

t_0	a	$5a$
t_1	$3b$	b
	s_0	s_1

(52)

At a big ski-center, you have N measurements in total, with counts for individual possibilities for t and s as follows:

observation counts

t_0	N_{00}	N_{01}
t_1	N_{10}	N_{11}
	s_0	s_1

(53)

What is the ML estimate for a and b?

Toy Example 2: Model Estimation

$p(t, s|a, b)$

t_0	a	$5a$
t_1	$3b$	b
	s_0	s_1

observation counts

t_0	N_{00}	N_{01}
t_1	N_{10}	N_{11}
	s_0	s_1

Likelihood is $P(\mathcal{T}|a, b) = a^{N_{00}}(5a)^{N_{01}}(3b)^{N_{10}}(b)^{N_{11}}$.

Log-likelihood is $\ell(\mathcal{T}|a, b) = N_{00} \ln a + N_{01} \ln 5a + N_{10} \ln 3b + N_{11} \ln b$. Maximize this log-likelihood s.t. $6a + 4b = 1$. The Lagrangian is $L(a, b, \lambda) = N_{00} \ln a + N_{01} \ln 5a + N_{10} \ln 3b + N_{11} \ln b + \lambda(6a + 4b - 1)$. Conditions of optimality are:

$$\frac{\partial L}{\partial a} = N_{00}\frac{1}{a} + N_{01}\frac{1}{a} + 6\lambda = 0 \tag{54}$$

$$\frac{\partial L}{\partial b} = N_{10}\frac{1}{b} + N_{11}\frac{1}{b} + 4\lambda = 0 \tag{55}$$

$$6a + 4b = 1 \tag{56}$$

and they have the solution ($N = N_{00} + N_{01} + N_{10} + N_{11}$):

$$a = \frac{N_{00} + N_{01}}{6N} \qquad b = \frac{N_{10} + N_{11}}{4N} \tag{57}$$

Toy Example 2: Model Estimation (Incomplete Data)

Now imagine you have data from little village in the mountains. Unfortunately, there is *no* measurement for which both temperature and snow amount would be available. The data consist only of T_0 reports of low temperature, T_1 of high temperature, S_0 of little snow and S_1 of lots of snow.

$p(t, s | a, b)$

t_0	a	$5a$
t_1	$3b$	b
	s_0	s_1

$\Rightarrow$

$p(t_0)$	$6a$
$p(t_1)$	$4b$
$p(s_0)$	$a + 3b$
$p(s_1)$	$5a + b$

observation counts

t_0	T_0
t_1	T_1
s_0	S_0
s_1	S_1

Log-likelihood is $\ell(\mathcal{T}|a, b) = T_0 \ln 6a + T_1 \ln 4b + S_0 \ln(a + 3b) + S_1 \ln(5a + b)$.
Maximize this log-likelihood s.t. $6a + 4b = 1$. The Lagrangian is
$L(a, b, \lambda) = T_0 \ln 6a + T_1 \ln 4b + S_0 \ln(a + 3b) + S_1 \ln(5a + b) + \lambda(6a + 4b - 1)$.
Conditions of optimality:

$$\frac{\partial L}{\partial a} = \frac{T_0}{a} + \frac{S_0}{a + 3b} + \frac{5S_1}{5a + b} + 6\lambda = 0 \tag{58}$$

$$\frac{\partial L}{\partial b} = \frac{T_1}{b} + \frac{3S_0}{a + 3b} + \frac{3S_1}{5a + b} + 4\lambda = 0 \tag{59}$$

$$6a + 4b = 1 \tag{60}$$

$\rightarrow$ Not as easy to solve as in the previous case!

Toy Example 2: Model Estimation using EM algorithm

This is what EM algorithm would do to maximize likelihood for these incomplete data.

1. Make initial estimate of a and b
2. **E**-step: For each observation, compute the distribution over the missing value, given the observed value and current estimate of a and b.
   E. g. observation $(\bullet, s_0)$ where '$\bullet$' is the unknown temperature t and s_0 is the observed low amount of snow. The distrib. $q(t) = p(t|s_0, a, b)$ is computed as follows:

   $p(t, s|a, b)$

t_0	a	$5a$
t_1	$3b$	b
	s_0	s_1

   $$q(t_0) = p(t_0|s_0, a, b) = \frac{a}{a + 3b} \tag{61}$$

   $$q(t_1) = p(t_1|s_0, a, b) = \frac{3b}{a + 3b} \tag{62}$$

3. **M**-step: Recompute parameters a, b:
   Use the distribution q computed in the previous step as weights.
   I.e. the considered incomplete observation $(\bullet, s_0)$ produces two complete observations:
   (t_0, s_0) with weight $q(t_0)$, and (t_1, s_0) with weight $q(t_1)$.
   Let w_{ij} be the sum of weights for observations (t_i, s_j) across the entire dataset. Then a and b are computed (using the result for complete data) as:

   $$a = \frac{w_{00} + w_{01}}{6N}, \qquad b = \frac{w_{10} + w_{11}}{4N} \tag{63}$$

4. Iterate (go to 2.)

EM algorithm - Derivation

- $\mathcal{T}$: training set
- $\mathbf{o}$: all observed values (no essential difference between $\mathcal{T}$ and $\mathbf{o}$, just notational convenience)
- $\mathbf{z}$: all unobserved values
- $\boldsymbol{\theta}$: model parameters to be estimated.

Goal: Find $\boldsymbol{\theta}^*$ using the Maximum Likelihood approach:

$$\boldsymbol{\theta}^* = \operatorname*{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \ln p(\mathbf{o}|\boldsymbol{\theta}) \tag{64}$$

Line of thought

Assume that solving this:

$$\operatorname*{argmax}_{\boldsymbol{\theta}} \ln p(\mathbf{o}, \mathbf{z}|\boldsymbol{\theta}) \tag{65}$$

is easy (that is, estimation of optimal parameters had the data been complete.)

Our goal will be to rewrite Eq. (64) in a way which will involve optimization terms of kind as in Eq. (65).

Lower Bound on the Log Likelihood

$$\ln p(\mathbf{o}|\boldsymbol{\theta}) = \ln \sum_{\mathbf{z}} p(\mathbf{o}, \mathbf{z}|\boldsymbol{\theta}) \tag{66}$$

$$= \ln \sum_{\mathbf{z}} q(\mathbf{z}) \frac{p(\mathbf{o}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} \tag{67}$$

Introduction of distribution $q(\mathbf{z})$
As $\forall \mathbf{z} : 0 \leq q(\mathbf{z}) \leq 1$ and $\sum_{\mathbf{z}} q(\mathbf{z}) = 1$, the sum is now a convex combination of $p(\mathbf{o}, \mathbf{z}|\boldsymbol{\theta})/q(\mathbf{z})$.

$$\geq \sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p(\mathbf{o}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} \tag{68}$$

Jensen's inequality. Here inequality holds because logarithm is a concave function.

Define

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p(\mathbf{o}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} . \tag{69}$$

This $\mathcal{L}(q, \boldsymbol{\theta})$ is the lower bound for $\ln p(\mathbf{o}|\boldsymbol{\theta})$ due to Eq. (68), for any distribution q.

Maximizing $\mathcal{L}(q, \boldsymbol{\theta})$ will also push the log likelihood $\ln p(\mathbf{o}|\boldsymbol{\theta})$ upwards.

How Tight Is This Bound? (1)

$$\ln p(\mathbf{o}|\boldsymbol{\theta}) - \mathcal{L}(q,\boldsymbol{\theta}) = \ln p(\mathbf{o}|\boldsymbol{\theta}) - \sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p(\mathbf{o},\mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} \tag{70}$$

$$= \ln p(\mathbf{o}|\boldsymbol{\theta}) - \sum_{\mathbf{z}} q(\mathbf{z}) \{\ln \underbrace{p(\mathbf{o},\mathbf{z}|\boldsymbol{\theta})}_{p(\mathbf{z}|\mathbf{o},\boldsymbol{\theta})p(\mathbf{o}|\boldsymbol{\theta})} - \ln q(\mathbf{z})\} \tag{71}$$

$$= \ln p(\mathbf{o}|\boldsymbol{\theta}) - \sum_{\mathbf{z}} q(\mathbf{z}) \{\ln p(\mathbf{z}|\mathbf{o},\boldsymbol{\theta}) + \ln p(\mathbf{o}|\boldsymbol{\theta}) - \ln q(\mathbf{z})\} \tag{72}$$

$$= \ln p(\mathbf{o}|\boldsymbol{\theta}) - \underbrace{\sum_{\mathbf{z}} q(\mathbf{z})}_{1} \ln p(\mathbf{o}|\boldsymbol{\theta}) - \sum_{\mathbf{z}} q(\mathbf{z}) \{\ln p(\mathbf{z}|\mathbf{o},\boldsymbol{\theta}) - \ln q(\mathbf{z})\} \tag{73}$$

$$= -\sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p(\mathbf{z}|\mathbf{o},\boldsymbol{\theta})}{q(\mathbf{z})} \tag{74}$$

This is the Kullback Leibler divergence between the two distributions $q(\mathbf{z})$ and $p(\mathbf{z}|\mathbf{o},\boldsymbol{\theta})$:

$$D_{\mathsf{KL}}(q||p) = \sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{o},\boldsymbol{\theta})} = -\sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p(\mathbf{z}|\mathbf{o},\boldsymbol{\theta})}{q(\mathbf{z})} \tag{75}$$

How Tight Is This Bound? (2)

$$\ln p(\mathbf{o}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + D_{\mathsf{KL}}(q||p) \tag{76}$$

$$\begin{array}{ccc} \uparrow & \uparrow & \uparrow \\ \text{log likelihood} & \text{lower bound} & \text{gap} \end{array}$$

We already know that due to Jensen's inequality, $\mathcal{L}(q, \boldsymbol{\theta})$ is indeed the lower bound. This is confirmed by the fact that $D_{\mathsf{KL}}(q||p) \geq 0$ for any q, p. Additionally,

$$D_{\mathsf{KL}}(q||p) = 0 \qquad \Leftrightarrow \qquad p = q. \tag{77}$$

When $q = p$, the bound is tight.

EM algorithm

$$\ln p(\mathbf{o}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + D_{\mathsf{KL}}(q||p) \tag{78}$$

$$\underset{\text{log likelihood}}{\uparrow} \quad \underset{\text{lower bound}}{\uparrow} \quad \underset{\text{gap}}{\uparrow}$$

EM algorithm attempts to maximize the log-likelihood by instead maximizing the lower bound (why 'attempts'? Because it may end up in local maximum).

1. Initialize $\boldsymbol{\theta} = \boldsymbol{\theta}^{(0)}$ $(t = 0)$
2. **E-step** (Expectation):
$$q^{(t+1)} = \operatorname*{argmax}_q \mathcal{L}(q, \boldsymbol{\theta}^{(t)}) \tag{79}$$
3. **M-step** (Maximization):
$$\boldsymbol{\theta}^{(t+1)} = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathcal{L}(q^{(t+1)}, \boldsymbol{\theta}) \tag{80}$$
4. If termination condition is not met, goto 2.

Expectation step

E-step: $\boldsymbol{\theta}^{(t)}$ is fixed

$$q^{(t+1)} = \operatorname*{argmax}_q \mathcal{L}(q, \boldsymbol{\theta}^{(t)}) \tag{81}$$

$$\mathcal{L}(q, \boldsymbol{\theta}^{(t)}) = \underbrace{\ln p(\mathbf{o}|\boldsymbol{\theta}^{(t)})}_{\text{const.}} - D_{\mathsf{KL}}(q||p) \tag{82}$$

Note: The distribution q maximizing this term is the one which minimizes the KL divergence. KL divergence is minimized when the two distributions are the same. Thus, the distribution maximizing Eq. (81) is

$$q^{(t+1)}(\mathbf{z}) = p(\mathbf{z}|\mathbf{o}, \boldsymbol{\theta}^{(t)}) \,. \qquad \left[D_{\mathsf{KL}}(q||p) = -\sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p(\mathbf{z}|\mathbf{o}, \boldsymbol{\theta})}{q(\mathbf{z})} \right] \tag{83}$$

Note that this corresponds to what we've obtained e.g. in our car/truck example,

$$q_i^{(t+1)}(\text{car}) = p(\text{car}|y_i^{\bullet}, \mu_{\mathsf{c}}, \mu_{\mathsf{t}}) \,, \qquad q_i^{(t+1)}(\text{truck}) = p(\text{truck}|y_i^{\bullet}, \mu_{\mathsf{c}}, \mu_{\mathsf{t}}) \tag{84}$$

Maximization step

M-step: $q^{(t+1)}$ is fixed

$$\boldsymbol{\theta}^{(t+1)} = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathcal{L}(q^{(t+1)}, \boldsymbol{\theta}) \tag{85}$$

$$\mathcal{L}(q^{(t+1)}, \boldsymbol{\theta}) = \sum_{\mathbf{z}} q^{(t+1)}(\mathbf{z}) \ln \frac{p(\mathbf{o}, \mathbf{z} | \boldsymbol{\theta})}{q^{(t+1)}(\mathbf{z})} \tag{86}$$

$$= \sum_{\mathbf{z}} q^{(t+1)}(\mathbf{z}) \ln p(\mathbf{o}, \mathbf{z} | \boldsymbol{\theta}) - \underbrace{\sum_{\mathbf{z}} q^{(t+1)}(\mathbf{z}) \ln q^{(t+1)}(\mathbf{z})}_{\text{const.}} \tag{87}$$

Result: The parameters $\boldsymbol{\theta}$ maximizing Eq. (85) are

$$\boldsymbol{\theta}^{(t+1)} = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{\mathbf{z}} q^{(t+1)}(\mathbf{z}) \ln p(\mathbf{o}, \mathbf{z} | \boldsymbol{\theta}) \, . \tag{88}$$

Note that this maximization is done as if all data were known (observed) and thus is often easy (has analytic solution.) E.g. in the case of estimating mean of Gaussian mixture component, it leads to weighted average of data.

Summary

- EM's most known application is estimating Gaussian Mixtures. M-step computes probabilities that a given point is generated by given components, and E-step computes the uknown parameters effectively (analytic solution). EM algorithm is similarly useful and effective for more exponential family distributions.
- EM cleverly maximizes likelihood by pushing its lower bound upwards.
- It is an iterative method and may not end up in the global maximum.
- Attention needs to be applied to parameter initialization, like with other methods we've already encountered (K-means, NNs, . . .)