

Expectation Maximization (EM) Algorithm

lecturer: [O. Drbohlav](#), drbohlav@cmp.felk.cvut.cz

author: J. Matas, O. Drbohlav

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>

14/Dec/2015

Last update: 14/Dec/2015, 4pm

LECTURE PLAN

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

Motivation. Example (1)

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

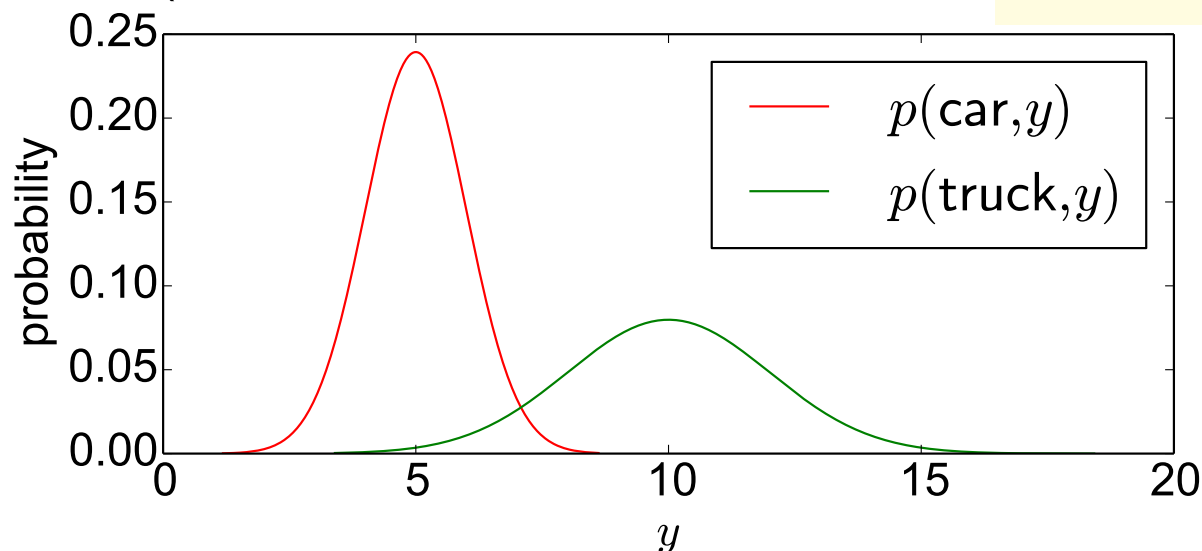
$$p(x, y) : \text{distribution}, \quad x \in \{\text{car, truck}\}, \quad y \in \mathbb{R} \quad (1)$$

$$p(\text{car}, y) = \pi_c \mathcal{N}(y | \mu_c, \sigma_c = 1) = \kappa_c \exp \left\{ -\frac{1}{2} (y - \mu_c)^2 \right\}, \quad (\kappa_c = \frac{\pi_c}{\sqrt{2\pi}}) \quad (2)$$

$$p(\text{truck}, y) = \pi_t \mathcal{N}(y | \mu_t, \sigma_t = 2) = \kappa_t \exp \left\{ -\frac{1}{8} (y - \mu_t)^2 \right\}, \quad (\kappa_t = \frac{\pi_c}{\sqrt{8\pi}}) \quad (3)$$

Parameters $\pi_c, \pi_t, \sigma_c, \sigma_t$ are assumed to be known. The **only unknowns** are μ_c and μ_t . We want to recover μ_c and μ_t using Maximum Likelihood.

Example ($\pi_c = 0.6, \pi_t = 0.4, \sigma_c = 1, \sigma_t = 2, \mu_c = 5, \mu_t = 10$)



Motivation. Example (2)

The observations are:

$$\mathcal{T} = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\} \quad (4)$$

$$= \underbrace{\{(\text{car}, y_1^{(c)}), (\text{car}, y_2^{(c)}), \dots, (\text{car}, y_C^{(c)})\}}_{C \text{ car observations}}, \underbrace{\{(\text{truck}, y_1^{(t)}), (\text{truck}, y_2^{(t)}), \dots, (\text{truck}, y_T^{(t)})\}}_{T \text{ truck observations}} \quad (5)$$

Log-likelihood $\ell(\mathcal{T}) = \ln p(\mathcal{T} | \mu_c, \mu_t)$:

$$\ell(\mathcal{T}) = \sum_{i=1}^N \ln p(x_i, y_i | \mu_c, \mu_t) = C \ln \kappa_c - \frac{1}{2} \sum_{i=1}^C (y_i^{(c)} - \mu_c)^2 + T \ln \kappa_t - \frac{1}{8} \sum_{i=1}^T (y_i^{(t)} - \mu_t)^2 \quad (6)$$

Estimation of μ_1, μ_2 is very easy:

$$\frac{\partial \ell(\mathcal{T})}{\partial \mu_c} = \sum_{i=1}^C (y_i^{(c)} - \mu_c) = 0 \quad \Rightarrow \quad \mu_c = \frac{1}{C} \sum_{i=1}^C y_i^{(c)} \quad (7)$$

$$\frac{\partial \ell(\mathcal{T})}{\partial \mu_t} = \frac{1}{4} \sum_{i=1}^T (y_i^{(t)} - \mu_t) = 0 \quad \Rightarrow \quad \mu_t = \frac{1}{T} \sum_{i=1}^T y_i^{(t)} \quad (8)$$

Motivation. Missing Values (3)

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

$$\mathcal{T} = \{(\text{car}, y_1^{(c)}), \dots, (\text{car}, y_C^{(c)}), (\text{truck}, y_1^{(t)}), \dots, (\text{truck}, y_T^{(t)}), \underbrace{(\bullet, y_1^\bullet), \dots, (\bullet, y_M^\bullet)}_{\text{data with unknown vehicle type}}\} \quad (9)$$

What is the probability of observing y^\bullet ?

$$p(y^\bullet) = p(\text{car}, y^\bullet) + p(\text{truck}, y^\bullet) \quad (\text{marginalizing over unknown value})$$

Log-likelihood:

$$\ell(\mathcal{T}) = \sum_{i=1}^N \ln p(x_i, y_i | \mu_c, \mu_t) = \overbrace{C \ln \kappa_c - \frac{1}{2} \sum_{i=1}^C (y_i^{(c)} - \mu_c)^2 + T \ln \kappa_t - \frac{1}{8} \sum_{i=1}^T (y_i^{(t)} - \mu_t)^2}^{\text{same term as before}} \quad (10)$$

$$+ \sum_{i=1}^M \ln \left(\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\} + \kappa_t \exp \left\{ -\frac{1}{8} (y_i^\bullet - \mu_t)^2 \right\} \right) \quad (11)$$

Motivation. Missing Values (4)

Log-likelihood:

$$\ell(\mathcal{T}) = C \ln \kappa_c - \frac{1}{2} \sum_{i=1}^C (y_i^{(c)} - \mu_c)^2 + T \ln \kappa_t - \frac{1}{8} \sum_{i=1}^T (y_i^{(t)} - \mu_t)^2 \quad (12)$$

$$+ \sum_{i=1}^M \ln \left(\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\} + \kappa_t \exp \left\{ -\frac{1}{8} (y_i^\bullet - \mu_t)^2 \right\} \right) \quad (13)$$

Optimality condition (shown for μ_c only):

$$0 = \frac{\partial \ell(\mathcal{T})}{\partial \mu_c} = \sum_{i=1}^C (y_i^{(c)} - \mu_c) + \quad (14)$$

$$+ \sum_{i=1}^M \frac{\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\}}{\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\} + \kappa_t \exp \left\{ -\frac{1}{8} (y_i^\bullet - \mu_t)^2 \right\}} (y_i^\bullet - \mu_c) \quad (15)$$

Motivation. Missing Values (5)

Log-likelihood:

$$\ell(\mathcal{T}) = C \ln \kappa_c - \frac{1}{2} \sum_{i=1}^C (y_i^{(c)} - \mu_c)^2 + T \ln \kappa_t - \frac{1}{8} \sum_{i=1}^T (y_i^{(t)} - \mu_t)^2 \quad (16)$$

$$+ \sum_{i=1}^M \ln \left(\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\} + \kappa_t \exp \left\{ -\frac{1}{8} (y_i^\bullet - \mu_t)^2 \right\} \right) \quad (17)$$

Optimality condition (shown for μ_c only):

$$0 = \frac{\partial \ell(\mathcal{T})}{\partial \mu_c} = \sum_{i=1}^C (y_i^{(c)} - \mu_c) + \underbrace{p(\text{car}, y_i^\bullet | \mu_c, \mu_t)}_{\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\}} \quad (18)$$

$$+ \sum_{i=1}^M \frac{\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\}}{\underbrace{\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\} + \kappa_t \exp \left\{ -\frac{1}{8} (y_i^\bullet - \mu_t)^2 \right\}}_{\substack{p(\text{car}, y_i^\bullet | \mu_c, \mu_t) \\ p(\text{truck}, y_i^\bullet | \mu_c, \mu_t)}}} (y_i^\bullet - \mu_c) \quad (19)$$

Motivation. Missing Values (6)

Log-likelihood:

$$\ell(\mathcal{T}) = C \ln \kappa_c - \frac{1}{2} \sum_{i=1}^C (y_i^{(c)} - \mu_c)^2 + T \ln \kappa_t - \frac{1}{8} \sum_{i=1}^T (y_i^{(t)} - \mu_t)^2 \quad (20)$$

$$+ \sum_{i=1}^M \ln \left(\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\} + \kappa_t \exp \left\{ -\frac{1}{8} (y_i^\bullet - \mu_t)^2 \right\} \right) \quad (21)$$

Optimality condition (shown for μ_c only):

$$0 = \frac{\partial \ell(\mathcal{T})}{\partial \mu_c} = \sum_{i=1}^C (y_i^{(c)} - \mu_c) + \underbrace{p(\text{car} | y_i^\bullet, \mu_c, \mu_t)} \quad (22)$$

$$+ \sum_{i=1}^M \frac{\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\}}{\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\} + \kappa_t \exp \left\{ -\frac{1}{8} (y_i^\bullet - \mu_t)^2 \right\}} (y_i^\bullet - \mu_c) \quad (23)$$

Motivation. Missing Values (7)

Optimality conditions (shown for both μ_c and μ_t):

$$0 = \frac{\partial \ell(\mathcal{T})}{\partial \mu_c} = \sum_{i=1}^C (y_i^{(c)} - \mu_c) + \underbrace{p(\text{car} | y_i^\bullet, \mu_c, \mu_t)}_{\text{from (25)}} \quad (24)$$

$$+ \sum_{i=1}^M \frac{\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\}}{\kappa_c \exp \left\{ -\frac{1}{2} (y_i^\bullet - \mu_c)^2 \right\} + \kappa_t \exp \left\{ -\frac{1}{8} (y_i^\bullet - \mu_t)^2 \right\}} (y_i^\bullet - \mu_c) \quad (25)$$

$$0 = 4 \frac{\partial \ell(\mathcal{T})}{\partial \mu_t} = \sum_{i=1}^T (y_i^{(t)} - \mu_t) + \sum_{i=1}^M p(\text{truck} | y_i^\bullet, \mu_c, \mu_t) (y_i^\bullet - \mu_t) \quad (26)$$

Things to note:

- ◆ Complicated equations for the unknowns μ_c, μ_t
- ◆ Both equations contain μ_c and μ_t (cf. case with no missing variables)

Motivation. Missing Values (8)

Optimality conditions (shown for both μ_c and μ_t):

$$\sum_{i=1}^C (y_i^{(c)} - \mu_c) + \sum_{i=1}^M p(\text{car} | y_i^\bullet, \mu_c, \mu_t) (y_i^\bullet - \mu_c) = 0 \quad (27)$$

$$\sum_{i=1}^T (y_i^{(t)} - \mu_t) + \sum_{i=1}^M p(\text{truck} | y_i^\bullet, \mu_c, \mu_t) (y_i^\bullet - \mu_t) = 0 \quad (28)$$

If $p(\text{car} | y_i^\bullet, \mu_c, \mu_t)$ and $p(\text{truck} | y_i^\bullet, \mu_c, \mu_t)$ **were** known, the estimation would've been easy:

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

$$\sum_{i=1}^C (y_i^{(c)} - \mu_c) + \sum_{i=1}^M q(z_i = \text{car}) (y_i^\bullet - \mu_c) = 0 \quad (29)$$

$$\Rightarrow \mu_c = \frac{\sum_{i=1}^C y_i^{(c)} + \sum_{i=1}^M q(z_i = \text{car}) y_i^\bullet}{C + \sum_{i=1}^M q(z_i = \text{car})} \quad (30)$$

and similarly,

$$\mu_t = \frac{\sum_{i=1}^T y_i^{(t)} + \sum_{i=1}^M q(z_i = \text{truck}) y_i^\bullet}{T + \sum_{i=1}^M q(z_i = \text{truck})} \quad (31)$$

Motivation. Missing Values (9)

$$\mu_c = \frac{\sum_{i=1}^C y_i^{(c)} + \sum_{i=1}^M q(z_i = \text{car}) y_i}{C + \sum_{i=1}^M q(z_i = \text{car})} \quad (32)$$

$$\mu_t = \frac{\sum_{i=1}^T y_i^{(t)} + \sum_{i=1}^M q(z_i = \text{truck}) y_i}{T + \sum_{i=1}^M q(z_i = \text{truck})} \quad (33)$$

- ◆ 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 $q(z_i) = p(z_i | y_i, \mu_c, \mu_t)$ for all $i = 1, 2, \dots, M$
 3. Recompute μ_c, μ_t according to Eqs.(32, 33)
 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:

$$\mu_c = \frac{\sum_{i=1}^M q(z_i = \text{car}) y_i}{\sum_{i=1}^M q(z_i = \text{car})} \quad (34)$$

$$\mu_t = \frac{\sum_{i=1}^M q(z_i = \text{truck}) y_i}{\sum_{i=1}^M q(z_i = \text{truck})} \quad (35)$$

1. Initialize μ_c, μ_t
2. Compute $q(z_i) = p(z_i | y_i, \mu_c, \mu_t)$ for all $i = 1, 2, \dots, M$
3. Recompute μ_c, μ_t according to Eqs.(36, 37)
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).

$$\mu_c = \frac{\sum_{i=1}^M q(z_i = \text{car}) y_i^\bullet}{\sum_{i=1}^M q(z_i = \text{car})} \quad (36)$$

$$\mu_t = \frac{\sum_{i=1}^M q(z_i = \text{truck}) y_i^\bullet}{\sum_{i=1}^M q(z_i = \text{truck})} \quad (37)$$

EM algorithm:

1. Initialize μ_c, μ_t
2. Compute $q(z_i) = p(z_i | y_i^\bullet, \mu_c, \mu_t)$
for all $i = 1, 2, \dots, M$
3. Recompute μ_c, μ_t according to Eqs.(36, 37)
4. If termination condition is met, finish.
Otherwise goto 2.

K-means:

1. ditto
2. $q(z_i = \text{car}) = \mathbb{I}[|y_i^\bullet - \mu_c| < |y_i^\bullet - \mu_t|]$
 $q(z_i = \text{truck}) = \mathbb{I}[|y_i^\bullet - \mu_t| \leq |y_i^\bullet - \mu_c|]$
for all $i = 1, 2, \dots, M$
3. ditto
4. ditto

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

EM algorithm

- ◆ \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
- ◆ θ : model parameters to be estimated.

Goal: Find θ^* using the Maximum Likelihood approach:

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \ell(\theta) = \underset{\theta}{\operatorname{argmax}} \ln p(\mathbf{o}|\theta) \quad (38)$$

Line of thought

Assume that solving this:

$$\underset{\theta}{\operatorname{argmax}} \ln p(\mathbf{o}, \mathbf{z}|\theta) \quad (39)$$

is easy (optimal parameters had \mathbf{z} been known.)

Our goal will be to rewrite Eq. (38) in a way which will involve optimization terms of kind as in Eq. (39).

Lower Bound on the Log Likelihood

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

Marginalizing over missing values (40)

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

Introduction of distribution $q(Z)$ (41)

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

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

Jensen's inequality. Here inequality holds because logarithm is a concave function. (42)

Define

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p(\mathbf{o}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})}. \quad (43)$$

This $\mathcal{L}(q, \boldsymbol{\theta})$ is the lower bound for $\ln p(\mathbf{o}|\boldsymbol{\theta})$ due to Eq. (42), for any distribution q .

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

How Tight Is This Bound? (1)

$$\ln p(\mathbf{o}|\boldsymbol{\theta}) - \mathcal{L}(q, \boldsymbol{\theta}) = \ln p(\mathbf{o}|\boldsymbol{\theta}) - \sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p(\mathbf{o}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} \quad (44)$$

$$= \ln p(\mathbf{o}|\boldsymbol{\theta}) - \sum_{\mathbf{z}} q(\mathbf{z}) \{ \underbrace{\ln p(\mathbf{o}, \mathbf{z}|\boldsymbol{\theta})}_{p(\mathbf{z}|\mathbf{o}, \boldsymbol{\theta})p(\mathbf{o}|\boldsymbol{\theta})} - \ln q(\mathbf{z}) \} \quad (45)$$

$$= \ln p(\mathbf{o}|\boldsymbol{\theta}) - \sum_{\mathbf{z}} q(\mathbf{z}) \{ \ln p(\mathbf{z}|\mathbf{o}, \boldsymbol{\theta}) + \ln p(\mathbf{o}|\boldsymbol{\theta}) - \ln q(\mathbf{z}) \} \quad (46)$$

$$= \ln p(\mathbf{o}|\boldsymbol{\theta}) - \underbrace{\sum_{\mathbf{z}} q(\mathbf{z}) \ln p(\mathbf{o}|\boldsymbol{\theta})}_1 - \sum_{\mathbf{z}} q(\mathbf{z}) \{ \ln p(\mathbf{z}|\mathbf{o}, \boldsymbol{\theta}) - \ln q(\mathbf{z}) \} \quad (47)$$

$$= - \sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p(\mathbf{z}|\mathbf{o}, \boldsymbol{\theta})}{q(\mathbf{z})} \quad (48)$$

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

$$D_{\text{KL}}(q||p) = \sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{o}, \boldsymbol{\theta})} = - \sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p(\mathbf{z}|\mathbf{o}, \boldsymbol{\theta})}{q(\mathbf{z})} \quad (49)$$

How Tight Is This Bound? (2)

$$\ln p(\mathbf{o}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + D_{\text{KL}}(q||p) \quad (50)$$

↑ ↑ ↑

log likelihood lower bound gap

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 $D_{\text{KL}}(q||p) \geq 0$ for any q, p . Additionally,

$$D_{\text{KL}}(q||p) = 0 \quad \Leftrightarrow \quad p = q. \quad (51)$$

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

EM algorithm

$$\ln p(\mathbf{o}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + D_{\text{KL}}(q||p) \quad (52)$$

↑ ↑ ↑
log likelihood lower bound gap

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):

$$q^{(t+1)} = \operatorname{argmax}_q \mathcal{L}(q, \boldsymbol{\theta}^{(t)}) \quad (53)$$

3. **M-step** (Maximization):

$$\boldsymbol{\theta}^{(t+1)} = \operatorname{argmax}_{\boldsymbol{\theta}} \mathcal{L}(q^{(t+1)}, \boldsymbol{\theta}) \quad (54)$$

4. If termination condition is not met, goto 2.

Expectation step

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

$$q^{(t+1)} = \operatorname{argmax}_q \mathcal{L}(q, \boldsymbol{\theta}^{(t)}) \quad (55)$$

$$\mathcal{L}(q, \boldsymbol{\theta}^{(t)}) = \underbrace{\ln p(\mathbf{o} | \boldsymbol{\theta}^{(t)})}_{\text{const.}} - D_{\text{KL}}(q || p) \quad (56)$$

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. (55) is

$$q^{(t+1)}(\mathbf{z}) = p(\mathbf{z} | \mathbf{o}, \boldsymbol{\theta}^{(t)}) . \quad (57)$$

$$\left[\text{Recall: } D_{\text{KL}}(q || p) = - \sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p(\mathbf{z} | \mathbf{o}, \boldsymbol{\theta})}{q(\mathbf{z})} \right] \quad (58)$$

Maximization step

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

$$\boldsymbol{\theta}^{(t+1)} = \operatorname{argmax}_{\boldsymbol{\theta}} \mathcal{L}(q^{(t+1)}, \boldsymbol{\theta}) \quad (59)$$

$$\mathcal{L}(q^{(t+1)}, \boldsymbol{\theta}) = \sum_{\mathbf{z}} q^{(t+1)}(\mathbf{z}) \ln \frac{p(\mathbf{o}, \mathbf{z} | \boldsymbol{\theta})}{q^{(t+1)}(\mathbf{z})} \quad (60)$$

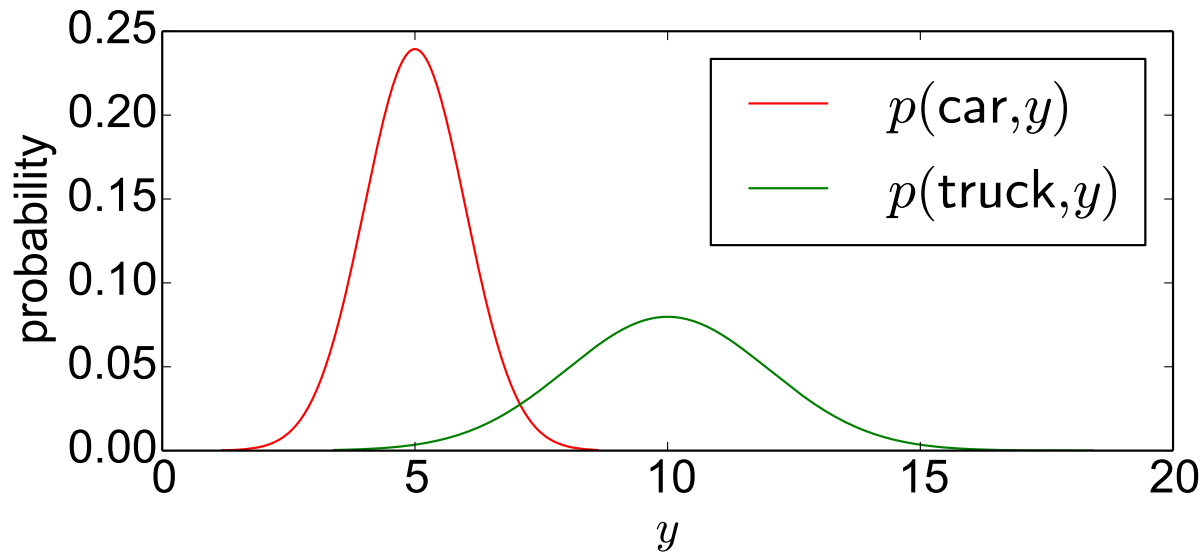
$$= \sum_{\mathbf{z}} q^{(t+1)}(\mathbf{z}) \ln p(\mathbf{o}, \mathbf{z} | \boldsymbol{\theta}) - \underbrace{\sum_{\mathbf{z}} q^{(t+1)}(\mathbf{z}) \ln q^{(t+1)}(\mathbf{z})}_{\text{const.}} \quad (61)$$

Result: The parameters $\boldsymbol{\theta}$ maximizing Eq. (59) are

$$\boldsymbol{\theta}^{(t+1)} = \operatorname{argmax}_{\boldsymbol{\theta}} \sum_{\mathbf{z}} q^{(t+1)}(\mathbf{z}) \ln p(\mathbf{o}, \mathbf{z} | \boldsymbol{\theta}). \quad (62)$$

Example 1 - Setting

$$\pi_c = 0.6, \pi_t = 0.4, \sigma_c = 1, \sigma_t = 2, \mu_c = 5, \mu_t = 10$$



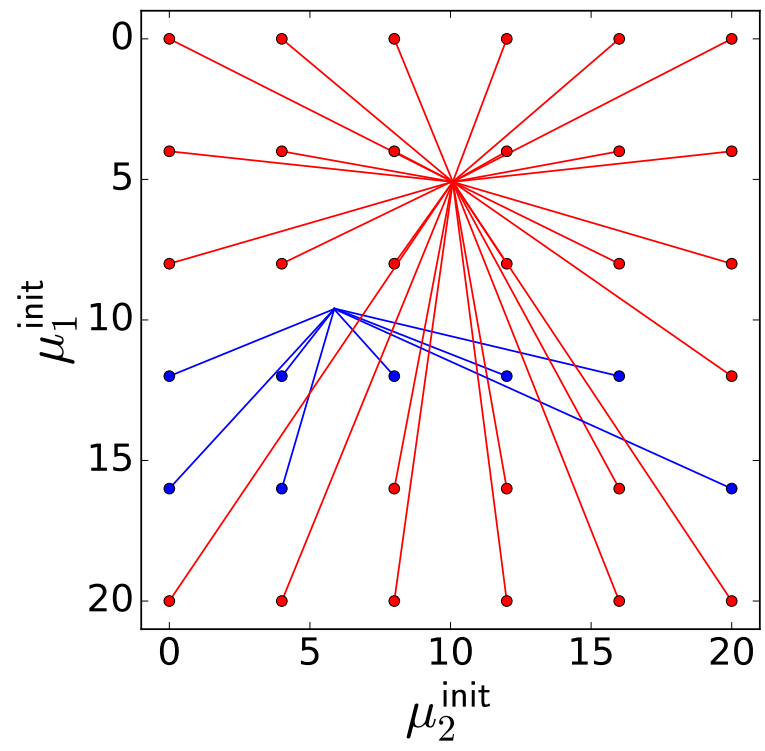
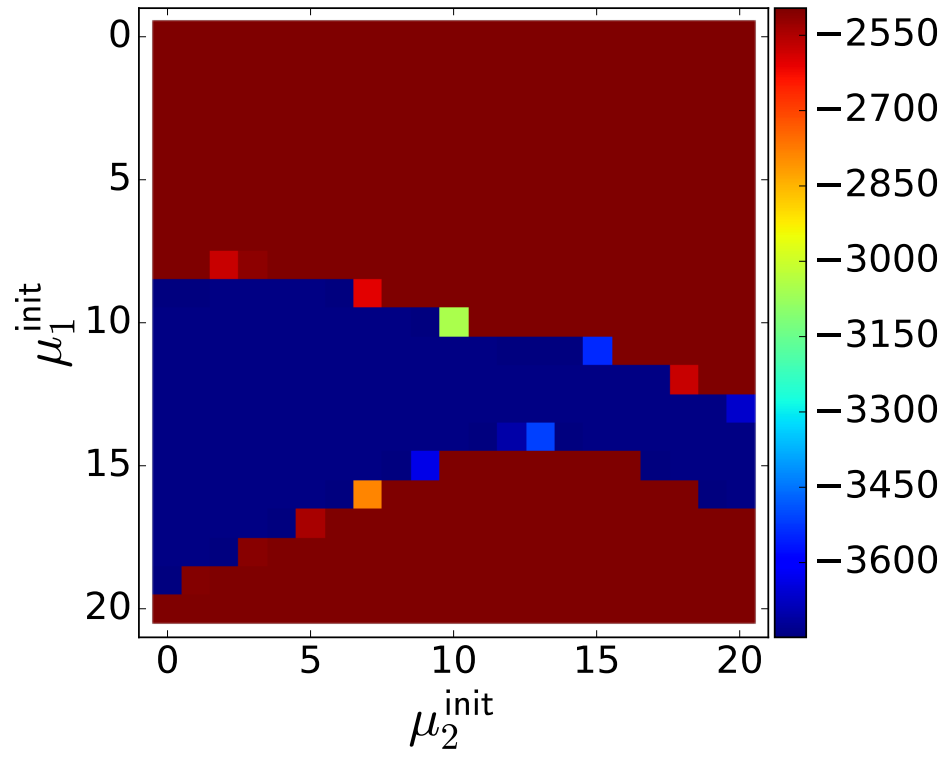
Data:

- ◆ 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 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.

$$\mu_c = \frac{\sum_{i=1}^M q(z_i = \text{car}) y_i^\bullet}{\sum_{i=1}^M q(z_i = \text{car})} \quad (63)$$

$$\sigma_c^2 = \frac{\sum_{i=1}^M q(z_i = \text{car}) (y_i^\bullet - \mu_c)^2}{\sum_{i=1}^M q(z_i = \text{car})} \quad (64)$$

$$\pi_c = \frac{\sum_{i=1}^M q(z_i = \text{car})}{M} \quad (65)$$