Expectation Maximization Algorithm

Intuition

- Consider coin-flipping experiment
  - Assume that we have 2 coins, A and B
  - The probability of getting heads with coin A is θ_A
  - The probability of getting heads with coin B is θ_B
  - Our goal is to estimate θ_A, θ_B by performing a number of trials

- Experiment detail
  - Choose one of the two coins (with equal probability)
  - Perform ten independent flipping trials with selected coin
  - Repeat the experiment five times
  - Keep track two vectors

$$x = (x_1, x_2, ..., x_5) \qquad z = (z_1, z_2, ..., z_5)$$

http://www.cmi.ac.in/~madhavan/courses/datamining12/reading/em-tutorial.pdf

Intuition

		Coin A	Coin B
B	H T T T H H T H T H		5 H, 5 T
A	H H H H T H H H H H	9 H, 1 T	
A	H T H H H H H T H H	8 H, 2 T	
B	H T H T T T H H T T		4 H, 6 T
A	T H H H T H H H T H	7 H, 3 T	
		24 H, 6 T	9 H, 11 T

http://www.cmi.ac.in/~madhavan/courses/datamining12/reading/em-tutorial.pdf

Intuition

$$\hat{\theta}_A = \frac{\text{\# of heads using coin A}}{\text{total \# of flips using coin A}}$$

$$\hat{\theta}_B = \frac{\text{\# of heads using coin B}}{\text{total \# of flips using coin B}}$$

$$\hat{\theta}_A = \frac{24}{24+6} = 0.80$$

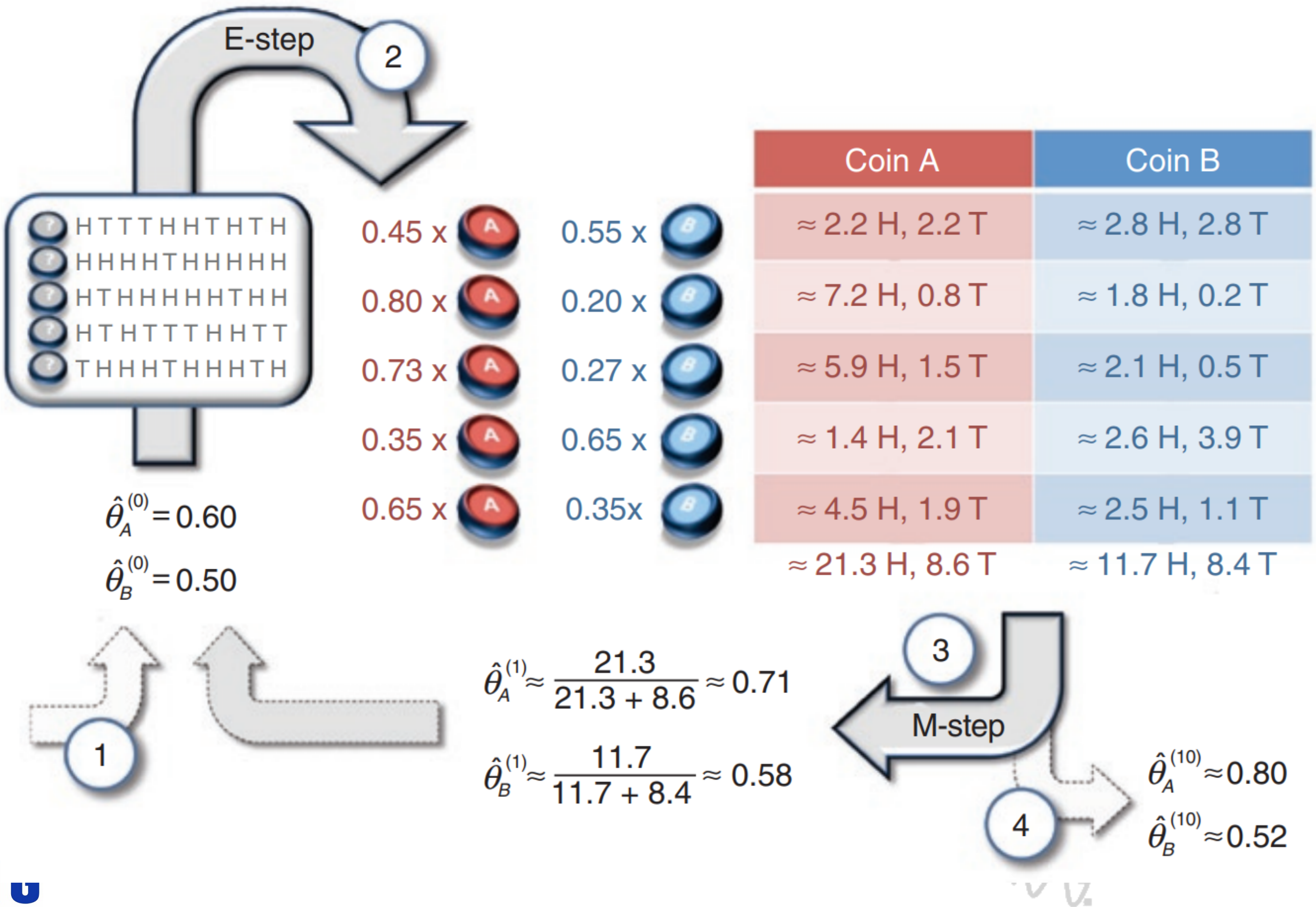
$$\hat{\theta}_B = \frac{9}{9+11} = 0.45$$

Coin A	Coin B
	5 H, 5 T
9 H, 1 T	
8 H, 2 T	
	4 H, 6 T
7 H, 3 T	
24 H, 6 T	9 H, 11 T

http://www.cmi.ac.in/~madhavan/courses/datamining12/reading/em-tutorial.pdf

Intuition

- What if we only know head counts x but not identities z
  - Computing proportions of heads for each coin is no longer possible
  - If we had some way of completing the data, then we could estimate maximum likelihood with complete data



Intuition

- Expectation step
  - Guessing a probability distribution over completions of missing data given the current model
  - Not to form the probability distribution over completion explicitly, but rather to only compute expected statistics over completions

- Maximization step
  - Re-estimating the model parameters using these complete data
  - Model re-estimation can be thought of as maximization of the expected log-likelihood of the data

http://www.cmi.ac.in/~madhavan/courses/datamining12/reading/em-tutorial.pdf

EM Algorithm in General

- Maximize the likelihood function

$$p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

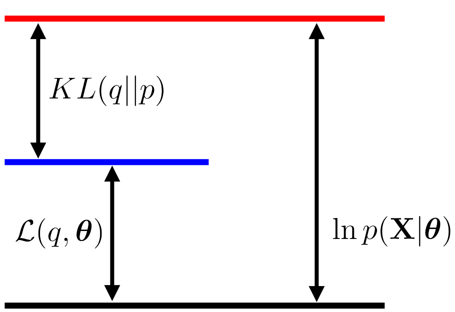
- Following decomposition holds for any choice of $q(\mathbf{Z})$

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \underbrace{\mathcal{L}(q, \boldsymbol{\theta})}_{\text{Lower bound on the likelihood}} + \underbrace{KL(q||p)}_{\text{Always } \geq 0}$$

where we have defined

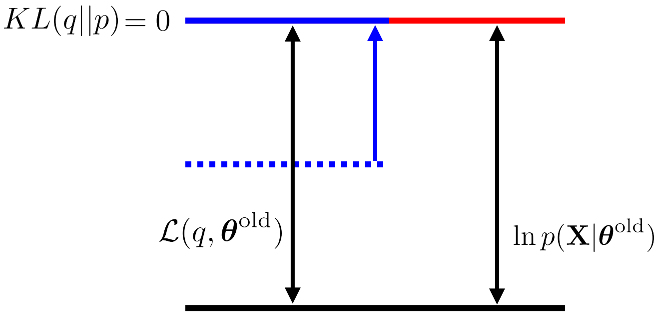
$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})}$$

$$KL(q||p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})}$$



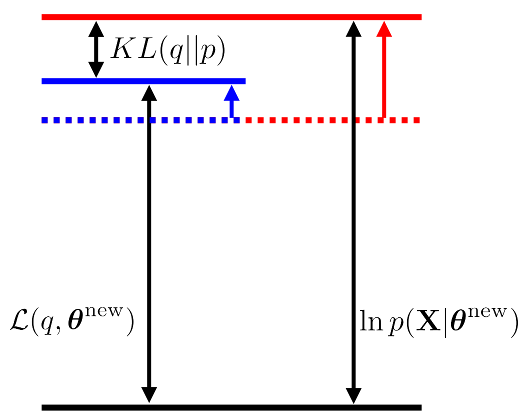
EM Algorithm in General

- Current value of model parameter is $\boldsymbol{\theta}^{\mathrm{old}}$



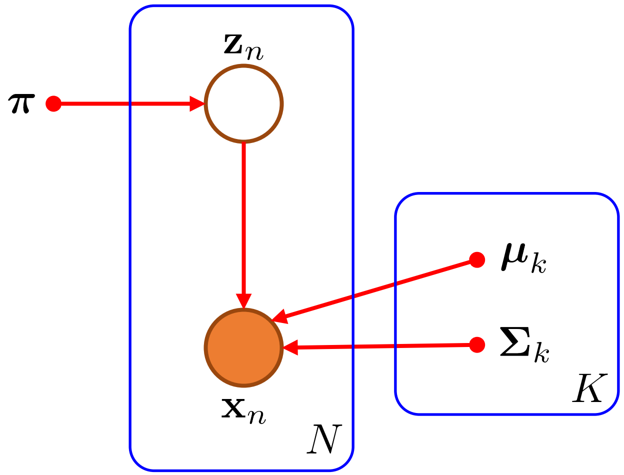
- E-step
  - The lower bound $\mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}})$ is maximized with respect to $q(\mathbf{Z})$ while holding $\boldsymbol{\theta}^{\mathrm{old}}$ fixed

- M-step
  - The distribution $q(\mathbf{Z})$ is fixed and the lower bound is maximized with respect to $\boldsymbol{\theta}$ to give new value $\boldsymbol{\theta}^{\mathrm{new}}$



Pattern Recognition and Machine Learning by Christopher M. Bishop

Mixtures of Gaussians



$$p(\mathbf{z}_n) = \prod_{k=1}^{K} \pi_k^{z_{nk}}$$

$$p(\mathbf{x}_n|\mathbf{z}_n) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

EM for Mixtures of Gaussians

① Initialize the means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$ and mixing coefficients π_k, and evaluate the initial value of the log likelihood

② E-step: evaluate the responsibilities using the current parameter values

$$p(z_{nk} = 1|\mathbf{x}_n) = \gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

③ M-step: Re-estimate the parameters using the current responsibilities

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}})^T$$

$$\pi_k^{\text{new}} = \frac{N_k}{N}$$

where $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$

④ Evaluate the log likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$