K-Means Clustering and Gaussian Mixture Model

Il-Chul Moon
Dept. of Industrial and Systems Engineering
KAIST
icmoon@kaist.ac.kr

Weekly Objectives

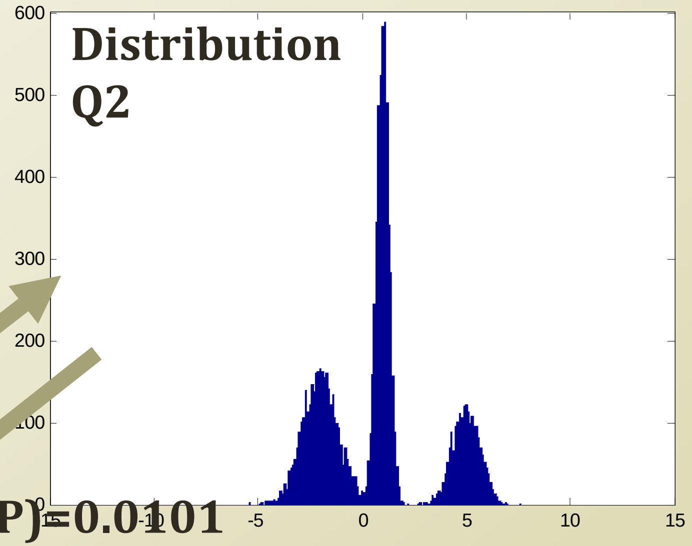
- Understand the clustering task and the K-means algorithm
  - Know what the unsupervised learning is
  - Understand the K-means iterative process
  - Know the limitation of the K-means algorithm
- Understand the Gaussian mixture model
  - Know the multinomial distribution and the multivariate Gaussian distribution
  - Know why mixture models are useful
  - Understand how the parameter updates are derived from the Gaussian mixture model
- Understand the EM algorithm
  - Know the fundamentals of the EM algorithm
  - Know how to derive the EM updates of a model

Maximizing the Lower Bound (1)

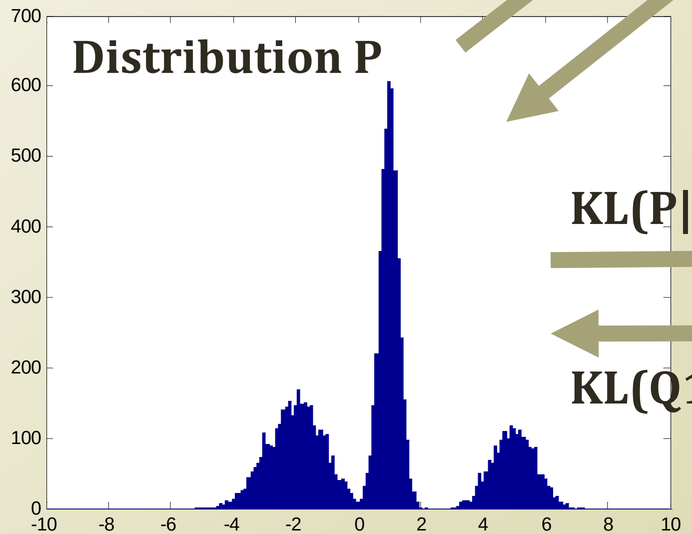
- $l(\theta) = \ln P(X|\theta) = \ln\left\{\sum_Z q(Z)\frac{P(X,Z|\theta)}{q(Z)}\right\} \geq \sum_Z q(Z)\ln\frac{P(X,Z|\theta)}{q(Z)} = Q(\theta, q)$
  - $Q(\theta, q) = E_{q(Z)}\ln P(X,Z|\theta) + H(q)$
- The other storyline is
  - $l(\theta) \geq \sum_Z q(Z)\ln\frac{P(X,Z|\theta)}{q(Z)} = \sum_Z q(Z)\ln\frac{P(Z|X,\theta)P(X|\theta)}{q(Z)}$
  - $= \sum_Z\{q(Z)\ln\frac{P(Z|X,\theta)}{q(Z)} + q(Z)\ln P(X|\theta)\} = \ln P(X|\theta) + \sum_Z\{q(Z)\ln\frac{P(Z|X,\theta)}{q(Z)}\}$
  - $L(\theta, q) = \ln P(X|\theta) - \sum_Z\{q(Z)\ln\frac{q(Z)}{P(Z|X,\theta)}\}$
- Here, the second term is a very special term
  - $KL(q(Z)||P(Z|X,\theta)) = \sum_Z\{q(Z)\ln\frac{q(Z)}{P(Z|X,\theta)}\}$
  - Kullback-Leiber divergence, or KL divergence: $KL(P||Q) = \sum_i P(i)\ln(\frac{P(i)}{Q(i)})$
  - Non-symmetric measure of the difference between two probability distributions, or KL(P||Q)
  - Measures the difference
    - $KL(P||Q) \geq 0$
    - When there is no difference between P and Q, KL(P||Q)=0

KL Divergence

- Kullback-Leiber divergence, or KL divergence: $KL(P||Q) = \sum_i P(i) \ln(\frac{P(i)}{Q(i)})$
  - Measures the matching performance of P and Q
  - Consider Gaussian distribution and Gaussian mixture distribution

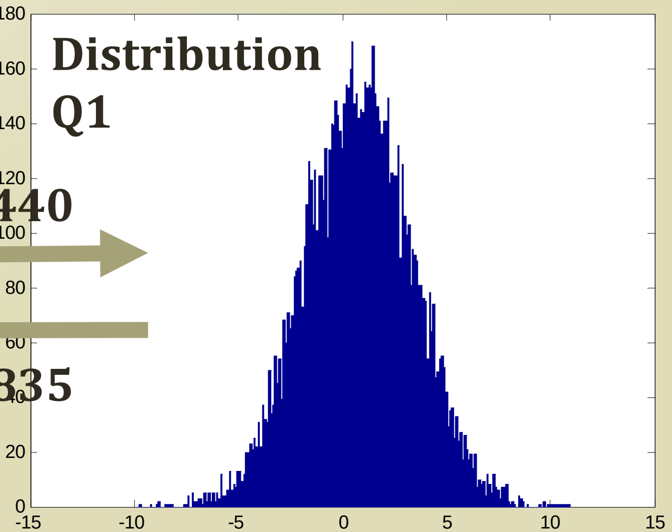


KL(P||Q2)=0.0104 KL(Q2||P)=0.0101



KL(P||Q1)=0.5440

KL(Q1||P)=0.9835



Maximizing the Lower Bound (2)

- $l(\theta) = \ln P(X|\theta) = \ln\left\{\sum_Z q(Z)\frac{P(X,Z|\theta)}{q(Z)}\right\} \geq \sum_Z q(Z)\ln\frac{P(X,Z|\theta)}{q(Z)} = Q(\theta,q)$
  - $Q(\theta,q) = E_{q(Z)}\ln P(X,Z|\theta) + H(q)$
  - $L(\theta,q) = \ln P(X|\theta) - \sum_Z\{q(Z)\ln\frac{q(Z)}{P(Z|X,\theta)}\}$
- Why do we compute $L(\theta,q)$?
  - We do not know how to optimize $Q(\theta,q)$ without further knowledge of $q(Z)$
  - The second term of $L(\theta,q)$ tells how to set $q(Z)$
    - The first term is fixed when θ is fixed ***at time t***
    - The second term can be minimized to maximize $L(\theta,q)$
      - $KL(q(Z)||P(Z|X,\theta)) = 0 \rightarrow q^t(Z) = P(Z|X,\theta^t)$
  - Now, the lower bound with optimized q is
    - $Q(\theta,q^t) = E_{q^t(Z)}\ln P(X,Z|\theta^t) + H(q^t)$
- Then, optimizing θ to retrieve the tight lower bound is
  - $\theta^{t+1} = argmax_\theta Q(\theta,q^t) = argmax_\theta E_{q^t(Z)}\ln P(X,Z|\theta)$
    - $q^t(Z) \rightarrow$ Distribution parameters for latent variable is at time t
    - $\ln P(X,Z|\theta) \rightarrow$ optimized log likelihood parameters is at time t+1

Tells how to setup Z by setting $q^t(Z) = P(Z|X,\theta^t)$

Relax the KL divergence by updating θ^t to θ^{t+1}

Graphical Interpretation of Lower Bound Maximization

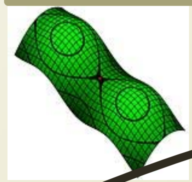
- $l(\theta) = \ln P(X|\theta) \geq L(\theta, q)$

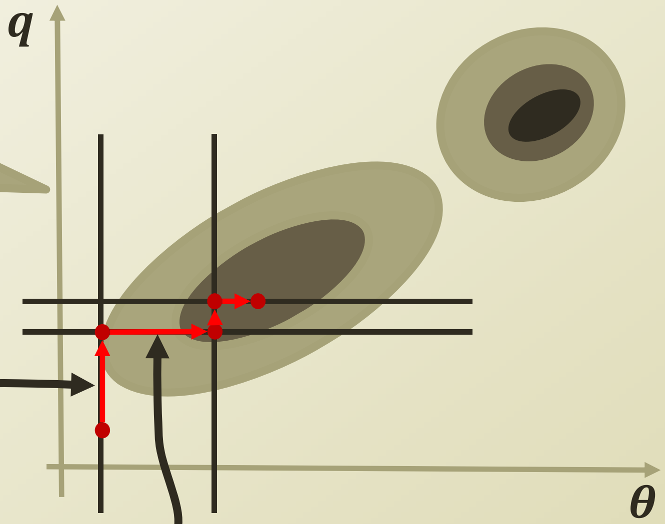
$$= \ln P(X|\theta) - \sum_{Z} \left\{ q(Z) \ln \frac{q(Z)}{P(Z|X,\theta)} \right\}$$

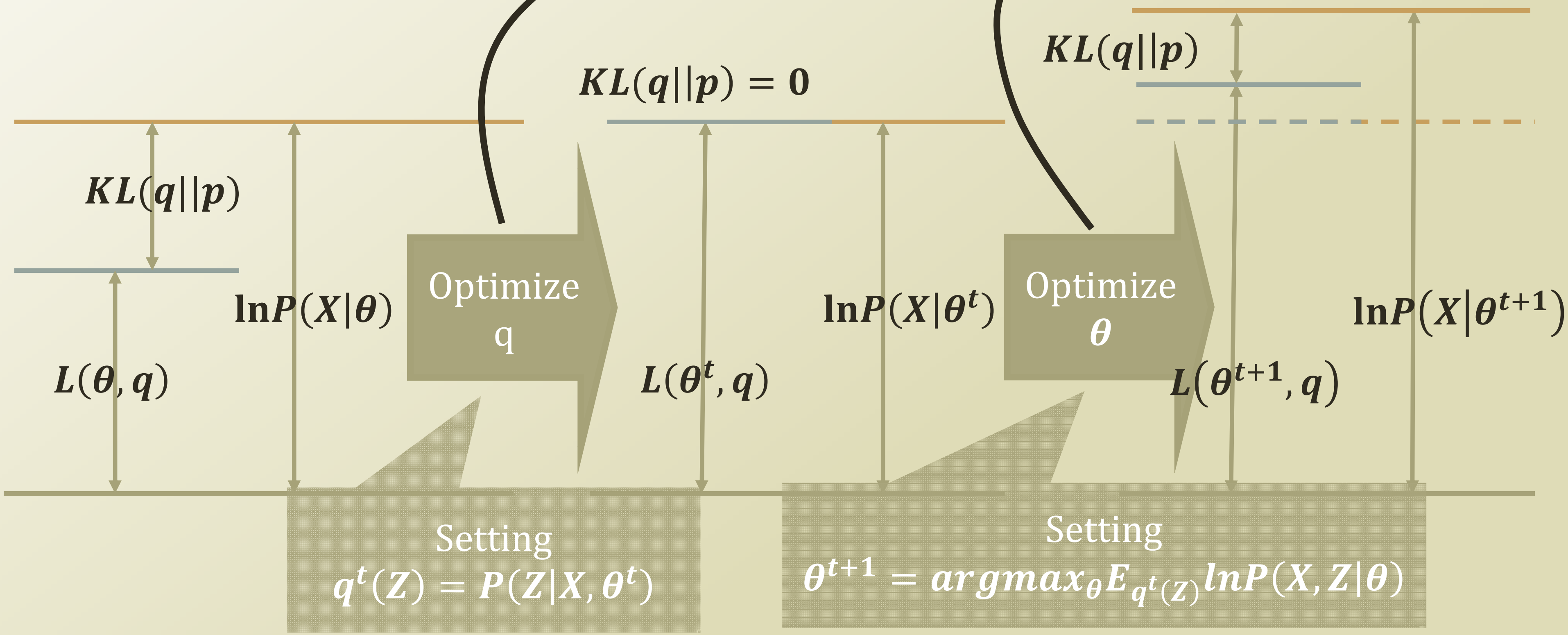
- $\ln P(X|\theta) = L(\theta, q) + \sum_Z \{ q(Z) \ln \frac{q(Z)}{P(Z|X,\theta)} \}$

$$= L(\theta, q) + KL(q||p)$$

Fall into a local maxima or ???







$$l(\theta) = \ln P(X|\theta) = \ln\left\{\sum_Z q(Z)\frac{P(X,Z|\theta)}{q(Z)}\right\} \geq \sum_Z q(Z)\ln\frac{P(X,Z|\theta)}{q(Z)} = Q(\theta,q)$$
$$Q(\theta,q) = E_{q(Z)}\ln P(X,Z|\theta) + H(q)$$
$$L(\theta,q) = \ln P(X|\theta) - \sum_Z \{q(Z)\ln\frac{q(Z)}{P(Z|X,\theta)}\}$$

EM Algorithm

- EM algorithm
  - Finds the maximum likelihood solutions for models with latent variables
  - $P(X|\theta) = \sum_Z P(X,Z|\theta) \rightarrow \ln P(X|\theta) = \ln\{\sum_Z P(X,Z|\theta)\}$
- EM algorithm
  - Initialize θ^0 to an arbitrary point
  - Loop until the likelihood converges
    - Expectation step
      - $q^{t+1}(z) = argmax_q Q(\theta^t,q) = argmax_q L(\theta^t,q) = argmin_q KL(q||P(Z|X,\theta^t))$
      - → $q^t(z) = P(Z|X,\theta)$→Assign Z by $P(Z|X,\theta)$
    - Maximization step
      - $\theta^{t+1} = argmax_\theta Q(\theta,q^{t+1}) = argmax_\theta L(\theta,q^{t+1})$
      - → fixed Z means that there is no unobserved variables
      - → Same optimization of ordinary MLE

Rethinking GMM Learning Process

- GMM, K-Means
  - We used EM algorithm to find the assignment of latent variables and the related distribution parameters
- EM algorithm
  - Initialize θ^0 to an arbitrary point
  - Loop until the likelihood converges
    - Expectation step
      - Assign Z by $P(Z|X,\theta)$
      - $\gamma(z_{nk}) \equiv p(z_k = 1|x_n) = \frac{P(z_k=1)P(x|z_k = 1)}{\sum_{j=1}^{K} P(z_j=1)P\left(x|z_j = 1\right)} = \frac{\pi_k N(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N\left(x|\mu_j, \Sigma_j\right)}$
    - Maximization step
      - Same optimization of ordinary MLE
      - $\frac{d}{d\mu_k}\ln P(X|\pi,\mu,\Sigma) = 0, \frac{d}{d\Sigma_k}\ln P(X|\pi,\mu,\Sigma) = 0, \frac{d}{d\pi_k}\ln P(X|\pi,\mu,\Sigma) + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right) = 0$
      - $\widehat{\mu_k} = \frac{\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{x}_n}{\sum_{n=1}^{N}\gamma(z_{nk})}, \Sigma_k = \frac{\sum_{n=1}^{N}\gamma(z_{nk})(\boldsymbol{x}_n - \widehat{\mu_k})(\boldsymbol{x}_n - \widehat{\mu_k})^{\boldsymbol{T}}}{\sum_{n=1}^{N}\gamma(z_{nk})}, \pi_k = \frac{\sum_{n=1}^{N}\gamma(z_{nk})}{N}$

Further Readings

- Bishop Chapter 2 and 9
- Murphy Chapter 11