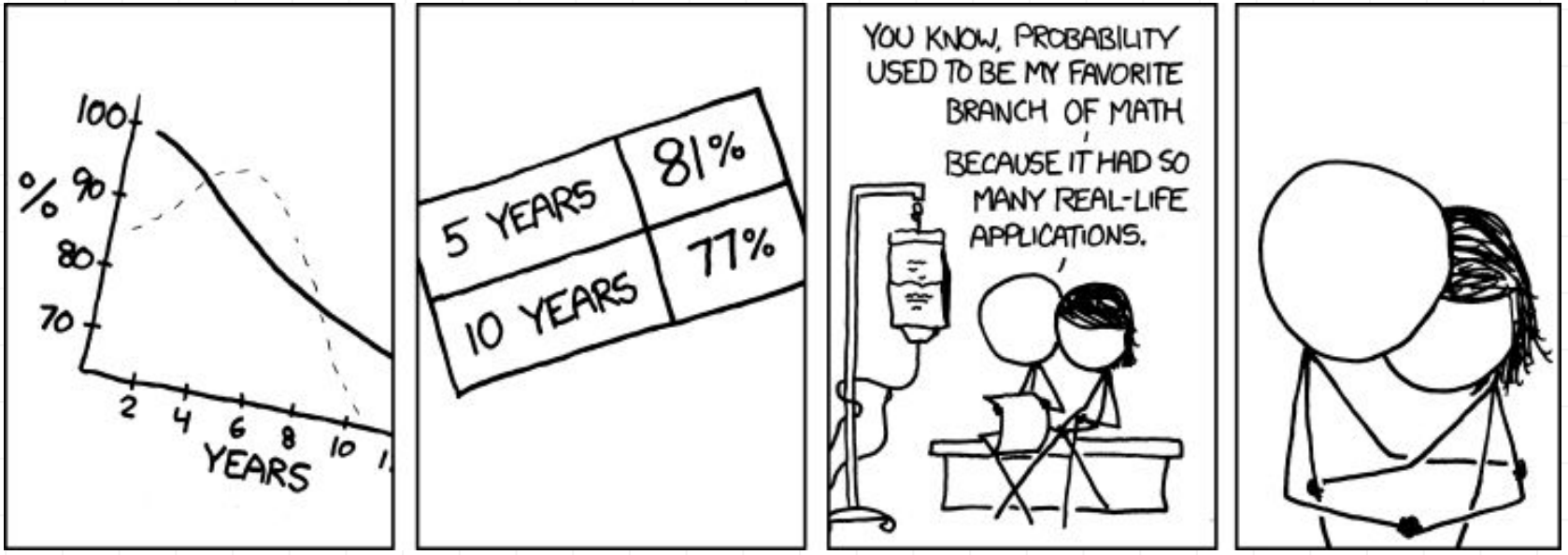
https://xkcd.com/881/



Mixture Models and Expectation Maximisation

Pre-read/watch: K-Means and Gaussian Mixture Models

EM revisited

- An alternative view of EM
- Connections between GMM and K-means
- Bernoulli mixture

TODO: write out the corresponding book section for pre-read and coverage

EM in general - does it really maximise likelihood, and why?

Practical considerations and other topics
- impossibility of clustering [Kleinberg 2003]

- Kmeans++ [Vassilvitskii and Arthur, 2006]

(initialise a set of cluster centers / component means)

Expectation step

K-means

Gaussian Mixture Models

Re-assign data points to clusters, determine r_{nk}

2. **E step**. Evaluate the responsibilities using the current parameter values

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \tag{9.23}$$

Maximisation step

Re-compute the cluster means - update $\{\boldsymbol{\mu}_k\}$

3. **M step**. Re-estimate the parameters using the current responsibilities

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \tag{9.24}$$

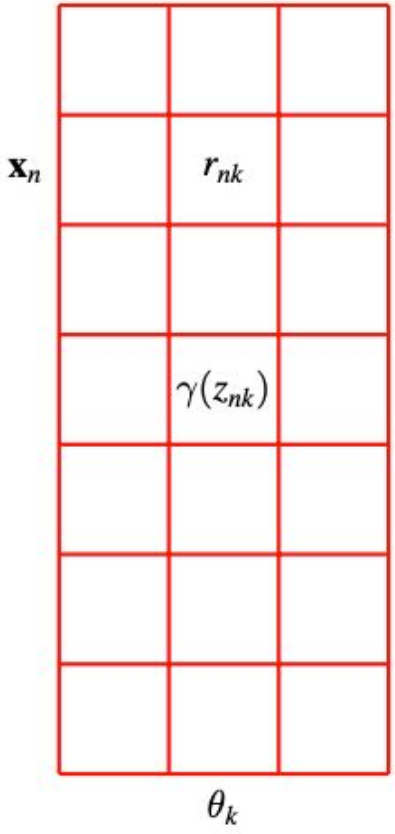
$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}) (\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}})^{\mathrm{T}} \tag{9.25}$$

$$\pi_k^{\text{new}} = \frac{N_k}{N} \tag{9.26}$$

where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}). \tag{9.27}$$

K-means and GMM - hard vs soft assignments



For k-means clustering,
we have hard assignments

For GMM,
we have soft assignments

Assume a Gaussian mixture model.

Covariance matrices given by $\epsilon\mathbf{I}$, where ϵ is shared by all components.

Then

$$p(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi\epsilon)^{D/2}} \exp\left\{-\frac{1}{2\epsilon}\|\mathbf{x} - \boldsymbol{\mu}_k\|^2\right\}.$$

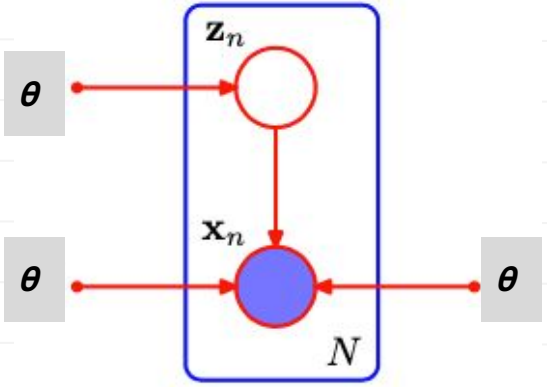
$$\gamma(z_{nk}) = \frac{\pi_k \exp\left\{-\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2/2\epsilon\right\}}{\sum_j \pi_j \exp\left\{-\|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2/2\epsilon\right\}}$$

Taking the limit $\epsilon \to 0$

$$\gamma(z_{nk}) = \begin{cases} 1 & \text{if } \|\mathbf{x}_n - \boldsymbol{\mu}_k\| < \|\mathbf{x}_n - \boldsymbol{\mu}_j\| \quad \forall j \neq k \\ 0 & \text{otherwise} \end{cases}$$

Expectation-maximization revisited

Fig 9.6



X observed, Z “latent”

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right\}. \qquad (9.29)$$

compute posterior P(Z | X, $\boldsymbol{\theta}$)

Take expectation of the complete data likelihood P(X, Z | $\boldsymbol{\theta}$) w.r.t. Z

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}). \qquad (9.30)$$

Maximise this expectation w.r.t. $\boldsymbol{\theta}$

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}). \qquad (9.31)$$

The General EM Algorithm

Given a joint distribution $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ over observed variables $\mathbf{X}$ and latent variables $\mathbf{Z}$, governed by parameters $\boldsymbol{\theta}$, the goal is to maximize the likelihood function $p(\mathbf{X}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

1. Choose an initial setting for the parameters $\boldsymbol{\theta}^{\text{old}}$.

2. **E step** Evaluate $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.

3. **M step** Evaluate $\boldsymbol{\theta}^{\text{new}}$ given by

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) \tag{9.32}$$

where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}). \tag{9.33}$$

4. Check for convergence of either the log likelihood or the parameter values. If the convergence criterion is not satisfied, then let

$$\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}} \tag{9.34}$$

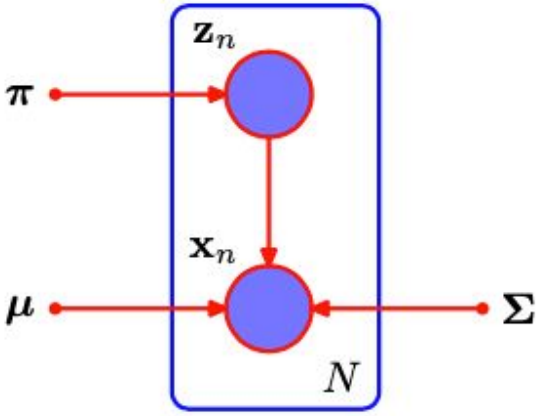
and return to step 2.

MAP objective

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) + \ln p(\boldsymbol{\theta})$$

EM for GMM, revisited

Figure 9.9 This shows the same graph as in Figure 9.6 except that we now suppose that the discrete variables $\mathbf{z}_n$ are observed, as well as the data variables $\mathbf{x}_n$.



$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}. \quad (9.36)$$

$$\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}. \quad (9.40)$$

K-means and GMM - differences & connections

	μ_k	Σ_k	π_k	N_k
GMM	$\frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n$	$\frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n-\boldsymbol{\mu}_k)(\mathbf{x}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}$	$\frac{N_k}{N}$	$\sum_{n=1}^{N}\gamma(z_{nk}).$
K-Means	$\frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}.$	ϵI	Does not matter if >0	$\Sigma_n\ r_{nk}$

Complexity: how many parameters for each?
Implementation: can a cluster center “die” in each model, how to handle?

K-means and GMM - differences & connections

GMM

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}. \tag{9.14}$$

K-Means

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2 \tag{9.1}$$

D separate Bernoulli Distributions

a set of D binary variables x_i , each governed by Bernoulli distribution with mean $\boldsymbol{\mu}_i$

$$\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}} \text{ and } \boldsymbol{\mu} = (\mu_1, \ldots, \mu_D)^{\mathrm{T}}$$

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{i=1}^{D} \mu_i^{x_i}(1-\mu_i)^{(1-x_i)} \tag{9.44}$$

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu} \tag{9.45}$$

$$\mathrm{cov}[\mathbf{x}] = \mathrm{diag}\{\mu_i(1-\mu_i)\}. \tag{9.46}$$

Mixture of Bernoulli Distributions

$$\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}}$$

$$\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}, \boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$$

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|\boldsymbol{\mu}_k) \tag{9.47}$$

$$p(\mathbf{x}|\boldsymbol{\mu}_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i}(1 - \mu_{ki})^{(1-x_i)}. \tag{9.48}$$

$$\mathbb{E}[\mathbf{x}] = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k \tag{9.49}$$

$$\mathrm{cov}[\mathbf{x}] = \sum_{k=1}^{K} \pi_k \left\{\boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^{\mathrm{T}}\right\} - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^{\mathrm{T}} \tag{9.50}$$

$$\boldsymbol{\Sigma}_k = \mathrm{diag}\left\{\mu_{ki}(1 - \mu_{ki})\right\}$$

Data likelihood :(

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k) \right\}. \tag{9.51}$$

$$p(\mathbf{x}|\mathbf{z}, \boldsymbol{\mu}) = \prod_{k=1}^{K} p(\mathbf{x}|\boldsymbol{\mu}_k)^{z_k} \qquad p(\mathbf{z}|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_k}.$$

Complete data likelihood

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left\{ \ln \pi_k + \sum_{i=1}^{D} [x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1 - \mu_{ki})] \right\} \tag{9.54}$$

Expectations of complete data likelihood

$$\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\pi})] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \left\{ \ln \pi_k + \sum_{i=1}^{D} [x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1 - \mu_{ki})] \right\} \tag{9.55}$$

where

$$\gamma(z_{nk}) = \mathbb{E}[z_{nk}] = \frac{\pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k)}{\sum_{j=1}^{K} \pi_j p(\mathbf{x}_n|\boldsymbol{\mu}_j)}. \tag{9.56}$$

Similar calculation as with mixture of Gaussian

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

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

$$\bar{\mathbf{x}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \qquad \mu_k = \bar{\mathbf{x}}$$

$$\pi_k = \frac{N_k}{N}$$

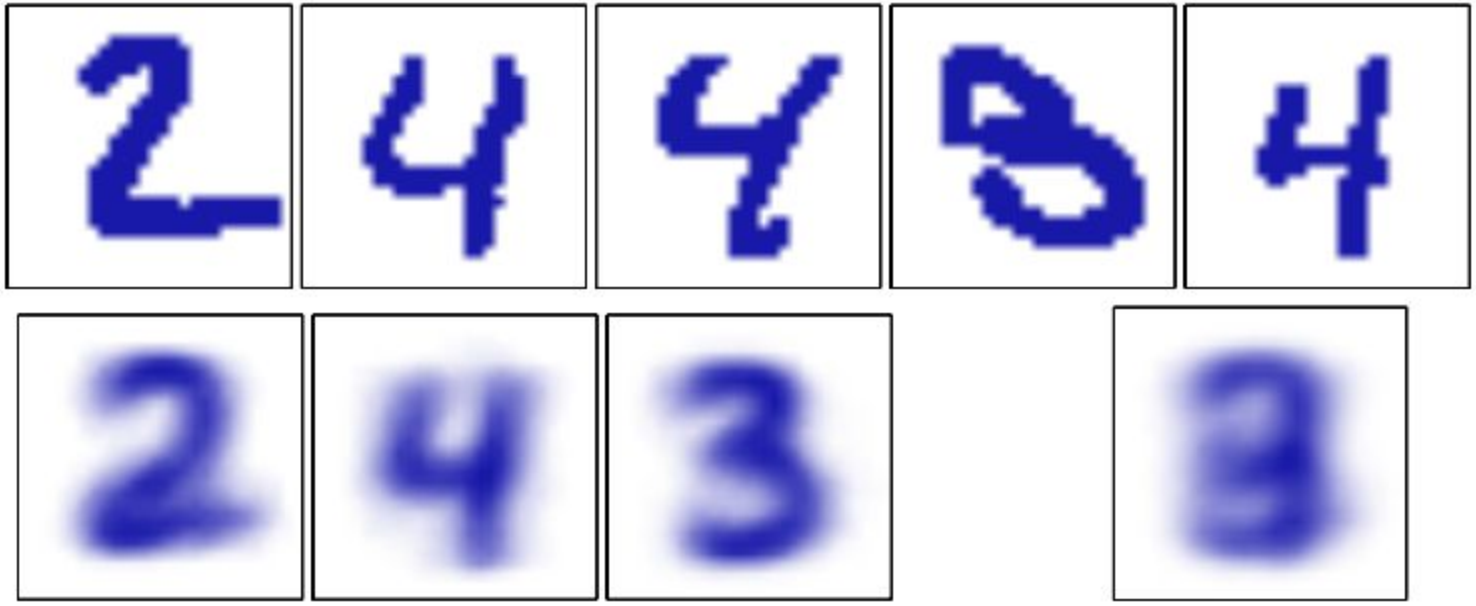


Figure 9.10 Illustration of the Bernoulli mixture model in which the top row shows examples from the digits data set after converting the pixel values from grey scale to binary using a threshold of 0.5. On the bottom row the first three images show the parameters μ_{ki} for each of the three components in the mixture model. As a comparison, we also fit the same data set using a single multivariate Bernoulli distribution, again using maximum likelihood. This amounts to simply averaging the counts in each pixel and is shown by the right-most image on the bottom row.

Outline

EM revisited

- Connections between GMM and K-means
- Bernoulli mixture

EM in general - does it really maximise likelihood, and why?

Practical considerations and other topics
- impossibility of clustering [Kleinberg 2003]

- Kmeans++ [Vassilvitskii and Arthur, 2006]

KL divergence (reminder)

$$\mathrm{KL}(p\|q) = -\int p(\mathbf{x}) \ln q(\mathbf{x})\, d\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x})\, d\mathbf{x}\right)$$
$$= -\int p(\mathbf{x}) \ln \left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\} d\mathbf{x}. \qquad (1.113)$$

If we have “incorrectly” represented p(x) with q(x), how much more *information* do we need to recover p(x)?

Apply Jensen’s inequality, -ln() is convex

$$f\left(\int \mathbf{x} p(\mathbf{x})\, d\mathbf{x}\right) \leqslant \int f(\mathbf{x}) p(\mathbf{x})\, d\mathbf{x}. \qquad (1.117)$$

$$\mathrm{KL}(p\|q) = -\int p(\mathbf{x}) \ln \left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\} d\mathbf{x} \geqslant -\ln \int q(\mathbf{x})\, d\mathbf{x} = 0 \qquad (1.118)$$

Equality will only hold iff. p(x) = q(x) for all x

The EM algorithm in general

Goal: show that the EM algorithm maximises the likelihood function.

$$p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}). \tag{9.69}$$

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q\|p) \tag{9.70}$$

where we have defined

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\} \tag{9.71}$$

$$\mathrm{KL}(q\|p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right\}. \tag{9.72}$$

Derive (9.70)

likelihood ← complete data likelihood

$$p(X|\theta) = \sum_Z p(X, Z|\theta)$$

sum rule

$$\ln p(X|\theta) + \ln p(Z|X,\theta) = \ln p(X, Z|\theta)$$

$$\ln p(X|\theta) = \ln p(X, Z|\theta) - \ln p(Z|X,\theta)$$

introduce $q(Z)$, $\sum_Z q(Z) = 1$

$$= \sum_Z q(Z) \ln p(X|\theta)$$

$$= \sum_Z q(Z) \left[\ln p(X, Z|\theta) - \ln p(Z|X,\theta) + \ln q(Z) - \ln q(Z) \right]$$

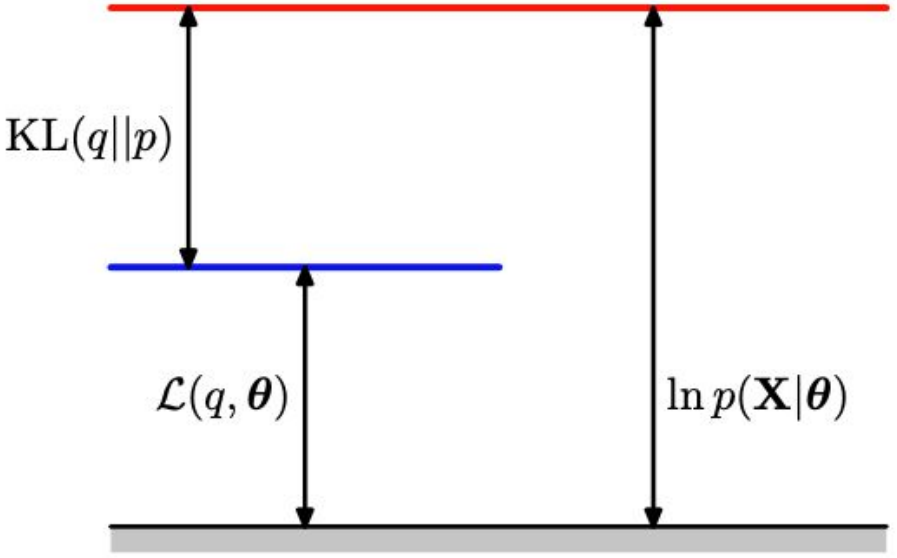
$$= \sum_Z q(Z) \left[\ln p(X, Z|\theta) - \ln q(Z) \right] - \sum_Z q(Z) \ln \frac{p(Z|X,\theta)}{q(Z)}$$

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

Illustrating the decomposition

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q\|p) \tag{9.70}$$

Figure 9.11 Illustration of the decomposition given by (9.70), which holds for any choice of distribution $q(\mathbf{Z})$. Because the Kullback-Leibler divergence satisfies $\mathrm{KL}(q\|p) \geqslant 0$, we see that the quantity $\mathcal{L}(q, \boldsymbol{\theta})$ is a lower bound on the log likelihood function $\ln p(\mathbf{X}|\boldsymbol{\theta})$.



E step

set q(Z) = p(Z | X, $\boldsymbol{\theta}$)

$$\mathrm{KL}(q\|p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right\}. \qquad (9.72)$$

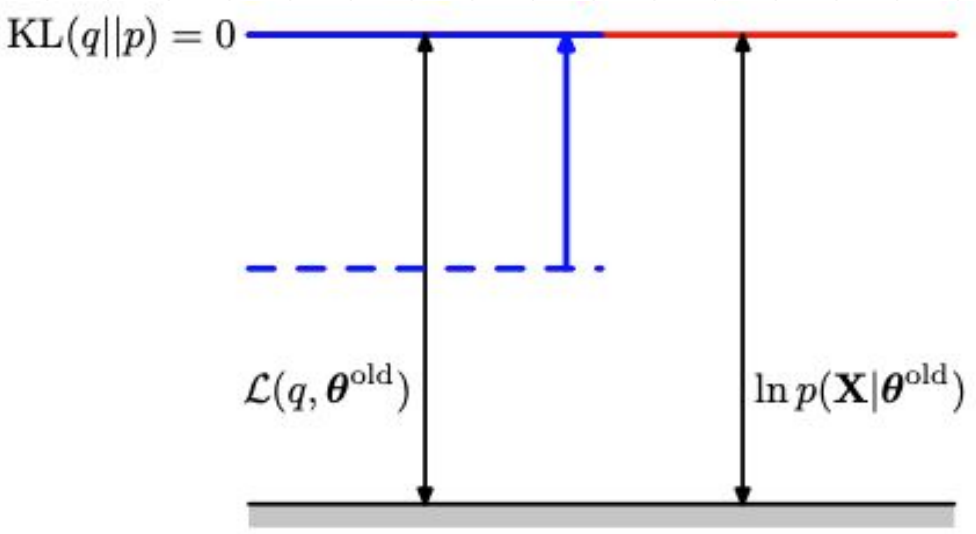


Figure 9.12 Illustration of the E step of the EM algorithm. The q distribution is set equal to the posterior distribution for the current parameter values $\boldsymbol{\theta}^{\mathrm{old}}$, causing the lower bound to move up to the same value as the log likelihood function, with the KL divergence vanishing.

3. **M step** Evaluate $\boldsymbol{\theta}^{\text{new}}$ given by

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) \tag{9.32}$$

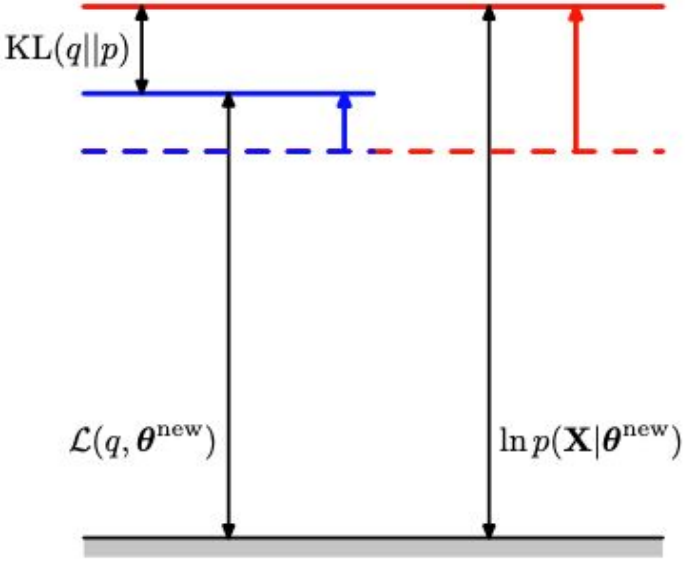
where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}). \tag{9.33}$$

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\} \tag{9.71}$$

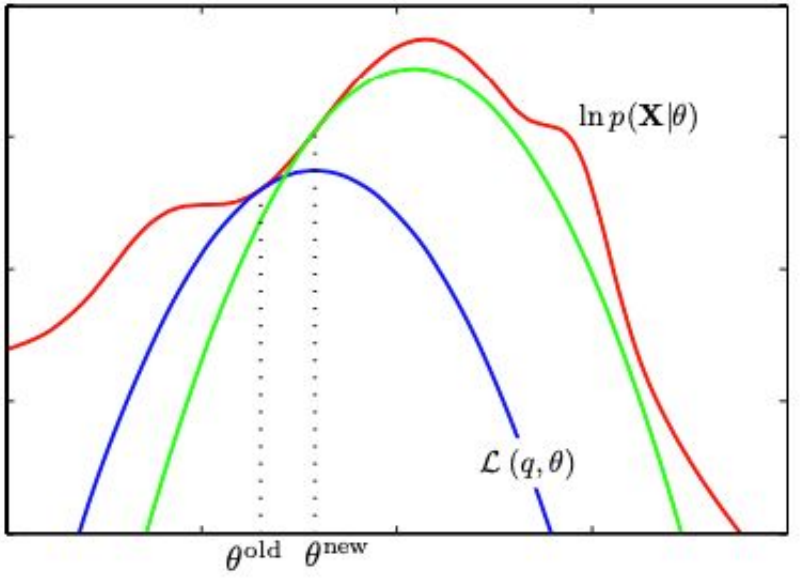
$q(\mathbf{Z})$ ↓ $p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}})$

Figure 9.13 Illustration of the M step of the EM algorithm. The distribution $q(\mathbf{Z})$ is held fixed and the lower bound $\mathcal{L}(q, \boldsymbol{\theta})$ is maximized with respect to the parameter vector $\boldsymbol{\theta}$ to give a revised value $\boldsymbol{\theta}^{\text{new}}$. Because the KL divergence is nonnegative, this causes the log likelihood $\ln p(\mathbf{X}|\boldsymbol{\theta})$ to increase by at least as much as the lower bound does.



EM as alternating maximization

Figure 9.14 The EM algorithm involves alternately computing a lower bound on the log likelihood for the current parameter values and then maximizing this bound to obtain the new parameter values. See the text for a full discussion.



Lower bound L(q, $\boldsymbol{\theta}$) is a convex function having a unique maximum (for mixture components from the exponential family).

Extensions: Generalised EM seeks to improve rather than maximise L(q, $\boldsymbol{\theta}$); expectation conditional maximisation seeks to maximise L(q, $\boldsymbol{\theta}$) for a subset of the parameters; Incremental algorithms also exist.

Outline

EM revisited

- Connections between GMM and K-means
- Bernoulli mixture

EM in general - does it really maximise likelihood, and why?

Practical considerations and other topics
- impossibility of clustering [Kleinberg 2003]

- Kmeans++ [Vassilvitskii and Arthur, 2006]

An Impossibility Theorem for Clustering

[NeuRIPS 2003]

Jon Kleinberg

Professor of Computer Science, Cornell Univer
Verified email at cs.cornell.edu - Homepage
algorithms data mining information network

Theorem 2.1 *For each $n \geq 2$, there is no clustering function f that satisfies Scale-Invariance, Richness, and Consistency.*

Single-linkage operates by initializing each point as its own cluster, and then repeatedly merging the pair of clusters whose distance to one another (as measured from their closest points of approach) is minimum.
Stopping conditions:

	K-clusters	Distance-r	Scale-α	K-means
SCALE-INVARIANCE. *For any distance function d and any $\alpha > 0$, we have $f(d) = f(\alpha \cdot d)$.*	✓		✓	
RICHNESS. Range(f) is equal to the set of all partitions of S.		✓	✓	
CONSISTENCY. Let d and d' be two distance functions. If $f(d) = \Gamma$, and d' is a Γ-transformation of d, then $f(d') = \Gamma$.	✓	✓		✗

K-means ++

Vassilvitskii, Sergei, and David Arthur.
"k-means++: The advantages of careful seeding."
In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027-1035. 2006.

Sergei Vassilvitskii

I am a Research Scientist at Google New York. Previously I was a Research Scientist at Yahoo! Research and an Adjunct Assistant Professor at Columbia University. I completed my PhD at Stanford Universty under the supervision of Rajeev Motwani. Prior to that I was an undergraduate at Cornell University.

sergei at cs.stanford.edu

Problem with K-means

Algorithm summary:

1. Choose one center uniformly at random among the data points.
2. For each data point x not chosen yet, compute $D(x)$, the distance between x and the nearest center that has already been chosen.
3. Choose one new data point at random as a new center, using a weighted probability distribution where a point x is chosen with probability proportional to $D(x)^2$.
4. Repeat Steps 2 and 3 until k centers have been chosen.
5. Now that the initial centers have been chosen, proceed using standard *k*-means clustering.

Summary for today

EM revisited

- Connections between GMM and K-means
- Bernoulli mixture

EM in general - does it really maximise likelihood, and why?

Practical considerations and other topics
- impossibility of clustering [Kleinberg 2003]

- Kmeans++ [Vassilvitskii and Arthur, 2006]

Assignment 1 “post mortem”

Gradescope logistics:

- Rubric list is available - feel free to look at what are the rest of the de-scoring points
- Do remember to link submission pages to questions

A few notable errors

- Theory Q4.2 - derivation error propagates through to rest of the question
- PCA - center data or not?
- Programming Q2 - use pytorch to find weights but did not show workings
- No grades from automarker if there is a python error - more access to auto-grader in assignment 2