# CS5487 Problem Set 4
## Mixture models and the EM algorithm

Antoni Chan
Department of Computer Science
City University of Hong Kong

## Mixture models

**Problem 4.1 Mean and variance of a mixture model**

Consider a d-dimensional vector r.v. $x$ with a mixture distribution given by

$$p(x) = \sum_{j=1}^{K} \pi_j p(x|j), \tag{4.1}$$

where the mean and covariance of each component $p(x|j)$ is $\mu_j$ and $\Sigma_j$. Show that the mean and covariance of $x$ is

$$\mathbb{E}[x] = \sum_{j=1}^{K} \pi_j \mu_j, \qquad \text{cov}(x) = \sum_{j=1}^{K} \pi_j (\Sigma_j + \mu_j \mu_j^T) - \mathbb{E}[x]\mathbb{E}[x]^T. \tag{4.2}$$

.....

**Problem 4.2 Direct derivation of MLE of a GMM**

In this problem we will directly derive the maximum likelihood estimate for a Gaussian mixture model. Let r.v. $x$ be distributed according to a GMM,

$$p(x|\theta) = \sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j, \sigma_j^2), \tag{4.3}$$

where $\theta = \{\pi_j, \mu_j, \sigma_j^2\}_{j=1}^K$ are the parameters. Given a set of i.i.d. samples, $\mathcal{D} = \{x_1, \cdots, x_n\}$, the maximum likelihood estimate is given by

$$\hat{\theta} = \underset{\theta}{\text{argmax}}\, \ell(\theta), \qquad \ell(\theta) = \sum_{i=1}^{n} \log p(x_i|\theta). \tag{4.4}$$

(a) Derive the ML estimate for the mean $\mu_j$ by taking the partial derivative of $\ell(\theta)$ w.r.t. $\mu_j$, setting it to zero, and solving for $\mu_j$. Repeat for the variance $\sigma_j^2$ and priors $\pi_j$. For the latter, use Lagrange multipliers to enforce the constraint $\sum_j \pi_j = 1$ (see Problem 4.12).

(b) Verify that your estimate in (a) is not a closed-form solution (i.e., each estimate depends on itself). Suggest an iterative scheme for finding a solution. How is it related to the EM algorithm for GMMs?

.....

**Problem 4.3 EM for multivariate GMMs**

Let $x$ be a d-dimensional vector r.v. distributed as a multivariate Gaussian mixture model,

$$p(x|\theta) = \sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j), \tag{4.5}$$

where $\theta = \{\pi_j, \mu_j, \Sigma_j\}_{j=1}^K$ are the parameters. Let $X = \{x_1, \cdots, x_n\}$ be a set of observed samples, and $Z = \{z_1, \cdots, z_n\}$ the set of corresponding hidden values.

(a) Write down the complete data log-likelihood, $\log p(X, Z|\theta)$.

(b) E-step: derive the $Q$ function, $Q(\theta; \hat{\theta}^{\text{old}})$, and show that the E-step consists of calculating the soft assignments,

$$\hat{z}_{ij} = p(z_i = j|x_i, \hat{\theta}^{\text{old}}) = \frac{\pi_j \mathcal{N}(x_i|\hat{\mu}_j, \hat{\Sigma}_j)}{\sum_{k=1}^K \pi_k \mathcal{N}(x_i|\hat{\mu}_k, \hat{\Sigma}_k)}. \tag{4.6}$$

(c) M-step: show that the parameters that maximize the $Q$ function are

$$\hat{\mu}_j = \frac{1}{\hat{N}_j} \sum_{i=1}^n \hat{z}_{ij} x_i, \quad \hat{\Sigma}_j = \frac{1}{\hat{N}_j} \sum_{i=1}^n \hat{z}_{ij}(x_i - \hat{\mu}_j)(x_i - \hat{\mu}_j)^T, \quad \hat{\pi}_j = \frac{\hat{N}_j}{n}, \quad \hat{N}_j = \sum_{i=1}^n \hat{z}_{ij}. \tag{4.7}$$

Hint: use Lagrange multipliers (Problem 4.12) to find $\hat{\pi}_j$. The derivatives in Problem 2.6 will be helpful.

. . . . . .

**Problem 4.4 EM for multivariate GMMs with shared covariance**

Repeat Problem 4.3 for the special case in which the covariance matrices $\Sigma_j$ of all components are constrained to have a common value of $\Sigma$, i.e.,

$$\Sigma_1 = \Sigma_2 = \ldots = \Sigma_K = \Sigma. \tag{4.8}$$

. . . . . .

**Problem 4.5 Flying Bombs, part II – EM for mixtures of Poissons**

Let's reconsider the Problem 2.1, where we fit a Poisson distribution to the numbers of flying bombs hitting different areas in London. If we assume that the Germans were indeed targeting specific areas, then the bomb hit rate $\lambda$ would be higher for some squares (the targets), and lower for others (not the targets). Hence, the distribution over all squares should be a mixture of Poissons, with each Poisson component corresponding to squares with a particular hit rate. For $K = 2$, the components would correspond to target squares and non-target squares. For $K > 2$, one component would correspond to the target hit rate, while the other (non-target) components would have some gradation of hit rates (with squares far away from the target squares having lower hit rates).

(a) Consider the mixture of Poisson distribution

$$p(x = k|\theta) = \sum_{j=1}^{K} \pi_j \frac{1}{k!} e^{-\lambda_j} \lambda_j^k, \tag{4.9}$$

where $\lambda_j$ is the rate parameter for component $j$, and $\theta = \{\lambda_j, \pi_j\}_{j=1}^K$ the parameters of the mixture. Derive the EM algorithm to estimate the parameters of the model given samples $X = \{x_1, \cdots, x_n\}$. How is the M-step related to the ML estimate for a Poisson (Problem 2.1)?

(b) Implement your algorithm and run it for different values of $K \in \{1, 2, 3, 4, 5\}$ on the following data obtained from 2 cities (learn a separate mixture model for each city):

| city | number of hits ($k$) | 0 | 1 | 2 | 3 | 4 | 5 and over |
|---|---|---|---|---|---|---|---|
| London | number of cells with $k$ hits | 229 | 211 | 93 | 35 | 7 | 1 |
| Antwerp | number of cells with $k$ hits | 325 | 115 | 67 | 30 | 18 | 21 |

What conclusions can you make about the attacks on each city? Is there any evidence to suggest there is specific targeting of areas in London or Antwerp?

.....

**Problem 4.6 Mixture of exponentials**

Consider a mixture of exponential densities,

$$p(x) = \sum_{j=1}^{K} \pi_j p(x|j), \qquad p(x|j) = \lambda_j e^{-\lambda_j x}. \tag{4.10}$$

where $\lambda_j > 0$ is the parameter of component $j$. Given a set of samples $\{x_1, \cdots, x_n\}$, derive the EM algorithm to estimate the parameters $\theta = \{\pi_j, \lambda_j\}_{j=1}^K$.

.....

## EM for other things

Besides mixture models, EM can also be used for learning when there is other missing information (e.g., missing feature values).

**Problem 4.7 EM and missing features**

In this problem, we will investigate using EM to learn a multivariate distribution from training data when some of the features are missing. Suppose we have a two r.v.'s, $x$ and $y$, that are jointly Gaussian,

$$p(x, y) = \mathcal{N}(\begin{bmatrix} x \\ y \end{bmatrix} |\mu, \Sigma), \tag{4.11}$$

where $\mu = [\mu_x, \mu_y]^T$ is the mean and $\Sigma$ the covariance. We have collected a set of samples $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$ to learn the parameters of the model. However, upon examining the samples, we find

that the sample $y_1$ is corrupted, while $x_1$ is still okay. Rather than throw away both $x_1$ and $y_1$, we can to treat the data point $y_1$ as "missing information" and then use EM to find the model parameters. Denote $Y_I = \{y_2, \cdots, y_n\}$ as the incomplete data of $Y$. Let's consider the case where the covariance is diagonal, $\Sigma = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix}$, i.e., $x$ and $y$ are independent univariate Gaussians.

(a) Write down the complete data log-likelihood, $p(X, Y)$.

(b) E-step: Show that the $Q$ function can be written as (up to an additive constant)

$$Q(\theta; \hat{\theta}^{\text{old}}) = \mathbb{E}_{y_1|X,Y_I}[\log p(X, Y|\theta)] \tag{4.12}$$

$$= -\frac{1}{2\sigma_x^2}\sum_{i=1}^{n}(x_i - \mu_x)^2 - \frac{1}{2\sigma_y^2}\sum_{i=2}^{n}(y_i - \mu_y)^2 - \frac{\hat{\sigma}_y^2 + (\hat{\mu}_y - \mu_y)^2}{2\sigma_y^2} - \frac{n}{2}\log\sigma_x^2 - \frac{n}{2}\log\sigma_y^2. \tag{4.13}$$

where $\{\hat{\mu}_y, \hat{\sigma}_y^2\}$ are the old parameters ($\hat{\theta}^{\text{old}}$).

(c) M-step: Show that the maximum of the $Q$ function in (4.13) is obtained with

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2, \tag{4.14}$$

$$\mu_y = \frac{1}{n}\left(\hat{\mu}_y + \sum_{i=2}^{n} y_i\right), \qquad \sigma_y^2 = \frac{1}{n}\left(\hat{\sigma}_y^2 + (\hat{\mu}_y - \mu_y)^2 + \sum_{i=2}^{n}(y_i - \mu_y)^2\right). \tag{4.15}$$

(d) Determine the fixed-point of this EM algorithm (i.e., the solution at convergence), by setting $\mu_y = \hat{\mu}_y$ and solving for $\mu_y$ (and similarly for $\sigma_y^2$). What does the fixed-point solution correspond to?

Now consider the case where $\Sigma$ is an arbitrary covariance matrix. For convenience, define the vector $z_i = \begin{bmatrix} x_i & y_i \end{bmatrix}^T$.

(e) Write down the complete data log-likelihood, $p(X, Y)$, using matrix notation.

(f) E-step: Show that the $Q$ function can be written as

$$Q(\theta; \hat{\theta}^{\text{old}}) = \mathbb{E}_{y_1|X,Y_I,\hat{\theta}^{\text{old}}}[\log p(X, Y)] \tag{4.16}$$

$$= -\frac{1}{2}\sum_{i=2}^{n}(z_i - \mu)^T\Sigma^{-1}(z_i - \mu) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\text{tr}\left(\Sigma^{-1}\mathbb{E}_{y_1|x_1,\hat{\theta}^{\text{old}}}\left[(z_1 - \mu)(z_1 - \mu)^T\right]\right). \tag{4.17}$$

(g) Show that the expectation in (4.17) is given by

$$\mathbb{E}_{y_1|x_1,\hat{\theta}^{\text{old}}}\left[(z_1 - \mu)(z_1 - \mu)^T\right] = (\hat{z}_1 - \mu)(\hat{z}_1 - \mu)^T + \hat{\Sigma}_1, \tag{4.18}$$

where

$$\hat{z}_1 = \begin{bmatrix} x_1 \\ \hat{\mu}_{y_1} \end{bmatrix}, \qquad \hat{\Sigma}_1 = \begin{bmatrix} 0 & 0 \\ 0 & \hat{\sigma}_{y_1}^2 \end{bmatrix} \tag{4.19}$$

and

$$\hat{\mu}_{y_1} = \mathbb{E}_{y_1|x_1,\hat{\theta}^{\text{old}}}[y_1], \qquad \hat{\sigma}_{y_1}^2 = \text{var}_{y_1|x_1,\hat{\theta}^{\text{old}}}(y_1). \tag{4.20}$$

In other words, the missing data point $y_1$ is replaced with the mean of $y_1$ conditioned on $x_1$, with some extra variance added due to uncertainty when recovering $y_1$ from $x_1$. (Hint: write out the terms of the outer-product, and then take the expectation.)

(h) M-step: Show that the maximum of the $Q$ function in (4.17) is

$$\mu = \frac{1}{n}\left(\hat{z}_1 + \sum_{i=2}^{n} z_i\right), \quad \Sigma = \frac{1}{n}\left[\hat{\Sigma}_1 + (\hat{z}_1 - \mu)(\hat{z}_1 - \mu)^T + \sum_{i=2}^{n}(z_i - \mu)(z_i - \mu)^T\right]. \tag{4.21}$$

Given the result in (c), how is this solution similar or different than what you expected?

.....

**Problem 4.8 Multinomial EM**

In this problem we consider an example where there is a closed-form solution to ML estimation from incomplete data. The goal is to compare with the EM solution and get some insight on how the steps of the latter can be substantially easier to derive than the former.

Consider our bridge example and let $U$ be the type of vehicle that crosses the bridge. $U$ can take 4 values, (*compact, sedan, station wagon, and pick-up truck*) which we denote by $U \in \{1, 2, 3, 4\}$. On a given day, an operator collects an iid sample of size $n$ from $U$ and the number of vehicles of each type is counted and stored in a vector $\mathcal{D} = [x_1, x_2, x_3, x_4]$. The resulting random variable $X$ (the histogram of vehicle classes) has a multinomial distribution

$$p(x_1, x_2, x_3, x_4|\theta) = \frac{n!}{x_1!x_2!x_3!x_4!}\left(\frac{1}{2} + \frac{1}{4}\theta\right)^{x_1}\left(\frac{1}{4} - \frac{1}{4}\theta\right)^{x_2}\left(\frac{1}{4} - \frac{1}{4}\theta\right)^{x_3}\left(\frac{1}{4}\theta\right)^{x_4}, \tag{4.22}$$

where $\theta$ is the parameter and $0 \leq \theta \leq 1$.

It is later realized that the operator included *motorcycles* in the *compact class*. Denote $x_{11}$ as the number of compact cars, and $x_{12}$ as the number of motorcycles. It is established that motorcycles have probability $\frac{1}{4}\theta$, which leads to a new model

$$p(x_{11}, x_{12}, x_2, x_3, x_4|\theta) = \frac{n!}{x_{11}!x_{12}!x_2!x_3!x_4!}\left(\frac{1}{2}\right)^{x_{11}}\left(\frac{1}{4}\theta\right)^{x_{12}}\left(\frac{1}{4} - \frac{1}{4}\theta\right)^{x_2}\left(\frac{1}{4} - \frac{1}{4}\theta\right)^{x_3}\left(\frac{1}{4}\theta\right)^{x_4}, \tag{4.23}$$

Determining the parameter $\theta$ from the available data is a problem of ML estimation with *missing data*, since we only have measurements for

$$x_1 = x_{11} + x_{12} \tag{4.24}$$

but not for $x_{11}$ and $x_{12}$ independently.

(a) Determine the value of $\theta$ that maximizes the likelihood of $\mathcal{D}$, i.e.

$$\theta_i^* = \underset{\theta}{\operatorname{argmax}}\, p(\mathcal{D}|\theta) \tag{4.25}$$

by using standard ML estimation procedures.

(b) Assume that we have the complete data, i.e. $\mathcal{D}_c = [x_{11}, x_{12}, x_2, x_3, x_4]$. Determine the value of $\theta$ that maximizes its likelihood, i.e.

$$\theta_c^* = \underset{\theta}{\operatorname{argmax}}\, p(\mathcal{D}_c; \theta), \tag{4.26}$$

by using standard ML estimation procedures. Compare the difficulty of obtaining this solution vs. that of obtaining the solution in (a). Does this look like a problem where EM might be helpful?

(c) Derive the E and M-steps of the EM algorithm for this problem.

(d) Using the equations for the EM steps, determine the fixed point of the algorithm (i.e. the solution) by making

$$\theta^{(k+1)} = \theta^{(k)} \tag{4.27}$$

where $k$ is the iteration number. Compare to the solution obtained in (a).

.....

**Problem 4.9 EM for hyperparameter estimation**

In this problem we will use EM to estimate the hyperparameters of Bayesian linear regression (Problem 3.10). As before, we want to estimate a function

$$f(x, \theta) = \phi(x)^T \theta \tag{4.28}$$

where $\theta \in \mathbb{R}^D$ is the parameter vector, $\phi(x)$ is a feature transformation of input $x \in \mathbb{R}^d$. We observe a noisy version of the function $y = f(x, \theta) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Hence, the likelihood of the output given the input $x$ and parameter $\theta$ is

$$p(y|x, \theta, \sigma^2) = \mathcal{N}(y|\phi(x)^T\theta, \sigma^2), \tag{4.29}$$

As it is a Bayesian model, we place a Gaussian prior on the parameter vector,

$$p(\theta|\alpha) = \mathcal{N}(\theta|0, \alpha I). \tag{4.30}$$

For this model, $\psi = \{\sigma^2, \alpha\}$ are the hyperparameters. One way to estimate the hyperparameters is to maximize the marginal likelihood of the data (e.g., see Programming Assignment 1, Q3),

$$\hat{\psi} = \underset{\sigma^2,\alpha}{\operatorname{argmax}} \log p(y|x, \sigma^2, \alpha) = \underset{\sigma^2,\alpha}{\operatorname{argmax}} \log \int p(y|x, \theta, \sigma^2) p(\theta|\alpha) d\theta. \tag{4.31}$$

Note that this is a maximization problem of a likelihood that has "missing data", i.e., the parameter vector $\theta$. Hence, the EM algorithm can be employed to optimize it! Let the observations $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$ be the incomplete data, and $\theta$ be the missing data (note there is only one $\theta$ for all the observations).

(a) Write down the complete data log-likelihood, $p(\mathcal{D}, \theta|\psi)$.

(b) E-step: Show that the $Q$ function is

$$Q(\psi; \hat{\psi}^{\text{old}}) = \mathbb{E}_{\theta|\mathcal{D},\hat{\psi}^{\text{old}}}[\log p(\mathcal{D}, \theta|\psi)] \tag{4.32}$$

$$= -\frac{D}{2}\log\alpha - \frac{1}{2\alpha}\mathbb{E}_{\theta|\mathcal{D},\hat{\psi}^{\text{old}}}[\theta^T\theta] - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\mathbb{E}_{\theta|\mathcal{D},\hat{\psi}^{\text{old}}}\left[\left\|y - \Phi^T\theta\right\|^2\right] \tag{4.33}$$

where the conditional expectations are

$$\mathbb{E}_{\theta|\mathcal{D},\hat{\psi}^{\text{old}}}[\theta^T\theta] = \hat{\mu}_\theta^T\hat{\mu}_\theta + \operatorname{tr}(\hat{\Sigma}_\theta), \tag{4.34}$$

$$\mathbb{E}_{\theta|\mathcal{D},\hat{\psi}^{\text{old}}}\left[\left\|y - \Phi^T\theta\right\|^2\right] = \left\|y - \Phi^T\hat{\mu}_\theta\right\|^2 + \operatorname{tr}(\Phi^T\hat{\Sigma}_\theta\Phi), \tag{4.35}$$

and $\{\hat{\mu}_\theta, \hat{\Sigma}_\theta\}$ are defined in Problem 3.10, and calculated using $\hat{\psi}^{\text{old}}$.

(c) M-step: show that the maximum of the $Q$ function in (4.33) is

$$\alpha = \frac{1}{D}\left(\hat{\mu}_\theta^T \hat{\mu}_\theta + \mathrm{tr}(\hat{\Sigma}_\theta)\right), \tag{4.36}$$

$$\sigma^2 = \frac{1}{n}\left\{\left\|y - \Phi^T \hat{\mu}_\theta\right\|^2 + \mathrm{tr}(\Phi^T \hat{\Sigma}_\theta \Phi)\right\}. \tag{4.37}$$

What is the interpretation of these estimates? Does it make sense?

.....

## EM algorithm and MAP estimation

**Problem 4.10 EM algorithm for MAP estimation**

In this problem, we extend the EM algorithm to perform MAP estimation for models with hidden variables. Let $x$ be the observation r.v., $z$ the hidden r.v., and $\theta$ the parameters r.v. The joint likelihood of $\{x, z\}$ is

$$p(x, z|\theta) = p(x|z, \theta)p(z|\theta) \tag{4.38}$$

Let $X$ be the observed data, and $Z$ the corresponding hidden values. We wish to use the EM algorithm to find the maximum of the posterior distribution over parameters $p(\theta|X)$.

(a) *MAP Q function*: rather than take the expectation of the complete data log-likelihood, $\log p(X, Z|\theta)$, we take the expectation of the posterior $\log p(\theta|X, Z)$. Show that the MAP $Q$ function can be written as

$$Q_{MAP}(\theta; \hat{\theta}^{\text{old}}) = \mathbb{E}_{Z|X,\hat{\theta}^{\text{old}}}[\log p(\theta|X, Z)] \tag{4.39}$$

$$\propto \mathbb{E}_{Z|X,\hat{\theta}^{\text{old}}}[\log p(X, Z|\theta)] + \log p(\theta), \tag{4.40}$$

where we have dropped terms that do not affect the M-step later.

(b) *MAP-EM algorithm*: Using the result in (a), what are the E- and M-steps of the MAP-EM algorithm?

(c) *MAP-EM, alternative view*: Note that the first term in (4.40) is the only term with an expectation, and is also the same as the $Q$ function of EM. Show that the MAP-EM algorithm from (b) can be written as

$$\begin{aligned} \text{E} - \text{step}: \quad & Q(\theta; \hat{\theta}^{\text{old}}) = \mathbb{E}_{Z|X,\hat{\theta}^{\text{old}}}[\log p(X, Z|\theta)], \\ \text{M} - \text{step}: \quad & \hat{\theta}^{\text{new}} = \operatorname*{argmax}_\theta Q(\theta; \hat{\theta}^{\text{old}}) + \log p(\theta). \end{aligned} \tag{4.41}$$

Hence, the MAP-EM adds the log prior to the M-step of standard EM, and the M-step looks like a standard Bayesian estimation procedure.

(d) *Example:* Derive the MAP-EM algorithm for a univariate GMM with 2 components,

$$p(x) = \pi_1 \mathcal{N}(x|\mu_1, \sigma_1^2) + (1 - \pi_1)\mathcal{N}(x|\mu_2, \sigma_2^2), \tag{4.42}$$

where $\theta = \{\pi_1, \mu_1, \mu_2\}$ are the parameters and $\sigma_j^2$ are known. The prior distributions are

$$p(\pi_1) = 1, \quad 0 \leq \pi_1 \leq 1, \tag{4.43}$$
$$p(\mu_1) = \mathcal{N}(\mu_1|\mu_0, \sigma_0^2), \tag{4.44}$$
$$p(\mu_2) = \mathcal{N}(\mu_2|\mu_0, \sigma_0^2). \tag{4.45}$$

(Hint: Use your results from Problem 3.4 and Problem 3.7.) How are these E- and M-steps similar and different to the standard EM for GMMs? What is the interpretation in terms of "virtual" samples added to standard EM for GMMs?

.....

**Problem 4.11 MAP-EM for multinomial**

Consider the multinomial distribution of Problem 4.8, and a Gamma prior

$$p(\theta) = \frac{\Gamma(\nu_1 + \nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} \theta^{\nu_1 - 1}(1 - \theta)^{\nu_2 - 1}. \tag{4.46}$$

Derive the equations of the EM algorithm for MAP estimation of the parameter $\theta$.

.....

## Optimization Theory

**Problem 4.12 Lagrange multipliers and equality constraints**

In the M-step of the EM algorithm for mixture models, we need to calculate an estimate of the component prior probabilities $\pi_j$ via the optimization problem,

$$\{\hat{\pi}_j\} = \underset{\{\pi_j\}}{\operatorname{argmax}} \sum_{j=1}^{K} z_j \log \pi_j, \quad \text{s.t.} \sum_{j=1}^{K} \pi_j = 1, \quad \pi_j \geq 0, \tag{4.47}$$

for some $z_j \geq 0$. Note that this optimization problem has an equality constraint, which is $\{\pi_j\}$ must sum to 1, since they represent a probability distribution.

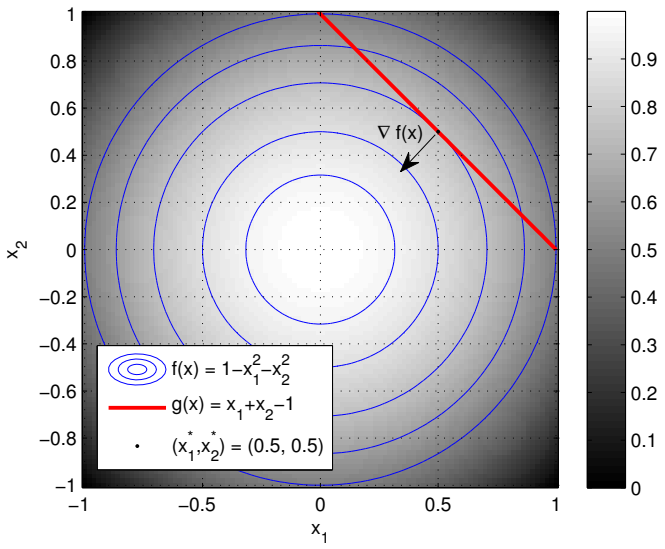
One method of solving an optimization problem with equality constraints is to use *Lagrange multipliers*. Consider the following problem,

$$\begin{aligned} x^* = &\underset{x}{\operatorname{argmax}}\, f(x), \\ &\text{s.t. } g(x) = 0, \end{aligned} \tag{4.48}$$

where $f(x)$ is the objective function and $g(x)$ is the constraint function. Let's look at two properties of these functions,

- First, the gradient $\nabla g(x)$ is orthogonal to the constraint surface, since $g(x)$ should be constant along the direction of the constraint surface (otherwise it would not be 0).
- Second, at the optimum, the gradient $\nabla f(x)$ must also be orthogonal to the constraint surface. Otherwise, we could move along the constraint surface to increase $f(x)$.

This is illustrated in the following figure[1]:




Hence, $\nabla f(x)$ and $\nabla g(x)$ must be parallel or anti-parallel, and by extension,

$$\nabla f(x) + \lambda \nabla g(x) = 0, \tag{4.49}$$

for some $\lambda \neq 0$. Define the *Lagrangian function*,

$$L(x, \lambda) = f(x) + \lambda g(x). \tag{4.50}$$

The optimality condition in (4.49) is obtained by setting $\frac{\partial L}{\partial x} = 0$. Furthermore, setting $\frac{\partial L}{\partial \lambda} = 0$ yields the equality constraint, $g(x) = 0$. Hence, to solve the constrained optimization problem (4.48), we form the Lagrangian function, and find the stationary point w.r.t. both $x$ and $\lambda$, by simultaneously solving

$$\frac{\partial}{\partial x} L(x, \lambda) = 0, \quad \frac{\partial}{\partial \lambda} L(x, \lambda) = 0. \tag{4.51}$$

(a) Use Lagrange multipliers to optimize (4.47), and show that the solution is

$$\pi_j = \frac{z_j}{\sum_{k=1}^{K} z_k}. \tag{4.52}$$

(b) Consider another optimization problem,

$$\{\hat{\pi}_j\} = \underset{\{\pi_j\}}{\operatorname{argmax}} \sum_{j=1}^{K} \pi_j (z_j - \log \pi_j), \quad \text{s.t.} \sum_{j=1}^{K} \pi_j = 1, \quad \pi_j \geq 0. \tag{4.53}$$

Show that the solution is $\pi_j = \frac{\exp z_j}{\sum_{k=1}^{K} \exp z_k}$.

More details about Lagrange multipliers can be found in Appendix E of Bishop's book, PRML.

. . . . . .

[1] The red line is straight! There is an optical illusion that makes it looked curved!

**Problem 4.13 The "log trick" and preventing numerical precision problems**

In this problem, we will see how to prevent numerical precision problems by doing calculations in the log-domain. When dealing with mixtures, we often need to calculate the likelihood of a sample under the full mixture distribution,

$$p(x) = \sum_{j=1}^{K} \pi_j p(x|j), \tag{4.54}$$

where $p(x|j)$ is the mixture component density and $\pi_j$ is the component prior probability. For example, $p(x)$ is needed for the denominator when calculating the soft assignment variables,

$$\hat{z}_{ij} = \frac{\pi_j p(x|j)}{\sum_k \pi_k p(x|k)} = \frac{\pi_j p(x|j)}{p(x)}. \tag{4.55}$$

However, when $x$ is high-dimensional (or the sample $x$ happens to be far away from all the components, or both), then the component likelihoods $p(x|j)$ can become very small. If the likelihood $p(x|j) < e^{-709}$, then the computer will interpret this as 0 due to the limited precision of the computer. If all the component likelihoods are this small, then $\hat{z}_{ij}$ cannot be calculated since the numerator and denominator will be both zero.

The solution to this problem is to do the calculations in the log domain. That is, use the *log-likelihood* of the components, $\ell(x|j) = \log p(x|j) + \log \pi_j$, to calculate the *log-likelihood* of the mixture $\ell(x) = \log p(x)$. We can then calculate the log probabilities of the assignment variables,

$$\hat{\gamma}_{ij} = \log \hat{z}_{ij} = \ell(x_i|j) - \ell(x_i), \tag{4.56}$$

which avoids the 0/0 problem, and take the exponent to get the assignment variable

$$\hat{z}_{ij} = e^{\hat{\gamma}_{ij}}. \tag{4.57}$$

Finally, we can use the data log-likelihood to test for convergence of the algorithm,

$$\log p(X) = \sum_{i=1}^{n} \log p(x_i) = \sum_{i=1}^{n} \ell(x_i). \tag{4.58}$$

The individual component log-likelihoods $\ell(x|j)$ can be calculated without a problem. However, when looking at $\ell(x)$,

$$\ell(x) = \log p(x) = \log \sum_{j=1}^{K} \pi_j p(x|j) = \log \sum_{j=1}^{K} e^{\ell(x|j)} \tag{4.59}$$

we still need to take the exponent before calculating the sum. If all $\ell(x|j)$ are small, then all $e^{\ell(x|j)}$ will be zero, the sum will be 0, and $\ell(x) = \log 0 = -\texttt{inf}$, resulting in the same precision problem.

We will now derive a way to calculate $\ell(x)$ in a numerically safe way, which is what I call the "log trick". Formally, given values $\{\ell_j\}_{j=1}^K$, we want to calculate the log-sum-exp of these values,

$$\ell = \log \sum_{j=1}^{K} e^{\ell_j}. \tag{4.60}$$

Let $\ell_* = \max(\ell_1, \cdots, \ell_K)$ be the maximum value over all the $\ell_j$.

(a) Show that $\ell$ in (4.60) can be rewritten as

$$\ell = \ell_* + \log \sum_{j=1}^{K} e^{(\ell_j - \ell_*)}. \tag{4.61}$$

(b) Why is the form in (4.61) more numerically stable than (4.60)? That is, how have we avoided taking the log of 0?

.....

**Problem 4.14 Calculating log determinant**

For very high-dimensional Gaussian densities, the likelihoods may be very close to zero (or even zero due to numerical precision). The numerical precision problems can be avoided by calculating the log-likelihood of a Gaussian, which requires the log-determinant of $\Sigma$. Consider the following two properties:

- The Choleksy decomposition of a positive definite matrix $\Sigma$ is

$$\Sigma = LL^T, \tag{4.62}$$

  where $L$ is a lower triangular matrix (a matrix where the upper triangle is zero, i.e. $L_{ij} = 0$ when $i < j$). A $3 \times 3$ lower triangular matrix looks like this $\begin{bmatrix} L_{11} & 0 & 0 \\ L_{21} & L_{22} & 0 \\ L_{31} & L_{32} & L_{33} \end{bmatrix}$.

- The determinant of a lower triangular matrix $L$ (or an upper triangular matrix) is the product of its diagonal elements

$$|L| = \prod_{i=1}^{n} L_{ii}. \tag{4.63}$$

Use these two properties to show that the log-determinant of $\Sigma$ is

$$\log |\Sigma| = 2 \sum_{i=1}^{n} \log L_{ii}, \tag{4.64}$$

where $\Sigma = LL^T$ is the Cholesky decomposition of $\Sigma$.

.....