Example—deriving the GMM M-step

Finding variances/covariances is similar.

Finding the mixing coefficients is also similar, except we also need to enforce that they sum to one.

(Here the equations for the *k*'s are coupled).

So we use Lagrange Multipliers.

Using Lagrange Multipliers

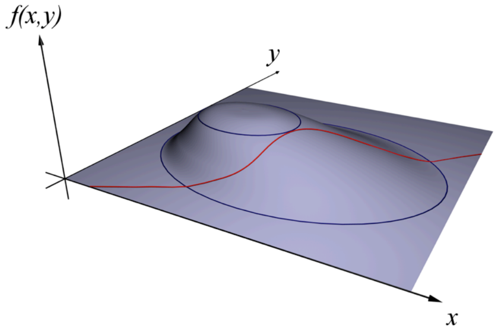
Now we find stationary points with respect to $\{\pi_k, \lambda\}$ of

$$Q\left(\theta^{(s+1)},\theta^{(s)}\right)+\lambda\left(\sum_k \pi_k - 1\right)$$

Note that differentiating with respect to λ, and setting the result to zero puts the constraint into the equations.

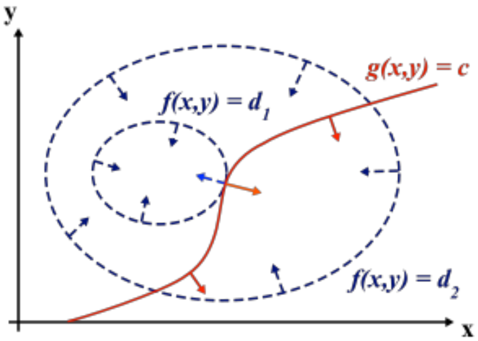
But the real problem is doing the optimization under the constraint.

Using Lagrange Multipliers



From WikiPedia

Using Lagrange Multipliers



From WikiPedia

Using Lagrange Multipliers

$\nabla f \parallel \nabla g$

$\nabla f = \lambda \nabla g$

So, $\nabla\left(f - \lambda g\right) = 0$

or, $\nabla\left(f + \lambda g\right) = 0$ (negate λ)

Using Lagrange Multipliers

Now we find stationary points with respect to $\{\pi_k, \lambda\}$ of

$$\mathrm{Q}\left(\theta^{(s+1)}, \theta^{(s)}\right) + \lambda\left(\sum_k \pi_k - 1\right)$$

$$\frac{\delta}{\delta\pi_k}\left\{\mathrm{Q}\left(\theta^{(s+1)}, \theta^{(s)}\right) + \lambda\left(\sum_k \pi_k - 1\right)\right\}$$

$$= \sum_n \left\{\gamma\left(z_{n,k}\right) \frac{\delta}{\delta\pi_k}\left(\log\left(\pi_k\right) + \log\left(\mathrm{N}\left(x_n \middle| \mu_k, \Sigma_k\right)\right)\right)\right\} + \lambda$$

$$= \sum_n \left\{\gamma\left(z_{n,k}\right) \frac{1}{\pi_k}\right\} + \lambda$$

Setting the result to zero, $\sum_n \left\{\gamma\left(z_{n,k}\right) \frac{1}{\pi_k}\right\} + \lambda = 0$

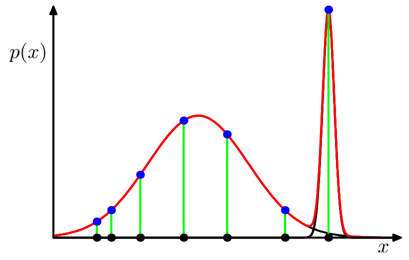
So $\quad \pi_k = \dfrac{\sum_n \left\{\gamma\left(z_{n,k}\right)\right\}}{-\lambda}$

Summing over k gives, $\quad 1 = \dfrac{\sum_k \sum_n \left\{\gamma\left(z_{n,k}\right)\right\}}{-\lambda} = \dfrac{N}{-\lambda}$

So, $\lambda = -N$, and $\pi_k = \dfrac{\sum_n \left\{\gamma\left(z_{n,k}\right)\right\}}{N}$ as before.

EM in practice

- For GMM we need to consider clusters that have essentially one point:



- Easily fixed by adding a constant to the variance (prior).

EM in practice

- Tying parameters (using GMM as an example)
  - Depending on the problem, it may make sense to assume that the variances (or covariances) for all clusters are the same by reducing the number of parameters.
    - This reduces the number of parameters, reducing the risk of over-fitting
  - Updates work as you expect. Instead of multiple weighted sums, you just use one big one.
  - In general, you would **not** tie variances over dimensions unless you know that the variables are semantically equivalent
    - Recall that one advantage of GMM over K-means is that the scale differences among dimensions is naturally taken care through the variance parameters

EM in practice

- You must check that the log likelihood increases!
- A simple way to compute it during an iteration:

Recall our objective function:

$$p(X)=\prod_{n}\sum_{k}p(k)p(x_n|k)$$

Consider how we might compute the responsibilities
$\gamma(n,k)\propto p(k)p(x_n|k)$
(Then normalize once you have them all).

So, make a running sum of the unormalized values

EM in practice

- Precision problems --> must work with logs
- But we need to exponentiate to normalize --> rescaling tricks

Let P$=\{p_i\}$.

Suppose we want Q$=\frac{1}{\sum_i p_i}\{p_i\}$

Where we need to use $V=\{\log(p_i)\}$

and $\exp(p_i)$ is too small, and the sum of them might be zero.

Let M = $\max\{\log(p_i)\}$

Observe that working with V′= $\{\log(p_i)-M\}$ does the trick.