Convergence is not gauranteed, however, if $g''(x)$ is bounded and does not change sign on $[a,b]$, then we can rescale nonconvergent problems by choosing $\alpha \neq 0$ and letting

$$G(x) = \alpha g'(x) + x.$$

This works since $\alpha g'(x) = 0$ iff $g'(x) = 0$.

There are ways to carefully calculate α, however it is often easier to just try a few values. (Ex pg 26).

Sec 2.2 Multivariate Problems.

In a multivariate opt. problem we seek to max/min a real valued function g of a p-dim vector $x = (x_1, \ldots, x_p)^T$. At iteration t, $x^{(t)} = (x_1^{(t)}, \ldots, x_p^{(t)})^T$

Many of the general principles still apply.
- iterative algorithms
- often take steps based on linearization of g' from Taylor series, secant approx, etc.
- convergence criteria are in the same spirit

Convergence criteria

L_p norms

Need $D(u,v)$, a distance measure for p-dim vectors.
Ex/ $D(u,v) = \sum_{i=1}^{p} |u_i - v_i|$ or $D(u,v) = \sqrt{\sum_{i=1}^{p} (u_i - v_i)^2}$
Then we form ab, rel convergence from:

$$D(x^{(t+1)}, x^{(t)}) < \varepsilon \quad \text{or} \quad \frac{D(x^{(t+1)}, x^{(t)})}{D(x^{(t)}, 0)} < \varepsilon$$

Sec. 2.2.1 Newton's Method & Fisher Scoring.

For Newton's method we approx $g(x^*)$ by the Taylor series

$$g(x^*) = g(x^{(t)}) + (x^* - x^{(t)})^T \vec{g}'(x^{(t)}) + (x^* - x^{(t)})^T \vec{g}''(x^{(t)})(x^* - x^{(t)})/2$$

and we max this function by setting the gradient of the r.h.s. equal to zero.

(Recall that the gradient of f at x is $f'(x) = \left(\frac{df(x)}{dx_1}, \dots \frac{df(x)}{dx_p}\right)$

& we have

$$0 = g'(x^{(t)}) + g''(x^{(t)})(x^* - x^{(t)})$$

and the algorithm becomes

$$x^{(t+1)} = x^{(t)} - g''(x^{(t)})^{-1} g'(x^{(t)}) \qquad \text{Newton's}$$

or similarly for an MLE we have

$$\theta^{(t+1)} = \theta^{(t)} + I(\theta^{(t)})^{-1} \ell'(\theta^{(t)}) \qquad \text{Fisher Scoring.}$$

Again the Methods are asymptotically equivalent (Ex. pg 33) and have similar problems to the univariate case.

Sec 2.2.2. Newton-like methods.

Computation of the Hessian, $g''(x^{(t)}) = \left[\frac{d^2 f(x)}{dx_i dx_j}\right]_{ij}$ can be expensive & so many methods rely upon eq. of the form

$$x^{(t+1)} = x^{(t)} - (M^{(t)})^{-1} g'(x^{(t)})$$

where $M^{(t)}$ is a $p \times p$ approx. (Ex Fisher Scoring).

2.2.2.1 Ascent Algorithms

With Newton's method the steps are not necess. uphill, i.e. $g(x^{(t+1)}) > g(x^{(t)})$. If we force this, than it is called an ascent algorithm.

what does this remind you of?

Method of steepest ascent: w/ $M^{(t)} = -I$.
$x^{(t+1)} = x^{(t)} + g'(x^{(t)})$ is taking a step in the steepest direction uphill (indicated by the gradient).

We can also use scaled steps
$$x^{(t+1)} = x^{(t)} + \alpha^{(t)} g'(x^{(t)}) \quad \text{for } \alpha^{(t)} > 0$$
to control convergence. or more generally
$$x^{(t+1)} = x^{(t)} - \alpha^{(t)} (M^{(t)})^{-1} g'(x^{(t)}).$$

Often w/ scaling $\alpha^{(t)} > 0$ is chosen to be a contraction or step length parameter whose value can shrink to ensure ascent at each step.
(See ex 2.7 pg 33).
In other words if a step turns out to be downhill, we adjust to ensure uphill.

Ex/ Backtracking:
- Start each step w/ $\alpha^{(t)} = 1$.
- if step is downhill ($g(x^{(t+1)}) < g(x^{(t)})$)
  let $\alpha^{(t)} = \frac{1}{2}\alpha^{(t)}$ & try again
- Repeat until step is uphill.

(See example pg 28)

Backtracking Will converge under formal conditions, can be slow.

pg39

Fixed-Point Methods.

If $M^{(t)} = M \ \forall t$ we have the fixed pt. method

$$x^{(t+1)} = x^{(t)} - M^{-1} g'(x^{(t)})$$

A reasonable choice is $M = g''(x^{(0)})$.
If M is diagonal, than this is eq. to applying the univariate-scaled fixed-point algorithm to each component.

Secant-Like Methods.

We can replace $g''(x^{(t)})$ w/ a matrix $M^{(t)}$ of finite discrete difference quotients.

Ex/ Let
$g'_i(x) = dg(x)/dx_i$ (ith element of $g'(x)$)
and
$$e_j = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow j\text{th position}$$

Then
$$m_{ij}^{(t)} = \frac{g'_i(x^{(t)} + h_{ij}^{(t)} e_j) - g'_i(x^{(t)})}{h_{ij}^{(t)}}$$

for some constants $h_{ij}^{(t)}$.
If $h_{ij}^{(t)} = h$ we get convergence order 1.
If $h_{ij}^{(t)} = x_j^{(t)} - x_j^{(t-1)} \ \forall i$ we get convergence order similar to secant method in univariate case.

There are quite a few other methods discussed in the text that we do not have time to cover.

pg 89

Chapter 4 EM (Expectation Maximization) Method

The EM algorithm is an iterative opt. strategy motivated by a notion of missingness & by consideration of the cond. dist of what is missing given what is observed.

Assume: we have observed data from r.v. X along with missining (unobserved or latent) data from r.v. Z. ∴ We wish to envision complete data from $Y = (X, Z)$.

$L(\theta|x)$ may be hard.

Given observed data x we want to maximize a likelihood $L(\theta|x)$, but we want to do so w/o calc $L(\theta|X)$ directly but rather working with $L(\theta|Y)$, $Y|\theta$ & $Z|(X,\theta)$. which may be more tractable.

Intuitively EM.
① Fills in z based on X & θ
② Restimates θ based on $Y=(X,Z)$

Notation.

X observed data
Y complete data
Z missing data
$f_X(x|\theta)$ density of observed data
$f_Y(y|\theta)$ density of complete data
M be the many to fewer mapping $X = M(Y)$

Then the missing data amounts to a marginalization model in which we observe X having density

$$f_X(x|\theta) = \int_{\{y:\, M(y)=x\}} f_Y(y|\theta)$$

and the conditional density of the missing z given observed x is

$$f_{Z|X}(z|x,\theta) = \frac{f_Y(y|\theta)}{f_X(x|\theta)}$$

And similarly we will view our likelihood $L(\theta|x)$ as a marginalization of the complete data Likelihood $L(\theta|y) = L(\theta|x,z)$.

Sec 4.2. The EM Algorithm

Let $\theta^{(t)}$ be our est at iteration $t = 0, 1, \ldots$.
Define $Q(\theta|\theta^{(t)})$ to be the expectation for the joint log likelihood of Y conditioned on $X = x$

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= E\left[\log L(\theta|Y) \,\middle|\, x, \theta^{(t)}\right] \\ &= E\left[\log f_Y(y|\theta) \,\middle|\, x, \theta^{(t)}\right] \\ &= \int \left(\log f_Y(y|\theta)\right] f_{Z|X}(z|x,\theta^{(t)})\,dz \end{aligned}$$

Recall Z is the only random part of Y once $X = x$.

Then the EM algorithm is (starting w/ $\theta^{(0)}$)
1. E step: Compute $Q(\theta|\theta^{(t)})$
2. M step: Maximize $Q(\theta|\theta^{(t)})$ w.r.t. θ and set $\theta^{(t+1)}$ equal to this maximizer.
3. Return to E step unless stopping criteria has been met.

Step 2 → use earlier methods.
Stopping criteria similar to before & built upon $|Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)})|$ or $(\theta^{(t+1)} - \theta^{(t)})^T(\theta^{(t+1)} - \theta^{(t)})$.

Ex/4.1 Very Simple

$Y_1, Y_2 \sim$ i.i.d $\text{Exp}(\theta)$ so $f(y_i) = \theta e^{-\theta y_i}$ $E[y_i] = \frac{1}{\theta}$

Suppose $y_1 = 5$ is observed (x) and y_2 is missing (z) & $Y = (Y_1, Y_2)$.

- write down expression for complete density
  $f_Y(Y|\theta) = \theta e^{-\theta y_1}\, \theta e^{-\theta y_2}$

also useful.
$f_{Z|X}(z|x, \theta^{(t)}) = f_Z(z|\theta^{(t)}) = \theta e^{-\theta y_2}$

- write down the log likelihood function of the complete Y
  $\log L(\theta|y) = \log f_Y(y|\theta) = 2\log\theta - \theta y_1 - \theta y_2$
- Find $Q(\theta|\theta^{(t)}) = E[\log L(\theta|y) \mid y_1, \theta^{(t)}]$

$$= E[2\log\theta - \theta y_1 - \theta y_2 \mid y_1, \theta^{(t)}]$$

$$= 2\log\theta - \theta \cdot 5 - \theta E[Y_2 \mid \theta^{(t)}].$$

$$= 2\log\theta - 5\theta - \theta/\theta^{(t)}.$$

Now we need to maximize $Q(\theta|\theta^{(t)})$ by solving for the root of $Q'(\theta|\theta^{(t)})$
$= 2/\theta - 5 - 1/\theta^{(t)} = 0.$

Hence $\theta^{(t+1)} = \dfrac{2(\theta^{(t)})}{5(\theta^{(t)}+1)}$.

and repeat.

Further comments:

① One of the most appealing & central results is that the sequence $\{\theta^{(k)}\}$ converges, at least to a local ~~minimum~~ maximum. proof pg 95. and for well behaved problems θ is a global max.

- We move uphill at each step. – The increase in the observed data log likelihood function $l(\theta|x)$ at each step is one of its most attractive features v/s Newton's Method.

- Rate of convergence: The rate is only linear (Newton quadratic) & is criticized as being slow. Convergence rate linked to proportion of data missing – more missing ⇒ slower convergence

However there are many techniques to speed up EM. (Sec 4.3).

- Starting Points: A main drawback is that its limiting position is often sensitive to initial guesses.

Pg 91 Ex 4.2 More complicated example.

C carbonia CC CI CT
I insularia II IT
T typica TT

$$p_C, p_I, p_T \Rightarrow \underset{CC}{p_C^2} \quad \underset{CI}{2p_Cp_I} \quad \underset{CT}{2p_Cp_T} \quad \underset{II}{p_I^2} \quad \underset{IT}{2p_Ip_T} \quad \underset{TT}{p_T^2}$$
$$\quad C \quad I \quad T$$

Also $p_C + p_I + p_T = 1$

$n = n_C + n_I + n_T$ $\quad X$ observed $= (n_C, n_I, n_T)$

Y = complete data $= (n_{CC}, n_{CI}, n_{CT}, n_{II}, n_{IT}, n_{TT})$

w/ $X = M(Y) = (n_{CC} + n_{CI} + n_{CT},\ n_{II} + n_{IT},\ n_{TT})$

Since $p_T = 1 - p_C - p_I$

$P = (p_C, p_I)$.

Multinomial R.V.

- complete log likelihood function.

$$\cancel{\log} f_Y(y|p) = \binom{n}{n_{CC} \cdots n_{TT}} (p_C^2)^{n_{CC}} \cdot (2p_Cp_I)^{n_{CI}} \cdot (2p_Cp_T)^{n_{CT}} \cdot (p_I^2)^{n_{II}} (2p_Ip_T)^{n_{IT}} (p_T^2)^{n_{TT}}$$

$\therefore$

$$\log f_Y(y|p) = n_{CC} \log p_C^2 + n_{CI} \log 2p_Cp_I + n_{CT} \log 2p_Cp_T + n_{II} \log p_I^2 + n_{IT} \log 2p_Ip_T + n_{TT} \log p_T^2 + \log \binom{n}{n_{CC} \cdots n_{TT}}$$

The complete data is

$$Y = (N_{CC}, N_{CI}, N_{CT}, N_{II}, N_{IT}, N_{TT})$$

and only $N_{TT} = n_{TT} = n_T$ is observed.

To find $Q(p|p^{(t)}) = E[\log f_Y(y|p) | x, p^{(t)}]$

we need

$E[N_{??} | n_C n_I, n_T, p^{(t)}]$ for each ?? pair

Each phenotype has its own dist.

Multinomial

$E[X_i] = np_i$

$$E[N_{CC} | x, p^{(t)}] = n_{CC}^{(t)} = \frac{n_C (p_C^{(t)})^2}{(p_C^{(t)})^2 + 2p_C^{(t)}p_I^{(t)} + 2p_C^{(t)}p_T^{(t)}}$$

$$E[N_{CI} | x, p^{(t)}] = n_{CI}^{(t)} = \frac{n_C (2p_C^{(t)}p_I^{(t)})}{(\quad)}$$

$$E[N_{CT} | x, p^{(t)}] = n_{CT}^{(t)} = \frac{n_C (2p_C^{(t)}p_T^{(t)})}{(\quad)}$$

$$E[N_{II} | X, p^{(t)}] = n_{II}^{(t)} = \frac{n_I (p_I^{(t)})^2}{(p_I^{(t)})^2 + 2p_I^{(t)} p_T^{(t)}}$$

$$E[N_{IT} | X, p^{(t)}] = n_{IT}^{(t)} = \frac{n_I 2(p_I^{(t)} p_T^{(t)})}{(\ \ )}$$

$$E[N_{TT} | X, p^{(t)}] = n_{TT} = n_T$$

all the exp are numbers //

and $E\left[\binom{n}{n_{CC} \cdots n_{PP}} \Big| x, p\right] = k(n_C, n_I, n_T; p_T)$

does not depend on p.

So

$$Q(p|p^{(t)}) = n_{CC}^{(t)} \log p_C^2 + n_{CI}^{(t)} \log 2p_C p_I + n_{CT}^{(t)} \log 2p_C p_T$$
$$+ n_{II}^{(t)} \log p_I^2 + n_{IT}^{(t)} \log 2 p_I p_T + n_{TT} \log p_T^{(2)}$$

and so we need to take partials w.r.t p_C & p_I

$$\frac{dQ(p|p^{(t)})}{dp_C} = \frac{2n_{CC}^{(t)} + n_{CI}^{(t)} + n_{CT}^{(t)}}{p_C} - \frac{n_{CT}^{(t)} + n_{IT}^{(t)} + 2n_{TT}^{(t)}}{1 - p_C - p_I}$$

$$\frac{dQ(p|p^{(t)})}{dp_I} = \frac{2n_{II}^{(t)} + n_{IT}^{(t)} + n_{CI}^{(t)}}{p_I} - \frac{2n_{TT}^{(t)} + n_{CT}^{(t)} + n_{IT}^{(t)}}{1 - p_C - p_I}.$$

We need to set these to 0 & solve.

$$p_C^{(t+1)} = \frac{2n_{CC}^{(t)} + n_{CI}^{(t)} + n_{CT}^{(t)}}{2n}$$

$$p_I^{(t+1)} = \frac{2n_{II}^{(t)} + n_{IT}^{(t)} + n_{CI}^{(t)}}{2n}$$

$$p_T^{(t+1)} = \frac{2n_{TT}^{(t)} + n_{CT}^{(t)} + n_{IT}^{(t)}}{2n}$$

Don't actually need missing info

Ex algorithm run on pg 94.

Final EM Example

Bayesian posterior mode w/ EM.

Review: Bayesian Inference Sec 1.5 pg 11

Consider a Bayesian problem with
- likelihood $L(\theta|x)$
- prior $f(\theta)$ - density assigned to θ before observing the data
- missing data or parameters Z

We wish to find the posterior density

$$f(\theta|x) = f(\theta) f(x|\theta) \cdot k = f(\theta) L(\theta|x) \cdot k$$

where k is a normalizing constant.

To use the EM method, the E-step requires

$$Q(\theta|\theta^{(t)}) = E\left[\log\{L(\theta|Y) f(\theta) \cdot k(Y) \,|\, x, \theta^{(t)}\}\right]$$

$$= E\left[\log L(\theta|Y) \,|\, x, \theta^{(t)}\right] + \log f(\theta)$$

$$+ E\left[\log k(Y) \,|\, x, \theta^{(t)}\right]$$

where the last term can be ignored when maximizing w.r.t. θ.

So, we have our standard MLE Q with the addition of the log prior.