Computational Aspects of MLE:

If the likelihood equations are non-linear fctns. of the parameters, then usually cannot get a "closed-form" solution to these first-order conditions. Use a numerical approximation to the solution. Lots of different techniques — illustrate with "Methods of Descent". (Max. Log L equivalent to min. $-\text{Log } L$.)

Exactly same problem encountered with non-linear least squares.

General approach: $\tilde{\underset{\sim}{\Theta}} = \underset{\sim}{\Theta}_0 + s \cdot \underset{\sim}{d}(\underset{\sim}{\Theta}_0)$

$\underset{\sim}{\Theta}_0$ = initial value

s = step-length (scalar, >0)

$\underset{\sim}{d}(\cdot)$ = direction vector

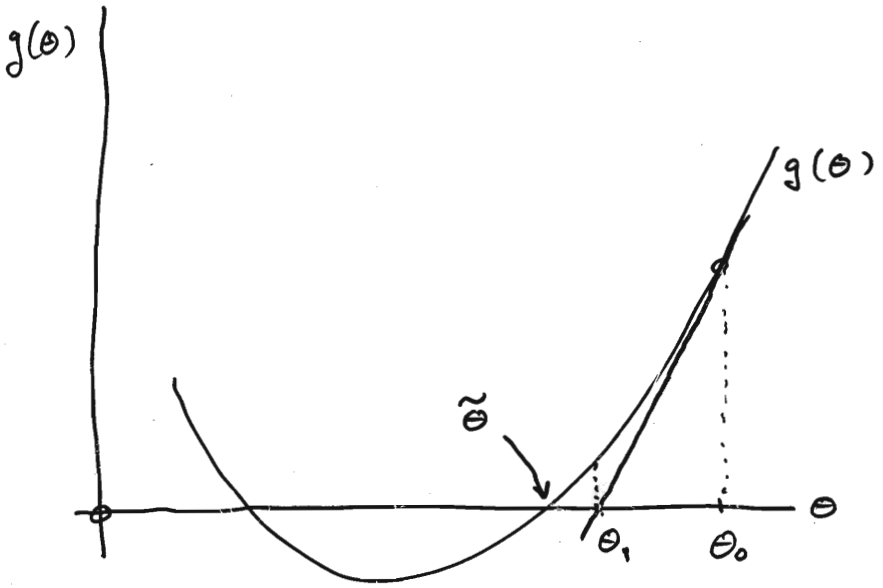
Usually $\underset{\sim}{d}(\cdot)$ depends on gradient vector (+ maybe Hessian matrix), at $\underset{\sim}{\Theta}_0$

Sometimes $s = s(\text{Hessian})$, too.

A specific member of family of descent methods is Newton-Raphson algorithm. First, lets look at a simple geometric example (scalar case). Let $g(\Theta) = -\frac{\partial \log L(\Theta)}{\partial \Theta}$. We want to find $\tilde{\Theta}$ s.t. $g(\tilde{\Theta}) = 0$, of course, also need to consider 2nd-order condition.

[N.B.: Max. $\log L \Leftrightarrow$ min. $-\log L$ = f.]

Start at an arbitrary point, Θ_0, move to Θ_1; Θ_2 using algorithm, until convergence.



How do we get from θ_0 to θ_1?

Slope of tangent $= \left(\frac{g(\theta_0)}{\theta_0 - \theta_1}\right)$

$= H(\theta_0) \quad \left[= \frac{\partial g(\theta)}{\partial \theta}\right]$

So, $\theta_1 = \theta_0 - g(\theta_0)/H(\theta_0)$

More generally:

$$\theta_{n+1} = \theta_n - H^{-1}(\theta_n) g(\theta_n)$$

[In this form, also o.k. for vector, $\underline{\theta}$.]

Clearly, this is just one possible approach

Where does this come from, mathematically?

(Knowing this will help us see when it may work well, or badly.)

Let $f(\underline{\theta}) = -\log L(\underline{\theta})$. Objective is to min. $f(\underline{\theta})$. Let's approximate $f(\cdot)$ using a 2nd-order Taylor's series expansion

$$f(\underline{\theta}) \simeq f(\tilde{\underline{\theta}}) + (\underline{\theta} - \tilde{\underline{\theta}})' \left(\frac{\partial f(\underline{\theta})}{\partial \underline{\theta}}\right)\Big|_{\underline{\theta}=\tilde{\underline{\theta}}}$$

$$+ \frac{1}{2}(\underline{\theta} - \tilde{\underline{\theta}})' \left[\frac{\partial^2 f(\underline{\theta})}{\partial \underline{\theta}\, \partial \underline{\theta}'}\right]\Big|_{\underline{\theta}=\tilde{\underline{\theta}}} (\underline{\theta} - \tilde{\underline{\theta}})$$

(Approximation is valid in neighbourhood of $\tilde{\underline{\theta}}$, the value that minimizes $f(\cdot)$)

Re-writing this:

$$f(\underline{\theta}) \simeq f(\tilde{\underline{\theta}}) + (\underline{\theta} - \tilde{\underline{\theta}})' g(\tilde{\underline{\theta}}) + \tfrac{1}{2}(\underline{\theta} - \tilde{\underline{\theta}})' H(\tilde{\underline{\theta}})(\underline{\theta} - \tilde{\underline{\theta}})$$

Differentiate:

$$\left[\frac{\partial f(\underline{\theta})}{\partial \underline{\theta}}\right] \approx 0 + g(\tilde{\underline{\theta}}) + (\tfrac{1}{2})(2) H(\tilde{\underline{\theta}})(\underline{\theta} - \tilde{\underline{\theta}})$$

$$= H(\tilde{\underline{\theta}})(\underline{\theta} - \tilde{\underline{\theta}})$$

(because $g(\tilde{\underline{\theta}}) = 0$.)

Re-arranging:

$$\tilde{\underline{\theta}} = \underline{\theta} - H^{-1}(\tilde{\underline{\theta}}) \left[\frac{\partial f(\underline{\theta})}{\partial \underline{\theta}}\right]$$

$$= \underline{\theta} - H^{-1}(\tilde{\underline{\theta}})\, g(\underline{\theta})$$

So, begin with $\underline{\theta} = \underline{\theta}_0$ & iterate:

$$\underline{\theta}_1 = \underline{\theta}_0 - H^{-1}(\underline{\theta}_1)\, g(\underline{\theta}_0)$$

$$\underline{\theta}_2 = \underline{\theta}_1 - H^{-1}(\underline{\theta}_2)\, g(\underline{\theta}_1)$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\underline{\theta}_{n+1} = \underline{\theta}_n - H^{-1}(\underline{\theta}_{n+1})\, g(\underline{\theta}_n)$$

or, approximately —

$$\underline{\theta}_{n+1} = \underline{\theta}_n - H^{-1}(\underline{\theta}_n)\, g(\underline{\theta}_n)$$

(i) Stop if $\left|\frac{\theta_{n+1}^i - \theta_n^i}{\theta_n^i}\right| < \varepsilon_i$ for $i = 1, 2, \ldots, k$.

(ii) Algorithm corresponds to case of $s = 1$.

(iii) $d(\underline{\theta}_n) = -H^{-1}(\underline{\theta}_n)\, g(\underline{\theta}_n)$

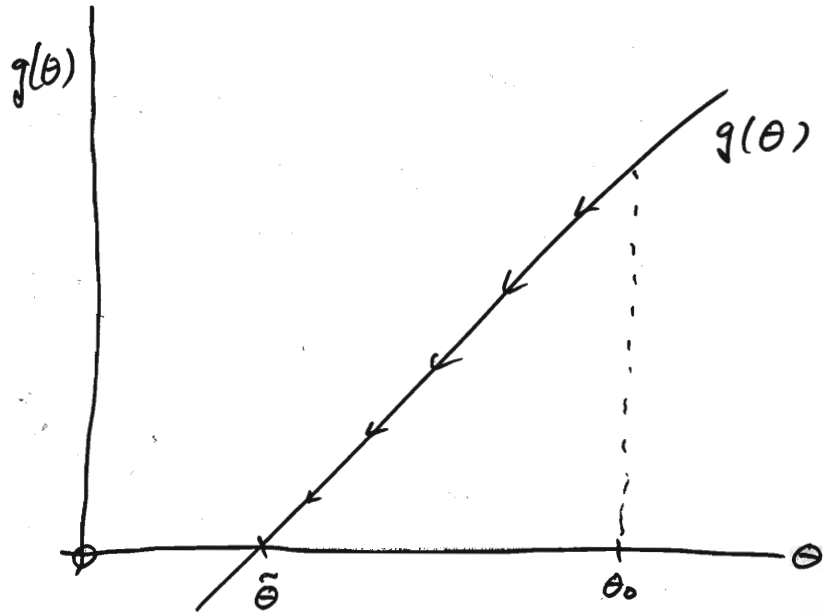
(iv) Algorithm will fail if $H(\cdot)$ singular.

(v) If $H(\cdot)$ becomes n.d., we'll locate a max. of $f(\cdot)$ & hence a min. of $\log L$. So, need to ensure that $H(\tilde{\underline{\theta}})$ is p.d.

(vi) Algorithm may locate a local min. (depending on initial value), or may oscillate & never converge.

(vii) Algorithm should work well if $\log L(\underline{\theta})$ is approximately quadratic, at least in neighbourhood of max.

In fact, if $f(\underset{\sim}{\theta}) = -\lg L(\underset{\sim}{\theta})$ is <u>exactly</u> quadratic, then convergence is immediate. (In this case, $g(\underset{\sim}{\theta})$ is <u>linear</u>.)



Need to experiment with different initial values to ensure we locate a global max. of $\log L$.

<u>Example</u>:

$$f(\theta) = \theta^3 - 2\theta^2 - 1$$

We can locate turning points analytically —

$$f'(\theta) = 3\theta^2 - 4\theta = \theta(3\theta - 4)$$
$$= 0$$

$$\Rightarrow \quad \theta = 0, \ 4/3.$$

$$f''(\theta) = 6\theta - 4 \quad ; \begin{cases} f''(0) = -4 & (\text{max.}) \\ f''(4/3) = 4 & (\text{min.}) \end{cases}$$

Now, let's see how the algorithm goes:

Start at $\theta_0 = 1$:

$$g(\theta_0) = f'(\theta_0) = -1 \quad ; \quad H(\theta_0) = f''(\theta_0) = 2$$

So, $\theta_1 = \theta_0 - H(\theta_0)^{-1} g(\theta_0) = 1 - (-1/2) = 1.5$

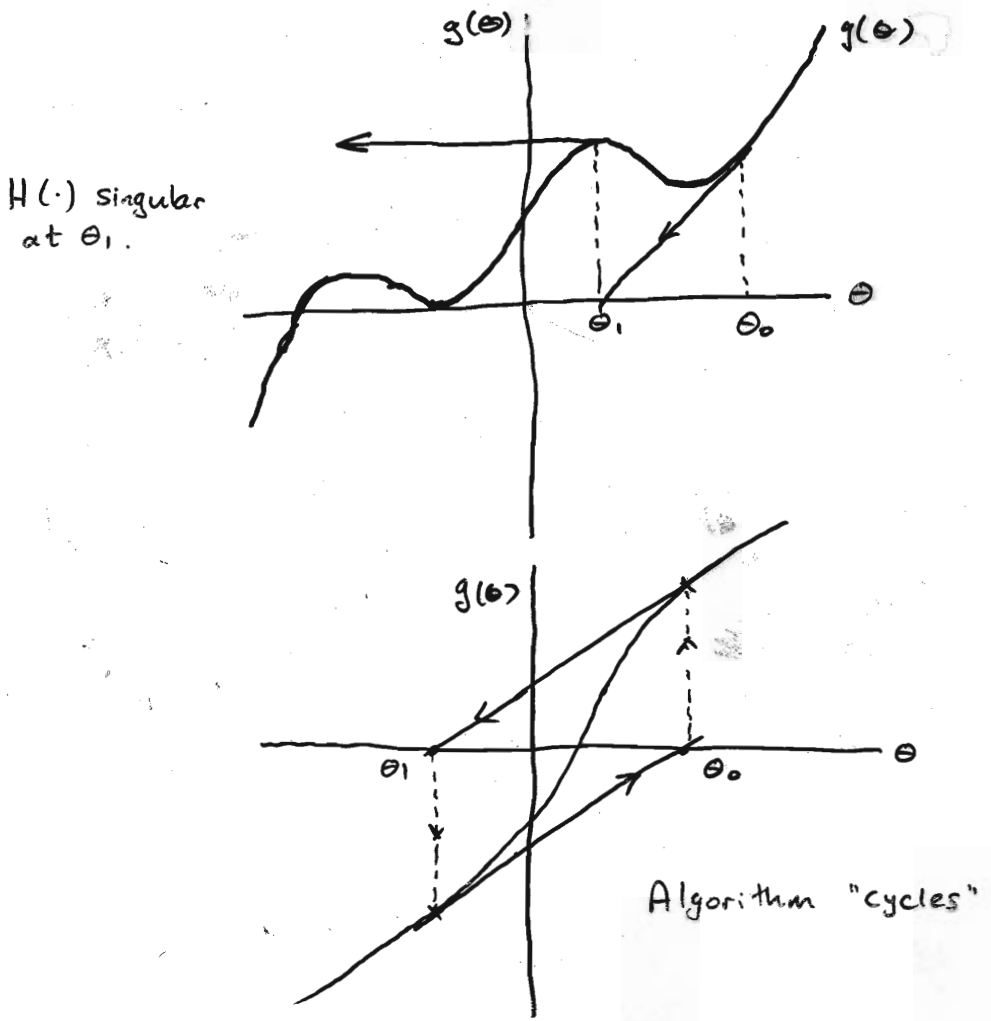
$$g(\theta_1) = 3/4 \quad ; \quad H(\theta_1) = 5$$

So, $\theta_2 = 1.5 - \left(\frac{3/4}{5}\right) = 1.35$

Rapidly get to $\tilde{\theta} = 1.333$.

What happens if we choose $\theta_0 = -1$?

Things that can go wrong:



"Concentrating" the Likelihood Fctn.

Sometimes the parameter vector falls into 2 natural parts: $\underset{\sim}{\theta} = (\underset{\sim}{\theta}_1, \underset{\sim}{\theta}_2)$. So, $L(\underset{\sim}{\theta} | \underset{\sim}{y}) = L(\underset{\sim}{\theta}_1, \underset{\sim}{\theta}_2 | \underset{\sim}{y})$. In addition, often the MLE for $\underset{\sim}{\theta}_2$ can be written as a fctn. of the MLE for $\underset{\sim}{\theta}_1$:

$\tilde{\underset{\sim}{\theta}}_2 = f(\tilde{\underset{\sim}{\theta}}_1)$. In this case,

$$L(\underset{\sim}{\theta}_1, \underset{\sim}{\theta}_2 | \underset{\sim}{y}) = L[\underset{\sim}{\theta}_1, f(\underset{\sim}{\theta}_1) | \underset{\sim}{y}]$$

$$= L_c[\underset{\sim}{\theta}_1 | \underset{\sim}{y}]$$

where $L_c(\cdot)$ is the concentrated likelihood fctn. The point is that we can then maximize $L_c(\cdot)$ with respect to just $\underset{\sim}{\theta}_1$. If we can get $\hat{\underset{\sim}{\theta}}_2 = f(\hat{\underset{\sim}{\theta}}_1)$ analytically this means that the numerical maximization has to be done only w.r.t. $\underset{\sim}{\theta}_1$ & not $\underset{\sim}{\theta}_2$, which simplifies matters significantly.

N.B.: "Concentrating" the L.F. is not always an option — worth keeping in mind, though.

Example: (concentrating is unnecessary, but it still works)

$$y = X\beta + \varepsilon \;;\; \varepsilon \sim N[0, \sigma^2 I]$$

$$p(y \mid \beta, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)\right]$$

$$\log L(\beta, \sigma \mid y) = -n/2 \log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)$$

$$(\partial \log L / \partial \sigma^2) = n/2\sigma^2 + \frac{1}{2\sigma^4}(y-X\beta)'(y-X\beta) = 0$$

$$\Rightarrow \tilde{\sigma}^2 = \frac{1}{n}(y-X\beta)'(y-X\beta)$$

$$\log L_c(\beta \mid y) = -n/2 \log\left[\frac{1}{n}(y-X\beta)'(y-X\beta)\right] - n/2 \log 2\pi - \frac{n}{2}$$

$$(\partial \log L_c / \partial \beta) = \frac{-n}{2\left(\frac{1}{n}(y-X\beta)'(y-X\beta)\right)} \cdot \left[\frac{\partial}{\partial\beta}(y-X\beta)'(y-X\beta)\right] = 0$$

$$\Rightarrow \frac{\partial}{\partial\beta}\left[(y-X\beta)'(y-X\beta)\right] = 0$$

$$\Rightarrow \tilde{\beta} = (X'X)^{-1}X'y$$

& so $\tilde{\sigma}^2 = \frac{1}{n}(y-X\tilde{\beta})'(y-X\tilde{\beta})$

So, "concentrating" the L.F. certainly works. Imagine that we knew how to differentiate w.r.t. a scalar, but not a vector. Then, we could have obtained $L_c(\beta \mid y)$ above & then would have had to $\max_{(\beta)} \{L_c(\beta \mid y)\}$ w.r.t. β.

That is, maximize

$$L_c(\beta \mid y) = -n/2 \log\left[\frac{1}{n}(y-X\beta)'(y-X\beta)\right] - n/2 \log(2\pi) - n/2$$

or, max. $\left\{-\frac{n}{2}\log(y-X\beta)'(y-X\beta)\right\}$ w.r.t. β.

We'll encounter situations where we can maximize partly analytically, but need to use numerical algorithm for the rest of problem. "Concentrating" will reduce dimension of space for numerical optimization.

Inequality Constraints:

Sometimes we'll want to impose an inequality constraint on one or more of our MLE's. For example: $\tilde{\theta}_1 > 0$; $0 < \tilde{\theta}_2 < 1$. Most packages don't allow this. However, there are some tricks we can use to impose such constraints.

Example: We want $\tilde{\theta}_1 \geq 0$. So, replace θ_1 with $(\phi_1)^2$, which will be ≥ 0. No constraint is placed on ϕ_1; $\tilde{\theta}_1 = \tilde{\phi}_1^2$.

Example: We want $0 \leq \theta_2 \leq 1$. Replace θ_2 by $1/(1+\phi^2)$, or replace θ_2 by $\exp(\phi)/(1+\exp(\phi))$ & then $\tilde{\theta}_2 = 1/(1+\tilde{\phi}^2)$, or $1/(1+\exp(\tilde{\phi}))$ will satisfy the constraint.