Convergence order, convergence constant and error estimates

Preliminaries

Assume that we have used an iterative method that generates a sequence $x_0, x_1, \ldots, x_k$ and that we again write the unknown exact solution as:

$$\tilde{x} \quad \text{meaning that} \quad f(\tilde{x}) = 0 \tag{1}$$

We write the error, ϵ_k, for iteration k:

$$\epsilon_k = x_k - \tilde{x} \tag{2}$$

And we write the change at the k'th iteration d_k:

$$d_k = x_k - x_{k-1} \tag{3}$$

Notice that if we assume that the method converges to our solution $\tilde{x}$, we may rewrite ϵ_k as

$$\epsilon_k = -\left[d_{k+1} + d_{k+2} + d_{k+3} + ...\right] \tag{4}$$

In the discussion below, we assume that we have this convergence. Otherwise it actually make little sense to discuss convergence properties and error estimates...

First order methods

If a method satisfies that

$$\frac{d_k}{d_{k-1}} \quad \rightarrow \quad C \quad \text{as} \quad k \rightarrow \infty \tag{5}$$

where $|C| < 1$, we can insert this in Eq.(4) to obtain

$$\epsilon_k \simeq -d_k \left[C + C^2 + C^3 + ...\right] = \frac{-C}{1-C} d_k \tag{6}$$

Hence, we can with this equation estimate the unknown error ϵ_k from the known d_k. Observe that

$$\frac{\epsilon_k}{\epsilon_{k-1}} \simeq \frac{d_k}{d_{k-1}} \simeq C \tag{7}$$

I.e. when Eq.(5) is satisfied, we have a first order method with convergence constant C. Sometimes we only have

$$\frac{|d_k|}{|d_{k-1}|} \quad \rightarrow \quad C < 1 \quad \text{as} \quad k \rightarrow \infty \tag{8}$$

We can insert this in Eq.4 in the same way to obtain

$$|\epsilon_k| \simeq \frac{C}{1-C} |d_k| \tag{9}$$

Higher order methods

For higher order methods, it is practically infeasible to use the same approach as above because convergence is so fast. Hence we do with an estimate of a supremum to the convergence constant in the following sense:

If a method satisfies that there is some k_0 so that

$$\frac{|d_{k+1}|}{|d_k|^\alpha} \leq C \quad \text{for all } k > k_0 \tag{10}$$

where $\alpha > 1$, we can rewrite this as

$$\frac{|d_{k+1}|}{|d_k|} \leq C|d_k|^{\alpha-1} \quad \text{for all } k > k_0$$

For convergence, there must be some $k \geq k_0$ so that

$$|d_{k+1}|, |d_{k+2}|, |d_{k+3}|, \ldots$$

will all be smaller than $|d_k|$. Thus, as $k \geq k_0$, we get

$$\frac{|d_{k+i+1}|}{|d_{k+i}|} \leq C|d_{k+i}|^{\alpha-1} \leq C|d_k|^{\alpha-1} \quad \text{for all } i \geq 0$$

and therefore we have

$$|d_{k+1+n}| \leq (C|d_k|^{\alpha-1})|d_{k+n}| \leq (C|d_k|^{\alpha-1})^2|d_{k+n-1}| \cdots \leq (C|d_k|^{\alpha-1})^n|d_{k+1}|$$

We then get by inserting into Eq.(4)

$$\begin{aligned} |\epsilon_k| &\leq |d_{k+1}|(1 + (C|d_k|^{\alpha-1}) + (C|d_k|^{\alpha-1})^2 + \cdots = \frac{C|d_k|^{\alpha-1}}{1 - C|d_k|^{\alpha-1}}|d_{k+1}| \\ &\leq \frac{C|d_k|^{\alpha-1}}{1 - C|d_k|^{\alpha-1}} C|d_k|^\alpha \end{aligned}$$

With the often reasonable assumption that $|d_k|$ is sufficiently small so that $C|d_k|^{\alpha-1} < \frac{1}{2}$, we get

$$|\epsilon_k| \leq C|d_k|^\alpha \tag{11}$$

In theory (when roundoff errors were not present), we would check

$$\frac{|d_k|}{|d_{k-1}|^\alpha} \to C \text{ for } k \to \infty \tag{12}$$

We would then get $|\epsilon_k| \simeq C|d_k|^\alpha$ and hence

$$\frac{|\epsilon_k|}{|\epsilon_{k-1}|^\alpha} \simeq \frac{C|d_k|^\alpha}{(C|d_{k-1}|^\alpha)^\alpha} = \frac{C}{C^\alpha}\frac{|d_k|^\alpha}{(|d_{k-1}|^\alpha)^\alpha} \simeq \frac{C}{C^\alpha}C^\alpha = C \tag{13}$$

Hence when Eq.(12) is satisfied, the method has convergence order α and convergence constant C.

How to apply this to obtain the errors

We typically don't know the exact solution(that is why we are looking for it...). But now we have the estimates in Eq.6 and Eq.9 for the first order methods and Eq.11 for higher order methods.

We know that Bisection is a first order method with $\overline{C} = \frac{1}{2}$. Regula Falsi is also *expected* to be a first order method with a problem dependent convergence constant C. Newton and Ridders are *expected* to be second and third order methods respectively with a problem dependent convergence constant C. Notice that both Newton (due to the derivative) and Ridders requires two function evaluations per iteration. Hence they can also be viewed as having order $\sqrt{2}$ and $\sqrt{3}$ respectively with convergence constant $\sqrt{C}$ per function evaluation. The Secant method is *expected* to be have order $\frac{1}{2}(1+\sqrt{5})$ also with a problem dependent convergence constant. The convergence properties is however not always as *expected*. Hence this should be checked using the output data.

We can summarize with the following table

Method	Expected Order	Estimate of C	Estimate of $\lvert\epsilon_k\rvert$
Newton	2	$\frac{\lvert d_k\rvert}{\lvert d_{k-1}\rvert^2}$	$C\lvert d_k\rvert^2$
Secant	$\frac{1}{2}(1+\sqrt{5}) \simeq 1.62$	$\frac{\lvert d_k\rvert}{\lvert d_{k-1}\rvert^{1.62}}$	$C\lvert d_k\rvert^{1.62}$
Bisection	1	$\frac{1}{2}$	$\lvert d_k\rvert$
False Position	1	$\frac{d_k}{d_{k-1}}$	$\frac{-C}{1-C}d_k$
Ridders	3	$\frac{\lvert d_k\rvert}{\lvert d_{k-1}\rvert^3}$	$C\lvert d_k\rvert^3$

For the higher order methods (Secant, Newton, Ridders), it can be difficult to estimate the convergence constant as convergence is so fast. Hence, a supremum may be guessed as outlined above. For first order methods, it is usually easy to estimate C.