Optimization

Class 9: How Methods Are Ranked

1 Convergence

Traditionally, we don't even bother to rank the performance of a method unless the guesses $\mathbf{x}$ exhibit *convergence* by becoming arbitrarily close to a "target" $\mathbf{x}^*$ (presumed to be a solution). The mathematical shorthand for this property is $\mathbf{x} \to \mathbf{x}^*$. In practice when solving a particular optimization problem, you might not need this degree of accuracy; but in general there is no fixed degree of accuracy that will cover all particular optimization problems.

If the guesses do not exhibit convergence, we say that they *diverge* (including aimless wandering, cycling, and approaching infinity).

2 Order of convergence

Assuming $\mathbf{x} \to \mathbf{x}^*$, we rank the method via the following two "speed" measures:

- time/effort to determine each new guess (Newton's takes a lot for $\nabla^2 f$, quasi-Newton methods take less)
- how quickly the guesses approach the target (via the concept of *order of convergence*)

We measure the approach to the target $\mathbf{x}^*$ via errors at consecutive steps:

$$\text{error}_{\text{old}} = ||\mathbf{x}_{\text{old}} - \mathbf{x}^*||$$
$$\text{error}_{\text{new}} = ||\mathbf{x}_{\text{new}} - \mathbf{x}^*||$$

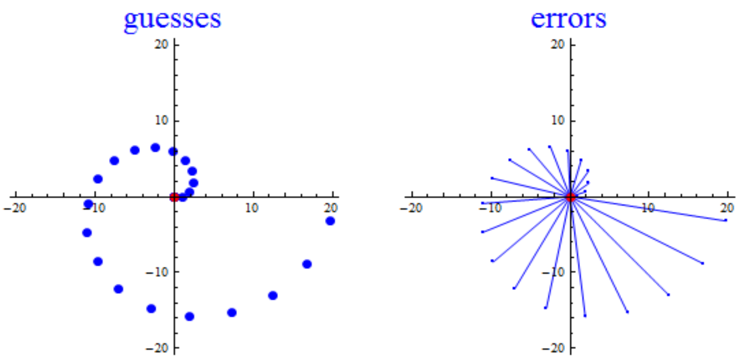

Figure 1: Guesses converging to $(0, 0)$, and the associated errors.

The order of convergence is determined by comparing consecutive errors. For example, a linear relationship between the consecutive errors means:

$$\text{error}_{\text{new}} = \delta\big(\text{error}_{\text{old}}\big) \quad \text{for some "slope" } \delta \geq 0 \tag{1}$$

Q: What does this say if $\delta > 1$?
A: New error larger than old. If this persists, the guesses diverge.

Q: What values of $\delta > 0$ are best?
A: The smaller the better, since then the error is reduced by a more at each step.

Traditionally, we use the fractional version of the linear relationship (1)

$$\frac{\text{error}_{\text{new}}}{\text{error}_{\text{old}}} = \delta,$$

and want this relationship to hold eventually for all guesses (possibly with different $\delta < 1$):

$$\textit{linear order of convergence} \quad \Leftrightarrow \quad \boxed{\limsup \frac{\text{error}_{\text{new}}}{\text{error}_{\text{old}}} \leq \delta < 1},$$

where the lim sup indicates the (smallest) eventual upper-bound on the error quotients. The lim sup is a technicality to cover situations when the error quotients don't actually converge (e.g., they cycle between 1/4 and 1/2, in which case the lim sup is 1/2). When the error quotients do converge, the lim sup is the same as the limit. Notice that $\mathbf{x} \to \mathbf{x}^*$ automatically follows from $\limsup \frac{\text{error}_{\text{new}}}{\text{error}_{\text{old}}} \leq \delta < 1$ since the new error is eventually at most a fraction (less than one) of the old error; which means the errors converge to zero.

When $\delta = 0$, we say the method exhibits a *super*linear order of convergence (since that's the best possible linear convergence constant).

2.1 Quadratic order of convergence

Since we are assuming $\mathbf{x} \to \mathbf{x}^*$, it is even better to have the quadratic relationship

$$\text{error}_{\text{new}} = \delta\big(\text{error}_{\text{old}}\big)^2,$$

which corresponds to

$$\textit{quadratic order of convergence} \quad \Leftrightarrow \quad \boxed{\limsup \frac{\text{error}_{\text{new}}}{\big(\text{error}_{\text{old}}\big)^2} \leq \delta < \infty} \;\; \& \;\; \boxed{\mathbf{x} \to \mathbf{x}^*}.$$

Group Problem 5.1.

(a) Use the identity

$$\big(\text{error}_{\text{old}}\big)\,\big(\text{error}_{\text{old}}\big) = \big(\text{error}_{\text{old}}\big)^2$$

to explain why a quadratic order of convergence is at least as good as a superlinear order of convergence.

(b) Generate the first three guesses in a sequence that appears to approach $x^* = 1$ and that exhibits a quadratic order of convergence with constant $\delta = 4$.

Group Problem 5.1. (solution)

(a) Applying this identity to the defining inequality for the quadratic order of convergence, we get

$$\limsup \frac{1}{(\text{error}_{\text{old}})} \frac{\text{error}_{\text{new}}}{\text{error}_{\text{old}}} \leq \delta < \infty. \tag{2}$$

Since $\mathbf{x} \to \mathbf{x}^*$, we know that the old errors approach zero (so their inverses $\frac{1}{(\text{error}_{\text{old}})}$ approach infinity). Thus, inequality (2) implies that the linear error ratios satisfy

$$\limsup \frac{\text{error}_{\text{new}}}{\text{error}_{\text{old}}} = 0.$$

This is the definition of a superlinear order of convergence.

(b) One sequence of guesses that appears to approach $x^* = 1$, and that exhibits a quadratic order of convergence with constant $\delta = 4$ is

$$1.1,\ 1.04,\ 1.0064,\ 1.00016384,\ 1.0000001074\ldots$$

the n-th term of which is obtained from $1 + 4^{2^n-1}(0.1)^{2^n}$. More generally, any sequence whose n-th term is the form $1 + 4^{2^n-1}(m)^{2^n}$ for some fixed $m < 0.25$ has these same properties.

Notice that the limiting quotient

$$\limsup \frac{\text{error}_{\text{new}}}{\left(\text{error}_{\text{old}}\right)^2} \leq \delta < \infty$$

alone (without $\mathbf{x} \to \mathbf{x}^*$) does not necessarily indicate a desirable situation. For example, the sequence of guesses 2^1, 2^2, 2^4, $2^8, \ldots$ does not converge to the target $x^* = 0$ (or to any target), yet the consecutive errors always satisfy $\frac{\text{error}_{\text{new}}}{\left(\text{error}_{\text{old}}\right)^2} = 1$.

3 Ranking Newton's method

If the guesses generated by Newton's optimization method converge, they will always exhibit a quadratic order of convergence. We'll use the following group problem to show this in the case of one-variable.

Group Problem

For the one-variable version of Newton's optimization method

$$x_{\text{new}} = x_{\text{old}} - \frac{f'(x_{\text{old}})}{f''(x_{\text{old}})},$$

use the remainder Taylor quadratic for f' at x_{old}:

$$f'(x) = f'(x_{\text{old}}) + f''(x_{\text{old}})\big(x - x_{\text{old}}\big) + \frac{f'''(\xi)}{2}\big(x - x_{\text{old}}\big)^2$$

to explain why

- $0 = f'(x_{\text{old}}) + f''(x_{\text{old}})\big(x^* - x_{\text{old}}\big) + \dfrac{f'''(\xi)}{2}\big(\text{error}_{\text{old}}\big)^2$

- $x_{\text{new}} - x^* = \dfrac{f'''(\xi)}{2f''(x_{\text{old}})}\big(\text{error}_{\text{old}}\big)^2$

- $\limsup \dfrac{\text{error}_{\text{new}}}{(\text{error}_{\text{old}})^2} = \left|\dfrac{f'''(x^*)}{2f''(x^*)}\right|$

Group Problem (solution)

We don't know anything about the target x^* except that it is a stationary point of f: $f'(x^*) = 0$. To use this information, we create the remainder form of the Taylor quadratic associated with the *derivative* function $f'(x)$:

$$f'(x) = f'(x_{\text{old}}) + f''(x_{\text{old}})\big(x - x_{\text{old}}\big) + \frac{f'''(\xi)}{2}\big(x - x_{\text{old}}\big)^2,$$

where ξ is some point between x and x_{old}. Evaluating this at $x = x^*$ gives

$$0 = f'(x_{\text{old}}) + f''(x_{\text{old}})\big(x^* - x_{\text{old}}\big) + \frac{f'''(\xi)}{2}\big(\text{error}_{\text{old}}\big)^2$$

where the term $\big(\text{error}_{\text{old}}\big)^2$ replaces $\big(x^* - x_{\text{old}}\big)^2$.

Moving the first two terms to the left side and dividing by $f''(x_{\text{old}})$ (assuming this term is not zero) yields

$$\begin{aligned}
x_{\text{old}} - \frac{f'(x_{\text{old}})}{f''(x_{\text{old}})} - x^* \quad &= \quad \frac{f'''(\xi)}{2f''(x_{\text{old}})}\big(\text{error}_{\text{old}}\big)^2 \\
&\updownarrow \\
x_{\text{new}} - x^* \quad &= \quad \frac{f'''(\xi)}{2f''(x_{\text{old}})}\big(\text{error}_{\text{old}}\big)^2 \\
&\updownarrow \\
\frac{\text{error}_{\text{new}}}{\big(\text{error}_{\text{old}}\big)^2} \quad &= \quad \left|\frac{f'''(\xi)}{2f''(x_{\text{old}})}\right|.
\end{aligned}$$

Since ξ is between x^* and x_{old}, and since $x_{\text{old}} \to x^*$, we know that in the limit there is no difference between the points ξ, x_{old}, and x^*:

$$\limsup \frac{\text{error}_{\text{new}}}{(\text{error}_{\text{old}})^2} = \left|\frac{f'''(x^*)}{2f''(x^*)}\right|,$$

which means a quadratic order of convergence with constant

$$\delta = \left|\frac{f'''(x^*)}{2f''(x^*)}\right|$$

as long as the denominator is not zero.

3.1 More variables

The argument is essentially similar (but more complicated because there isn't a simple multivariable analog to the third derivative), and yields:

$$\delta = \frac{K}{\min\{|\epsilon| \text{ over all eigenvalues } \epsilon \text{ of } \nabla^2 f(\mathbf{x}^*)\}}$$

for some positive constant K (a surrogate for the third derivative in the one-variable formula).

4 Ranking Quasi-Newton methods

In general, the guesses generated by quasi-Newton methods do not exhibit quadratic orders of convergence (e.g., steepest descent gets only linear).

MORAL: Nothing's perfect. Newton's method is worse for per-step time/effort, but better for order of convergence. Steepest descent is better for per-step time/effort, but worse for order of convergence. Other quasi-Newton methods may fall between these two extremes.

5 Lab Exercise

Use a spreadsheet to compute the error fraction $\dfrac{\text{error}_{\text{new}}}{(\text{error}_{\text{old}})^2}$ for Newton's optimization method applied to

(i) the two-variable discretization function $f(x, y) = x^3 - x\,y^2 + y^3 - y + 1$

(ii) the monthly payment function

$$f(x, y) = \frac{(30 - y)\left(1 + \frac{10^9}{(108-x)^2\,(30-y)^5}\right)^{\frac{x}{12}}}{x}$$

You should observe a quadratic order of convergence in each case. What are the corresponding constants δ?

For solutions, see Class9.nb or Lab Exercise (solutions).xlsx.