

C477: Computing for Optimal Decisions

The Newton-Raphson and Related Methods

Panos Parpas
Department of Computing
Imperial College London

www.doc.ic.ac.uk/~pp500
p.parpas@imperial.ac.uk

Review of 1-D Newton-Raphson Method

$$\min f(x)$$

- ❶ Minimising a **general** non-linear function is difficult
- ❷ **Basic idea:** minimise a quadratic approximation

$$\min q(x)$$

where $q(x) = f(x_k) + f'(x_k)(x - x_k) + \frac{1}{2}f''(x_k)(x - x_k)^2$

- ❸ Minimise **quadratic approximation**

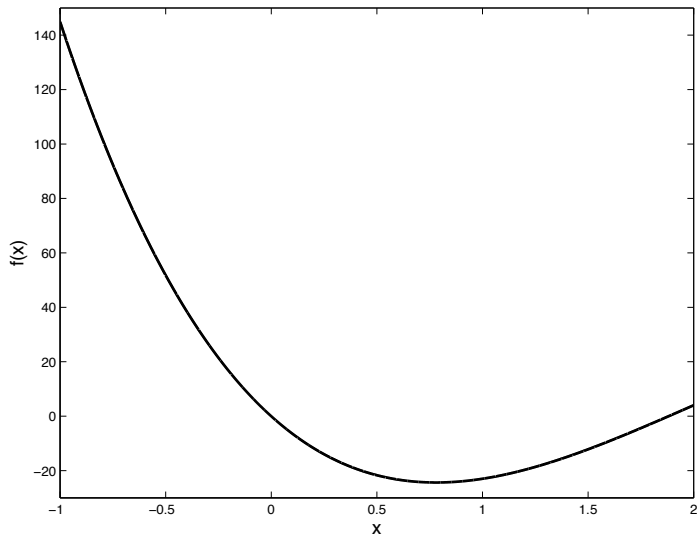
$$0 = q'(x) = f'(x_k) + f''(x_k)(x - x_k)$$

- ❹ **Iterate**

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$$

Example: One Dimensional Newton's Method

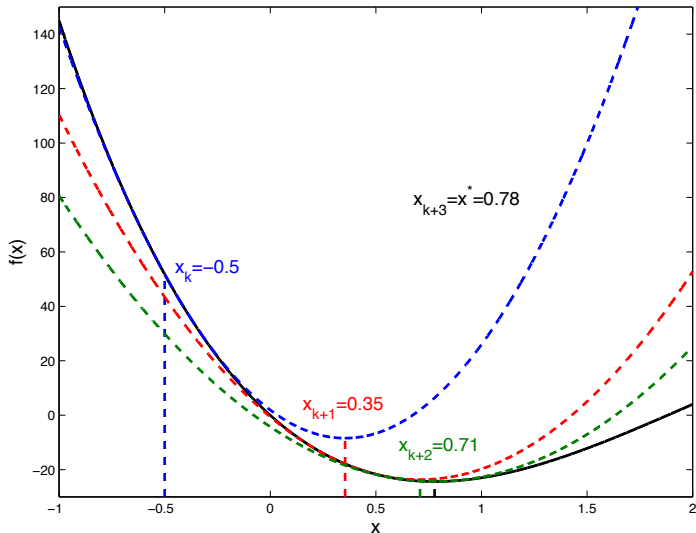
$$\min f(x) = x^4 - 14x^3 + 60x^2 - 70x$$



Example: One Dimensional Newton's Method

$$x_{k+3} = \arg \min f(0.71) + f'(0.71)(x - 0.71) + \frac{1}{2}f''(0.71)(x - 0.71)^2$$

Convergence after 3 iterations!



Towards a general Newton-Raphson Method

Issues with the Newton-Raphson Method we studied so far,

- (a) Only applicable to single dimension
- (b) The algorithm may cycle
- (c) It may fail to find a descent direction
- (d) It may converge to a saddle point or a local maximum

In this lecture:

- (a) Multivariate extension
- (b) Discuss conditions & modifications for guaranteed convergence
- (c) Discuss convergence rates & practical implementation

In the next lecture

- (a) Constrained Optimality Conditions
- (b) Constrained Optimisation Algorithms

Multivariate Newton-Raphson Method

General problem,

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x})$$

1. As in 1-d case we construct a **quadratic** approximation around the current iterate \mathbf{x}_k (second order Taylor series expansion)

$$\begin{aligned} f(\mathbf{x}) &\approx f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T (\mathbf{x} - \mathbf{x}_k) + \frac{1}{2} (\mathbf{x} - \mathbf{x}_k)^T \nabla^2 f(\mathbf{x}_k) (\mathbf{x} - \mathbf{x}_k) \\ &\triangleq q(\mathbf{x}) \end{aligned}$$

2. Apply the FONC to $q(\mathbf{x})$,

$$0 = \nabla q(\mathbf{x}) = \nabla f(\mathbf{x}_k) + \nabla^2 f(\mathbf{x}_k) (\mathbf{x} - \mathbf{x}_k)$$

3. Assume that $\nabla^2 f(\mathbf{x}_k) \succ 0$ (i.e. positive definite), then

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \nabla^2 f(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k)$$

Why is the assumption $\nabla^2 f(\mathbf{x}_k) \succ 0$ needed?

If $\nabla^2 f(\mathbf{x}_k)$ is positive definite then the Newton direction

$$\mathbf{d}_k = -\nabla^2 f(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k)$$

is a descent direction,

$$\nabla f(\mathbf{x}_k)^T \mathbf{d}_k = -\nabla f(\mathbf{x}_k)^T \nabla^2 f(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k) < 0.$$

Remark: Note that if a matrix is positive definite then so is its inverse, see for example any Linear algebra book e.g. *Matrix Analysis*, R.A. Horn, C.R. Johnson

Convergence Theory

Theorem

Suppose that f is three times continuously differentiable and that $\mathbf{x}^* \in \mathbb{R}^n$ satisfies,

$$\nabla f(\mathbf{x}^*) = 0$$

and that $\nabla^2 f(\mathbf{x}^*)$ is invertible. Then for all \mathbf{x}_0 (starting point) sufficiently close to \mathbf{x}^* the following holds,

- (1) Newton's method is **well defined** for all k .
- (2) The method **converges to \mathbf{x}^*** .
- (3) The order of **convergence is quadratic**.

Remarks:

- (a) Conditions $\nabla f(\mathbf{x}^*) = 0$ & $\nabla^2 f(\mathbf{x}^*)$ invertible hold for *local maxima as well*. **The theorem does not say the method will converge to a minimum.**
- (b) The starting point needs to be close to the solution

Is the Newton algorithm a descent algorithm?

- 1 Given a point \mathbf{x}_k .
- 2 Derive a **descent** direction $\mathbf{d}_k \in \mathbb{R}^n$, i.e.

$$\nabla f(\mathbf{x}_k)^T \mathbf{d}_k < 0.$$

- 3 Decide on a step-size α_k .
- 4 Transition to the next point,

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$$

Is the Newton algorithm a descent algorithm?

Theorem

Suppose that $\{\mathbf{x}_k\}$ is a sequence generated by the algorithm. If the Hessian $\nabla^2 f(\mathbf{x}^k) \succ 0$ and $\nabla f(\mathbf{x}^k) \neq 0$ then the search direction

$$\mathbf{d}_k = -\nabla^2 f(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k) = \mathbf{x}_{k+1} - \mathbf{x}_k$$

is a descent direction for f in the sense that there exists an $\alpha \in (0, \bar{\alpha})$ such that for all $\alpha \in (0, \bar{\alpha})$,

$$f(\mathbf{x}_k + \alpha \mathbf{d}_k) < f(\mathbf{x}_k).$$

The Newton algorithm is a descent algorithm with a descent direction given by

$$-\nabla^2 \mathbf{f}(\mathbf{x})^{-1} \nabla \mathbf{f}(\mathbf{x})$$

Line Search & Backtracking

Exact line search: The result in previous slide motivates the modification of the Newton method,

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla^2 f(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k)$$

where $\alpha_k = \arg \min_{\alpha \geq 0} f(\mathbf{x}_k - \alpha \nabla^2 f(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k))$ (exact line search)
Other types of line search algorithms are also used.

Backtracking:

while

Do not have sufficient decrease in objective function value

do

Reduce step size

Line Search & Backtracking

Exact line search: The result in previous slide motivates the modification of the Newton method,

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla^2 f(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k)$$

where $\alpha_k = \arg \min_{\alpha \geq 0} f(\mathbf{x}_k - \alpha \nabla^2 f(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k))$ (exact line search)

Other types of line search algorithms are also used.

Backtracking: Given two constants $0 < \beta < 0.5$, and $0 < \gamma < 1$ and a descent direction d then

while

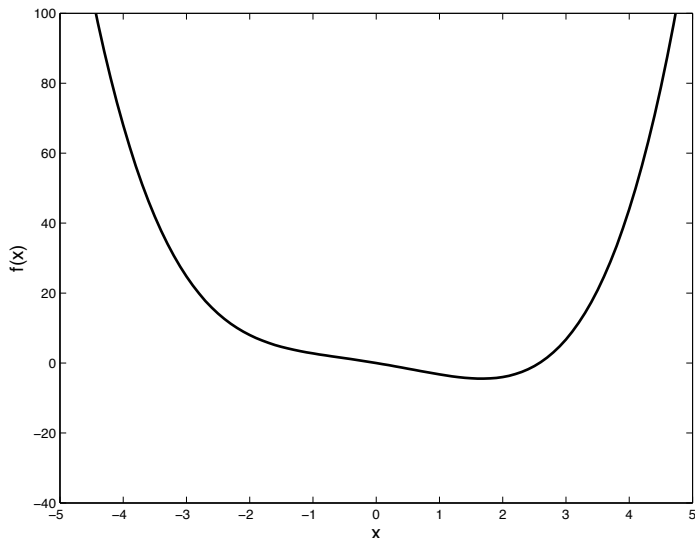
$$f(\mathbf{x} + \alpha d) > f(\mathbf{x}) + \alpha \beta \nabla f(\mathbf{x})^T d$$

do

$$\alpha \leftarrow \gamma \alpha$$

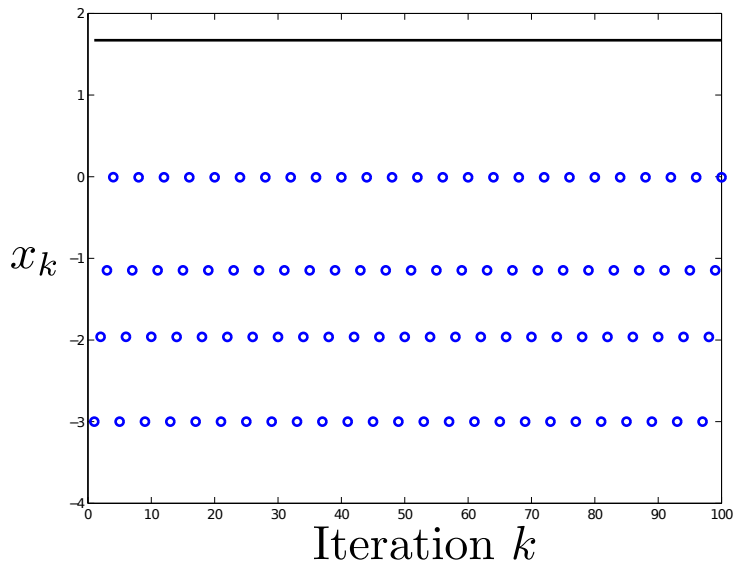
Example: Convergence Problems

$$\min \frac{1}{4}x^4 - \frac{1}{2}x^2 - 3x \quad \text{Initial Point } x_0 = -3.0$$



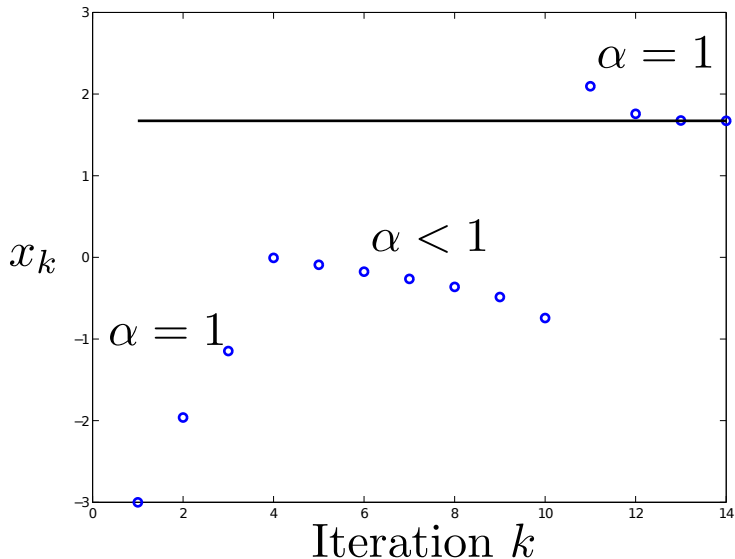
Example: newtonExample0.m

$\min x^4/4 - x^2/2 - 3x$ $x_0 = -3$ no line search



Example: newtonExample0.m

$\min x^4/4 - x^2/2 - 3x$ $x_0 = -3$ with line search



Convergence Theory: Positive Hessian

Key Assumption: The Hessian satisfies,

$$m\mathbf{I} \preceq \nabla^2 f(x)$$

for some scalar $m > 0$ (this implies that the function is strongly convex and that it has a unique global minimum).

There exists a constants $\eta > 0$ and $\theta > 0$ such that

- If $\|\nabla f(\mathbf{x}_k)\|_2 > \eta$ (far away from the solution) then

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq -\theta$$

i.e. the objective function is reduced at every iteration.

- If $\|\nabla f(\mathbf{x}_k)\|_2 \leq \eta$ (close to a solution) then the algorithm converges to the minimum with a quadratic rate.

Illustration (a, b, c) randomly generated

$$\min c^T x - \sum_{i=1}^{500} \ln(b_i - a_i^T x) \quad \text{Backtracking } \beta = 0.01, \gamma = 0.5$$

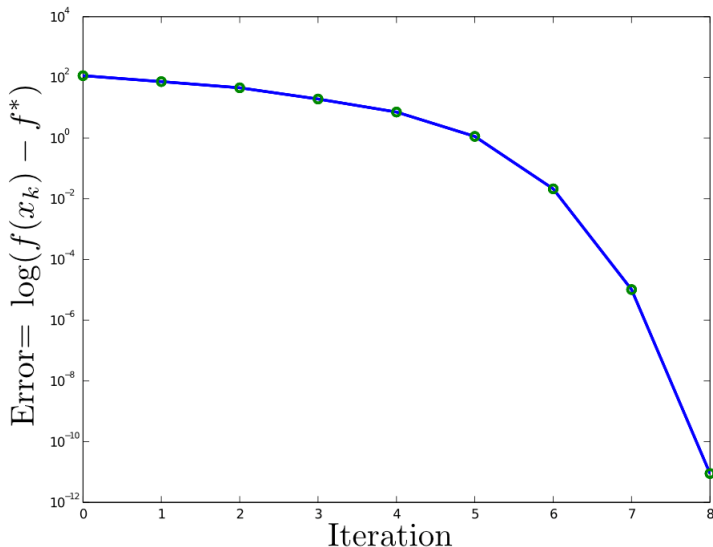
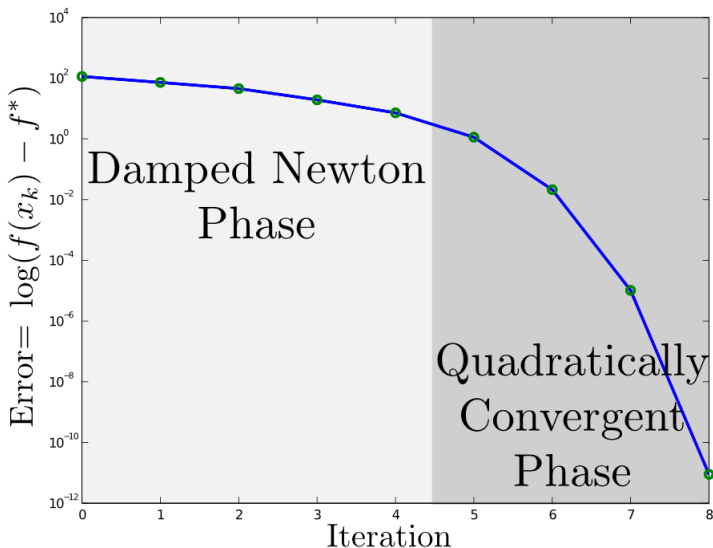


Illustration (a, b, c) randomly generated

$$\min c^T x - \sum_{i=1}^{500} \ln(b_i - a_i^T x) \quad \text{Backtracking } \beta = 0.01, \gamma = 0.5$$



Levenberg–Marquardt Modification

If the Hessian $\nabla^2 f(\mathbf{x}_k)$ is not positive definite then the search direction

$$\mathbf{d}_k = \nabla^2 f(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k)$$

may not be a descent direction.

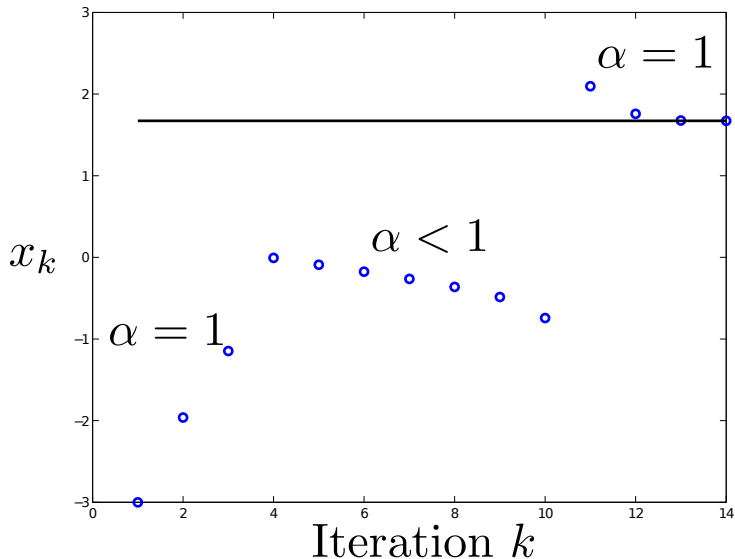
Levenberg–Marquardt Modification:

$$\mathbf{d}_k = (\nabla^2 f(\mathbf{x}_k) + \mu_k \mathbf{I})^{-1} \nabla f(\mathbf{x}_k)$$

- As $\mu_k \rightarrow \infty$ method is like steepest descent with a small step size
- As $\mu_k \rightarrow 0$ method is like Newton Raphson
- In practice, start with a small μ and increase it until a descent condition is satisfied

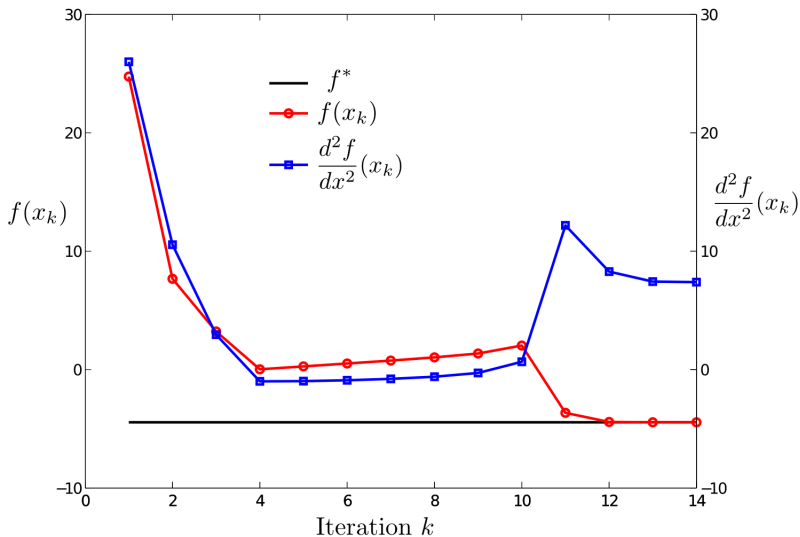
Example: newtonExample0.m

$\min x^4/4 - x^2/2 - 3x$ $x_0 = -3$ with line search



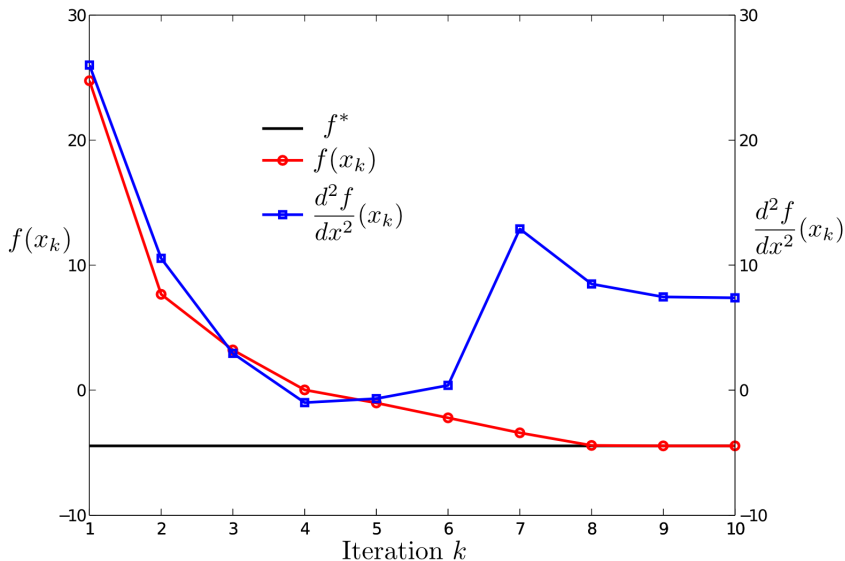
Example: newtonExample0LV.m

$\min x^4/4 - x^2/2 - 3x$ $x_0 = -3$ with line search



Example: newtonExample0LV.m

With line search & Levenberg–Marquardt Modification ($\mu = 10$)



Quasi Newton Methods

- If the function is convex then Newton-Raphson with a line-search works well:
 - (1) Guaranteed to converge from any starting point (globally convergent)
 - (2) Quadratic rate of convergence
 - (3) Careful/Robust implementations available
- In general the method is not guaranteed to converge from any starting point (usually only locally convergence can be guaranteed)
- Computationally expensive if Hessian is large & dense

Quasi Newton Methods: (not covered in this course)

- Iteratively construct an approximation of $\nabla^2 f(\mathbf{x}_k)^{-1}$.
- Most methods generate positive definite approximations
- Algorithms are globally convergent
- State-of-the-art in unconstrained optimisation