Numerical Methods I: Newton and nonlinear least squares

Georg Stadler
Courant Institute, NYU
stadler@cims.nyu.edu

October 12, 2017
Newton’s method to solve $F(x) = 0$, $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$
Newton’s method: Example

In one dimension, solve \( f(x) = 0 \) with \( f : \mathbb{R} \to \mathbb{R} \):

Start with \( x_0 \), and compute \( x_1, x_2, \ldots \) from

\[
x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, \ldots
\]

Requires \( f(x_k) \neq 0 \) to be well-defined (i.e., tangent has nonzero slope).
Newton’s method

Let \( F : \mathbb{R}^n \rightarrow \mathbb{R}^n, \ n \geq 1 \) and solve

\[
F(x) = 0.
\]

Taylor expansion about starting point \( x^0 \):

\[
F(x) = F(x^0) + F'(x^0)(x - x^0) + o(|x - x^0|) \quad \text{for} \ x \rightarrow x^0.
\]

Hence:

\[
x^1 = x^0 - F'(x^0)^{-1} F(x^0)
\]

Newton iteration: Start with \( x^0 \in \mathbb{R}^n \), and for \( k = 0, 1, \ldots \) compute

\[
F'(x^k) \Delta x^k = -F(x^k), \quad x^{k+1} = x^k + \Delta x^k
\]

Requires that \( F'(x^k) \in \mathbb{R}^{n \times n} \) is invertible.
**Newton’s method**

Newton iteration: Start with $x^0 \in \mathbb{R}^n$, and for $k = 0, 1, \ldots$ compute

$$F'(x^k) \Delta x^k = -F(x^k), \quad x^{k+1} = x^k + \Delta x^k$$

Equivalently:

$$x^{k+1} = x^k - F'(x^k)^{-1} F(x^k)$$

Newton’s method is **affine invariant**, that is, the sequence is invariant to affine transformations:

Instead of solving $F(x) = 0$ solve $A F(x) - G(x) = 0$

Newton iteration for $G(x) = 0$:

$$\Delta x = -G'(x^k)^{-1} G(x^k) = (A F'(x^k))^{-1} A F(x^k)$$

$$= F'(x^k)^{-1} A^{-1} A F(x^k) = F'(x^k)^{-1} F(x^k)$$

Newton step computed using $F$
Newton’s method

Intersection point between

\[ F_i(x, y) = y - x^2 + x = 0 \text{ parabola} \]
\[ F_2(x, y) = \frac{x^2}{16} + y^2 - 1 = 0 \text{ ellipse} \]

\[ F(x, y) = \begin{pmatrix} F_1(x, y) \\ F_2(x, y) \end{pmatrix} = \begin{pmatrix} y - x^2 + x \\ \frac{x^2}{16} + y^2 - 1 \end{pmatrix}, \quad F : \mathbb{R}^2 \to \mathbb{R}^2 \]

\[ F'(x, y) = \begin{pmatrix} \frac{\partial F_1(x, y)}{\partial x} & \frac{\partial F_1(x, y)}{\partial y} \\ \frac{\partial F_2(x, y)}{\partial x} & \frac{\partial F_2(x, y)}{\partial y} \end{pmatrix} = \begin{pmatrix} -2x + 1 & 1 \\ \frac{x}{8} & 2y \end{pmatrix} \]

Newton iterations

\[ x^0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \text{Solve} \quad \begin{pmatrix} -1 & 1 \\ \frac{1}{8} & 0 \end{pmatrix} \Delta x = -F(x) \]

\[ x' = x^0 + \Delta x \text{ repeat, stop if } \|F(x^k)\| < \varepsilon \]
Convergence of Newton’s method

Assumptions on $F$: $D \subset \mathbb{R}^n$ open and convex, $F : D \to \mathbb{R}^n$ continuously differentiable with $F'(x)$ is invertible for all $x$, and there exists $\omega \geq 0$ such that

$$
\| F'(x)^{-1}(F'(x + sv) - F'(x))v \| \leq s\omega \| v \|^2
$$

for all $s \in [0, 1]$, $x \in D$, $v \in \mathbb{R}^n$ with $x + v \in D$.

Assumptions on $x^*$ and $x^0$: There exists a solution $x^* \in D$ and a starting point $x^0 \in D$ such that

$$
\rho := \| x^* - x^0 \| \leq \frac{2}{\omega} \text{ and } B_\rho(x^*) \subset D
$$

Theorem: Then, the Newton sequence $x^k$ stays in $B_\rho(x^*)$ and

$$
\lim_{k \to \infty} x^k = x^*, \text{ and }
$$

$$
\| x^{k+1} - x^* \| \leq \frac{\omega}{2} \| x^k - x^* \|^2
$$
Convergence of Newton’s method

$x, y \in \mathbb{D}$

Mean theorem for integrals:

$$F(y) - F(x) = F'(x)(y - x) - \int_0^1 (F'(x + s(y - x)) - F'(x))(y - x) \, ds$$

Take norms:

$$\|F(x)^{-1} \| \leq \int_0^1 s \|y - x\|^2 \, ds = \frac{w}{2} \|y - x\|^2$$

$$\left\| F'(x)^{-1} (F(y) - F(x) - F'(x)(y - x)) \right\| \leq \frac{w}{2} \|y - x\|^2 \quad (1)$$

$$x^{k+1} - x^* = x^k - F(x^k)^{-1} F(x^k) - x^* = 0$$

$$= x^k - x^* - F'(x^k)^{-1} (F(x^k) - F(x^*))$$

$$= F'(x^k)^{-1} \left[ F(x^*) - F(x^k) - F'(x^*) (x^* - x) \right] \quad (2)$$

$$(1) + (2)$$

$$\|x^{k+1} - x^*\| \leq \frac{w}{2} \|x^k - x^*\|^2 \rightarrow \text{Quadratic Convergence}$$
Convergence of Newton’s method

\[ 0 < \| x^{k+1} - x^* \| \leq 8 \]

\[ \| x^{k+1} - x^* \| \leq \frac{w}{2} \| x^k - x^* \| \| x^k - x^\circ \| \leq \frac{w}{2} \cdot 8 \leq 1 \]

\[ \Rightarrow \text{ I'm staying in } B_8(x^\circ) \text{ if } x^k \text{ is in } B_8(x^\circ) \]

Uniqueness \( x^* \) solution

\[ \| x^k - x^\circ \| \leq \frac{w}{2} \| x^k - x^\circ \| \| x^k - x^\circ \| \leq 1 \]

\[ \Rightarrow \| x^k - x^* \| \leq \| x^k - x^\circ \| \]

\[ \Rightarrow x^* = x^k \]

\[ \square. \]
Newton’s method—when does convergence theorem apply?

Example 1: \( f(x) = x^3 \)

\[ x^* = 0 \text{ solution} \]

\( f'(x^*) = 0 \)

1. \( x: \ 0.666666666666 \)
2. \( x: \ 0.444444444444 \)
3. \( x: \ 0.296296296296 \)

... 

17. \( x: \ 0.001014959227 \)
18. \( x: \ 0.000676639485 \)
19. \( x: \ 0.000451092990 \)
20. \( x: \ 0.000300728660 \)

Example 2: \( f(x) = x^{3/2} \)

\[ x^* = 0 \text{ solution} \]

\( f'(x^*) = 0 \)

1. \( x: \ 0.333333333333 \)
2. \( x: \ 0.111111111111 \)
3. \( x: \ 0.037037037037 \)
4. \( x: \ 0.012345679012 \)

... 

16. \( x: \ 0.0000000023231 \)
17. \( x: \ 0.0000000007744 \)
18. \( x: \ 0.0000000002581 \)
19. \( x: \ 0.0000000000860 \)
20. \( x: \ 0.0000000000287 \)
Choice of initialization $x^0$ is critical. Depending on the initialization, the Newton iteration might

- not converge (it could “blow up” or “oscillate” between two points)
- converge to different solutions
- fail cause it hits a point where the Jacobian is not invertible (this cannot happen if the conditions of the convergence theorem are satisfied)

...
Newton’s method
Convergence of Newton’s method

- The “more nonlinear” a problem, the harder it is to solve.

\[ \|F'(x)^{-1}(F'(x + sv) - F'(x))v\| \leq sw\|v\|^2 \]

- Computation of Jacobian \( F'(x^k) \) can be costly/complicated

(sometimes approximations of \( F'(x^k) \) are used)
Newton’s method
Convergence of Newton’s method

- There’s no reliable black-box solver for nonlinear problems; at least for higher-dimensional problems, the structure of the problem must be taken into account.

- Sometimes, continuation ideas must be used to find good initializations: Solve simpler problems first and use solution as starting point for harder problems.
Newton’s method
Robustification

Monotonicity test (affine invariant):

$$\| F'(x^k)^{-1} F(x^{k+1}) \| \leq \bar{\Theta} \| F'(x^k)^{-1} F(x^k) \|, \quad \bar{\Theta} < 1$$

Damping:

$$x^{k+1} = x^k + \lambda_k \Delta x^k, \quad 0 < \lambda_k \leq 1$$

For difficult problems, start with small $\lambda_k$ and increase later in the iteration (close to the solution $\lambda_k$ should be 1).

Approximative Jacobians: Use approximative Jacobians $\tilde{F}'(x^k)$, e.g., computed through finite differences.
“Classification of mathematical problems as linear and nonlinear is like classification of the Universe as bananas and non-bananas.”

or (according to Stanislav Ulam):

Using a term like nonlinear science is like referring to the bulk of zoology as the study of non-elephant animals.
Overview

Nonlinear least squares—Gauss-Newton
Assume a least squares problem, where the parameters $x$ do not enter linearly into the model. Instead of

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|^2,$$

we have with $F : D \to \mathbb{R}^n$, $D \subset \mathbb{R}^n$:

$$\min_{x \in \mathbb{R}^n} g(x) := \frac{1}{2} \|F(x)\|^2,$$

where $F(x)_i = \varphi(t_i, x) - b_i$, $1 \leq i \leq m$.

The (local) minimum $x^*$ of this optimization problem satisfies:

$$g'(x) = 0, \quad g''(x) \text{ is positive definite.}$$
Nonlinear least-squares problems

\[ g(x) = \frac{1}{2} \| F(x) \|^2 \]

The derivative of \( g(\cdot) \) is

\[ G(x) := g'(x) = F'(x)^T F(x) \]

This is a nonlinear system in \( x \), \( G : D \to \mathbb{R}^n \). Let’s try to solve it using Newton’s method:

\[ G'(x^k) \Delta x^k = -G(x^k), \quad x^{k+1} = x^k + \Delta x^k \]

where

\[ G''(x) = F'(x)^T F'(x) + F''(x^k)F(x). \]
Nonlinear least-squares problems: Example

\[(t_i, b_i) \quad i = 1, 2, 3 \text{ data points}\]

\[\varphi(t_i x_1 x_2) = e^{\varphi(x_i) t_i^2 + x_2^2 \sin(t_i)}\]

\[x = (x_1, x_2)\]

\[F(x) = \begin{bmatrix}
\varphi(t_1 x_1 x_2) - b_1 \\
\varphi(t_2 \ldots) - b_2 \\
\varphi(t_3 \ldots) - b_3
\end{bmatrix}\]

\[\min \frac{1}{2} \| F(x) \|^2 = \frac{1}{2} \left\| \begin{bmatrix}
2 e^{\varphi(x_1) t_1^2 + x_2^2 \sin(t_1)} - b_1 \\
-b_2 - b_2 \\
-b_3 - b_3
\end{bmatrix}\right\|^2\]
Nonlinear least-squares problems: Example

\[ G(x) = F(x)^{T} F(x) = \frac{\text{Jacobian}}{\in \mathbb{R}^{2 \times 3}} \]

Gauss–Newton method

\[ F(x^{k}) \Delta x = -F'(x^{k}) F(x^{k}) \]
\[ x^{k+1} = x^{k} + \Delta x \]

Linear least squares:

\[ A^{T} A x = -A^{T} b \]
Nonlinear least-squares problems

$F''(\mathbf{x})$ is a tensor. It is often neglected due to the following reasons:

- It's difficult to compute and we can use an approximate Jacobian in Newton's method.

- If the data is compatible with the model, then $F(\mathbf{x}^\star) = 0$ and the term involving $F''(\mathbf{x})$ drops out. If $\|F(\mathbf{x}^\star)\|$ is small, neglecting that term might not make the convergence much slower.

- We know that $g''(\mathbf{x}^\star)$ must be positive. If $F'(x^k)$ has full rank, then $F'(\mathbf{x})^T F'(\mathbf{x})$ is positive and invertible.
The resulting Newton method for the nonlinear least squares problem is called **Gauss-Newton method**: Initialize $x^0$ and for $k = 0, 1, \ldots$ solve

$$F'(x^k)^TF'(x^k)\Delta x^k = -F'(x^k)^TF(x^k) \quad \text{(solve)}$$

$$x^{k+1} = x^k + \Delta x^k. \quad \text{(update step)}$$
The resulting Newton method for the nonlinear least squares problem is called\textit{ Gauss-Newton method}: Initialize $x^0$ and for $k = 0, 1, \ldots$ solve

$$F'(x^k)^T F'(x^k) \Delta x^k = -F'(x^k)^T F(x^k)$$ \hspace{1cm} \text{(solve)}

$$x^{k+1} = x^k + \Delta x^k.$$ \hspace{1cm} \text{(update step)}

The solve step is the normal equation for the linear least squares problem

$$\min_{\Delta x} \| F'(x^k) \Delta x^k + F(x^k) \|.$$
Convergence of Gauss-Newton method

Assumptions on $F$: $D \subset \mathbb{R}^n$ open and convex, $F : D \rightarrow \mathbb{R}^m$, $m \geq n$ continuously differentiable with $F'(x)$ has full rank for all $x$, and let $\omega \geq 0$, $0 \leq \kappa^* < 1$ such that

$$
\|F'(x)^+(F'(x + sv) - F'(x))v\| \leq s\omega\|v\|^2
$$

for all $s \in [0, 1]$, $x \in D$, $v \in \mathbb{R}^n$ with $x + v \in D$.

Assumptions on $x^*$ and $x^0$: Assume there exists a solution $x^* \in D$ of the least squares problem and a starting point $x^0 \in D$ such that

$$
\|F'(x)^+F(x^*)\| \leq \kappa^*\|x - x^*\|
$$

$$
\rho := \|x^* - x^0\| \leq \frac{2(1 - \kappa^*)}{\omega} := \sigma
$$

Theorem: Then, the sequence $x^k$ stays in $B_\rho(x^*)$ and

$$
\lim_{k \to \infty} x^k = x^*, \text{ and}
$$

$$
\|x^{k+1} - x^*\| \leq \frac{\omega}{2}\|x^k - x^*\|^2 + \kappa^*\|x^k - x^*\|
$$
Convergence of Gauss-Newton method

- Role of $\kappa^*$: Represents omission of $F''(x)$
  - $\kappa^*$ can be chosen as 0 $\Rightarrow$ local quadratic convergence
  - $\kappa^* > 0$ linear convergence (thus we require $\kappa^* < 1$).

- Damping strategy as before (better: linesearch to make guaranteed progress in minimization problem)
- There can, in principle, be multiple solutions.