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 $F : \mathbb{R}^n \to \mathbb{R}^n$, $n \geq 1$ and solve

$$F(x) = 0.$$ 

Taylor expansion about starting point $x^0$:

$$F(x) = F(x^0) + F'(x^0)(x - x^0) + o(|x - x^0|) \quad \text{for} \ x \to x^0.$$

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.

Terminate iteration when $\|F(x^k)\| < \varepsilon \quad \text{or} \quad \frac{\|F(x^k)\|}{\|F(x)\|} < \varepsilon$
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 $G(x) = 0$:

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

$$= F'(x^k)^{-1} A^{-1} AF(x^k) = F'(x^k)^{-1} F(x^k) = \text{Newton step computed using } F$$
Newton's method

Intersection point between \( F(x, y) = y - x^2 + x = 0 \) parabola and \( F(x, y) = \frac{x^2}{16} + y^2 - 1 = 0 \) 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 iteration:

\( x^0 = (1, 0) \), solve

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

\( x' = x^0 + \Delta x \) 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

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

and

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

\[ x, y \in \mathcal{D} \]

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:

\[ ||F'(x)^{-1}\left[\int_0^1 \ldots \right]|| \leq \int_0^1 s \, ||y - x||^2 \, ds = \frac{w}{2} \, ||y - x||^2 \]

\[ \Rightarrow \left||\int_0^1 F'(x)^{-1}(F(y) - F(x) - F'(x)(y - x))d\sigma\right|| \leq \frac{w}{2} \, ||y - x||^2 \quad (1) \]

\[ \left||x^{k+1} - x^*\right|| = \left||x^k - F(x^k)^{-1}F(x^k) - x^*\right|| = 0 \]

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

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

\[ \Rightarrow \left||x^{k+1} - x^*\right|| \leq \frac{w}{2} \, \left||x^k - x^*\right||^2 \rightarrow \text{Quadratic Convergence} \]
Convergence of Newton’s method

\[ 0 < \|x^k - x^*\| \leq 8 \]

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

--- I’m staying in \( B_8(x^*) \) if \( x^k \) is in \( B_8(x^*) \)

Uniqueness: \( x^* \) is a solution,

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

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

\[ \rightarrow \quad x^k = x^* \]

\( \square \)
Newton’s method—when does convergence theorem apply?

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

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

| n   | x |  
|-----|---|---|
| 1   | 0.66666666666667 |
| 2   | 0.44444444444444 |
| 3   | 0.296296296296296 |
| ... |    |    |
| 17  | 0.001014959227 |
| 18  | 0.000676639485 |
| 19  | 0.000451092990 |
| 20  | 0.000300728660 |

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

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

| n   | x |  
|-----|---|---|
| 1   | 0.33333333333333 |
| 2   | 0.111111111111111 |
| 3   | 0.037037037037037 |
| 4   | 0.012345679012 |
| ... |    |    |
| 16  | 0.0000000023231 |
| 17  | 0.000000007744 |
| 18  | 0.000000002581 |
| 19  | 0.000000000860 |
| 20  | 0.000000000287 |
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.
Nonlinear least squares—Gauss-Newton
Assume a least squares problem, where the parameters $\mathbf{x}$ do not enter linearly into the model. Instead of

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

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

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

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

The (local) minimum $\mathbf{x}^*$ of this optimization problem satisfies:

$$g'(\mathbf{x}) = 0, \quad g''(\mathbf{x}) \text{ is positive definite.}$$
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

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

\((t_i, b_i)\) \(i = 1, 2, 3\) data points

\( \varphi(t_i x_i x_2) = \exp(x_i) t_i^2 + x_2^2 \sin(t_i) \)

\( x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \)

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

\[ \min \frac{1}{2} \| F(x) \|^2 = \frac{1}{2} \left\| \begin{bmatrix} \exp(x_1) t_1^2 + x_2^2 \sin(t_1) - b_1 \\ -t_2 - b_2 \\ -t_3 - b_3 - b_3 \end{bmatrix} \right\|^2 \]
Nonlinear least-squares problems: Example

\[ G(x) = F'(x)^T F(x) = \begin{bmatrix} \text{Jacobian} \\ \in \mathbb{R}^{2 \times 3} \end{bmatrix} \]

Gauss-Newton method

\[ F'(x^k)^T F(x^k) \Delta x = -F'(x^k)^T F(x^u) \]

\[ x^{u+1} = x^u + \Delta x \]

linear least squares: \[ A^T A x = -A^T b \]
Nonlinear least-squares problems

\( F''(x) \) 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 \( F(x^*) = 0 \) and the term involving \( F''(x) \) drops out. If \( \| F(x^*) \| \) is small, neglecting that term might not make the convergence much slower.
- We know that \( g''(x^*) \) must be positive. If \( F'(x^k) \) has full rank, then \( F'(x)^T F'(x) \) is positive and invertible.

We neglect \( F''(x) F(x) \) if compatible: \( \| F(x) \| \) is small.
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

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

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

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

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

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 $F$: $D \subseteq \mathbb{R}^n$ 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

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

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.
Why do outliers in data not matter when using the functional \( \sum_{i=1}^{m} |\Delta_i| \)?

Data fitting with a constant model \( q(t) = x \cdot 1 \)

Optimal model separates the points such that the same number of points is on both sides, so outliers have (almost) no influence.

If I minimize \( \sum_{i=1}^{m} \Delta_i^2 \), outliers have a big influence as we minimize the squares.