Numerical Methods I: Newton and nonlinear least squares

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

October 12, 2017

Overview

Newton's method to solve $F(\boldsymbol{x}) = \mathbf{0}$, $F : \mathbb{R}^n \to \mathbb{R}^n$

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

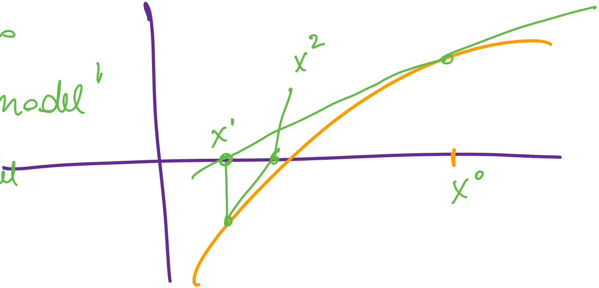
$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, \ldots$$

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

Initialization x^0

Compute 'linear model'

Solve linear model and iterate



Newton's method

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

$$F(\boldsymbol{x}) = 0.$$

Taylor expansion about starting point $\boldsymbol{x}^0$:

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

Hence:

Newton step / increment

$$\boldsymbol{x}^1 = \boldsymbol{x}^0 \overbrace{- F'(\boldsymbol{x}^0)^{-1} F(\boldsymbol{x}^0)}^{\Delta x}$$

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

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

Requires that $F'(\boldsymbol{x}^k) \in \mathbb{R}^{n \times n}$ is invertible.

Terminate iteration when $\|F(x^k)\| < \varepsilon$ or: $\dfrac{\|F(x^k)\|}{\|F(x^0)\|} < \tilde{\varepsilon}$

Newton's method

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

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

Equivalently:

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

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

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

$A \in \mathbb{R}^{n\times n}$, $\det(A)\neq 0$

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

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

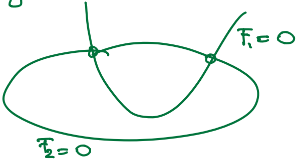
$$= F'(x^k)^{-1}\cancel{A^{-1}A}F(x^k) = F'(x^k)^{-1}F(x^k) = \text{Newton step computed using } F$$

Newton's method

Intersection point between $F_1(x,y) = y - x^2 + x = 0$ parabola

$F_2(x,y) = \frac{x^2}{16} + y^2 - 1 = 0$ ellipse

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



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

<u>Newton iterations</u>

$x^0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, Solve $\underbrace{\begin{pmatrix} -1 & 1 \\ \frac{1}{8} & 0 \end{pmatrix}}^{F'(x^0)} \Delta x = -F(x^0)$

$\longrightarrow x^1 = x^0 + \Delta x$ 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 $F'(\boldsymbol{x})$ is invertible for all $\boldsymbol{x}$, and there exists $\omega \geq 0$ such that

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

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

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

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

Theorem: Then, the Newton sequence $\boldsymbol{x}^k$ stays in $B_\rho(\boldsymbol{x}^*)$ and $\lim_{k\to\infty} \boldsymbol{x}^k = \boldsymbol{x}^*$, and

$$\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\| \leq \frac{\omega}{2}\|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2$$

Convergence of Newton's method

$x, y \in D$

Mean theorem for integrals:

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

take norms:

$$\left\| F'(x)^{-1}\left[\int \ldots\right]\right\| \le \int_{s=0}^{1} s\,\omega \|y-x\|^2\,ds = \frac{\omega}{2}\|y-x\|^2$$

$$\left\| F'(x)^{-1}\left(F(y) - F(x) - F'(x)(y-x)\right)\right\| \le \frac{\omega}{2}\|y-x\|^2 \qquad (1)$$

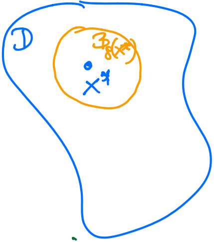
$$x^{k+1} - x^* = x^k - F'(x^k)^{-1}F(x^k) - x^*$$

$$= x^k - x^* - F'(x^k)^{-1}\left(F(x^k) - \underbrace{F(x^*)}_{=0}\right)$$

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

$$(1)+(2) \Longrightarrow \|x^{k+1} - x^*\| \le \frac{\omega}{2}\|x^k - x^*\|^2 \longrightarrow \text{quadratic convergence}$$

Convergence of Newton's method



$0 < \|x^k - x^*\| \le \rho$

$$\|x^{k+1} - x^*\| \le \underbrace{\frac{W}{2}\|x^k - x^*\|}_{\le \frac{W}{2}\rho \le 1} \|x^k - x^*\|$$

$\Longrightarrow$ I'm staying in $B_\rho(x^*)$ if x^k is in $B_\rho(x^*)$

<u>Uniqueness:</u> x^*, x^{**} solutions

$$\|x^* - x^{**}\| \le \underbrace{\frac{W}{2}\|x^* - x^{**}\|}_{<1} \|x^* - x^{**}\|$$

(from general result on previous page)

$$\Longrightarrow \|x^* - x^{**}\| < \|x^* - x^{**}\|$$

$$\longrightarrow x^* = x^{**}$$

$\square$.

Newton's method–when does convergence theorem apply?

- Example 1: $f(x) = x^3$

$x^* = 0$ solution

$f'(x^*) = 0$

```
1, x:  0.666666666667
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: $f(x) = x^{3/2}$

$x^* = 0$ solution

$f'(x^*) = 0$

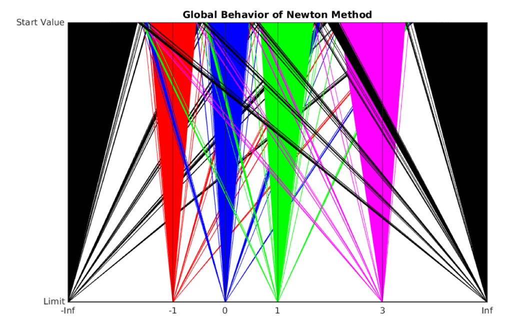
```
1, x:  0.333333333333
2, x:  0.111111111111
3, x:  0.037037037037
4, x:  0.012345679012
...
16, x:  0.000000023231
17, x:  0.000000007744
18, x:  0.000000002581
19, x:  0.000000000860
20, x:  0.000000000287
```

Newton's method

Role of initialization

Choice of initialization $\boldsymbol{x}^0$ 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)
- ...


Global Behavior of Newton Method
Start Value
Limit
-Inf
-1
0
1
3
Inf

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

very nonlinear $\Rightarrow F'(x)$ changes a lot

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

$\Rightarrow \omega$ large

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

(sometimes approximations of $F'(x^k)$ are used)

- 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.

Monotonicity test (affine invariant):

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

Damping:

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

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

Approximative Jacobians: Use approximative Jacobians $\tilde{F}'(\boldsymbol{x}^k)$, e.g., computed through finite differences.

Nonlinear versus linear problems

"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

Nonlinear least-squares problems

Assume a least squares problem, where the parameters $\boldsymbol{x}$ do *not* enter linearly into the model. Instead of

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

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

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} g(\boldsymbol{x}) := \frac{1}{2}\|F(\boldsymbol{x})\|^2, \quad \text{where } F(\boldsymbol{x})_i = \varphi(t_i,\boldsymbol{x}) - b_i, 1 \le i \le m$$

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

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

Nonlinear least-squares problems

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

The derivative of $g(\cdot)$ is

$$G(\boldsymbol{x}) := g'(\boldsymbol{x}) = F'(\boldsymbol{x})^T F(\boldsymbol{x})$$

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

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

where

$$G'(\boldsymbol{x}) = F'(\boldsymbol{x})^T F'(\boldsymbol{x}) + F''^T(\boldsymbol{x}) F(\boldsymbol{x}).$$

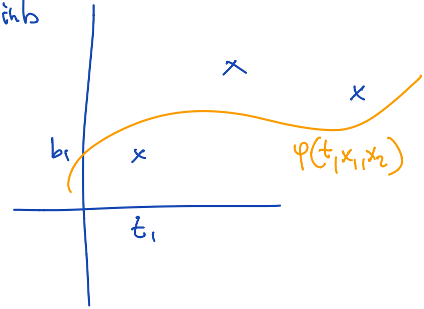
tensor

Nonlinear least-squares problems: Example

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

$$\varphi(t; x_1, x_2) = \exp(x_1)\, t^2 + x_2^2 \sin(t)$$

nonlinear in x



$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

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

$$\min_{x_1, x_2} \frac{1}{2} \| F(x) \|^2 = \frac{1}{2} \left\| \begin{bmatrix} \exp(x_1) t_1^2 + x_2^2 \sin(t_1) - b_1 \\ \text{—''—} \; t_2 \quad \text{—''—} \; t_2 - b_2 \\ \text{—''—} \; t_3 \quad \text{—''—} \; t_3 - b_3 \end{bmatrix} \right\|^2$$

Nonlinear least-squares problems: Example

$$G(x) = F'(x)^T F(x) = \boxed{\begin{matrix}\text{Jacobian} \\ \in \mathbb{R}^{2\times 3}\end{matrix}}$$

Gauss-Newton method

$$F'(x^k)^T F'(x^k)\, \Delta x = -F'(x^k)^T F(x^k)$$

$$x^{k+1} = x^k + \Delta x$$

linear least squares: $A^T A x = -A^T b$

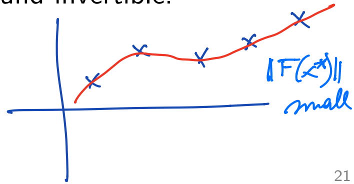
Nonlinear least-squares problems

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

We neglect $F''(x)\,F(x)$

compatible :



Nonlinear least-squares problems—Gauss-Newton

The resulting Newton method for the nonlinear least squares problem is called Gauss-Newton method: Initialize $\boldsymbol{x}^0$ and for $k = 0, 1, \ldots$ solve

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

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

Nonlinear least-squares problems—Gauss-Newton

The resulting Newton method for the nonlinear least squares problem is called Gauss-Newton method: Initialize $\boldsymbol{x}^0$ and for $k = 0, 1, \dots$ solve

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

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

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

$$\min_{\Delta\boldsymbol{x}} \|F'(\boldsymbol{x}^k)\Delta\boldsymbol{x}^k + F(\boldsymbol{x}^k)\|.$$

Convergence of Gauss-Newton method

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

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

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

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

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

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

Theorem: Then, the sequence $\boldsymbol{x}^k$ stays in $B_\rho(\boldsymbol{x}^*)$ and $\lim_{k\to\infty} \boldsymbol{x}^k = \boldsymbol{x}^*$, and

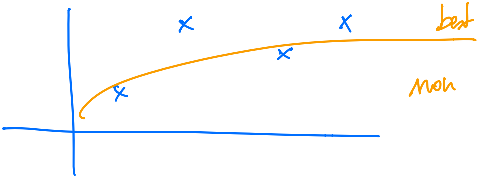
$$\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\| \leq \frac{\omega}{2}\|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 + \kappa^*\|\boldsymbol{x}^k - \boldsymbol{x}^*\|$$

Convergence of Gauss-Newton method

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

model curve can fit data exactly

$\kappa^* > 0$:

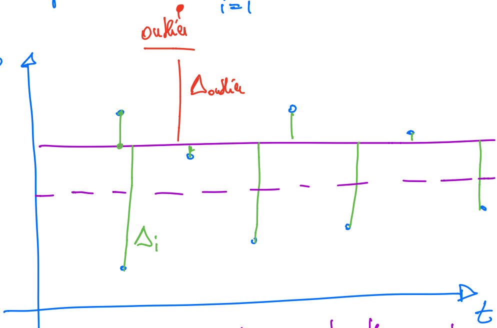


best model

non compatible with data

- 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 constant model $\varphi(t) = x \cdot 1$

optimal model seperates 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.