Convex Optimization
CMU-10725

Newton Method

Barnabás Póczos & Ryan Tibshirani

ML MACHINE LEARNING DEPARTMENT
Carnegie Mellon.
School of Computer Science

Administrivia

- Scribing
- Projects
- HW1 solutions
- Feedback about lectures / solutions on blackboard

Books to read

- **Boyd and Vandenberghe**: Convex Optimization, Chapters 9.5
- **Nesterov**: Introductory lectures on convex optimization
- **Bazaraa, Sherali, Shetty**: Nonlinear Programming
- **Dimitri P. Bestsekas**: Nonlinear Programming
- **Wikipedia**
- http://www.chiark.greenend.org.uk/~sgtatham/newton/

Goal of this lecture

Newton method

- ❑ Finding a root
- ❑ Unconstrained minimization
  - Motivation with quadratic approximation
  - Rate of Newton's method
- ❑ Newton fractals

Next lectures:

- ❑ Conjugate gradients

- ❑ Quasi Newton Methods

Newton method for finding a root

Newton method for finding a root

- Newton method: originally developed for finding a root of a function

- also known as the **Newton–Raphson method**

$$\phi : \mathbb{R} \to \mathbb{R}$$

$$\phi(x^*) = 0$$

$$x^* = ?$$

History

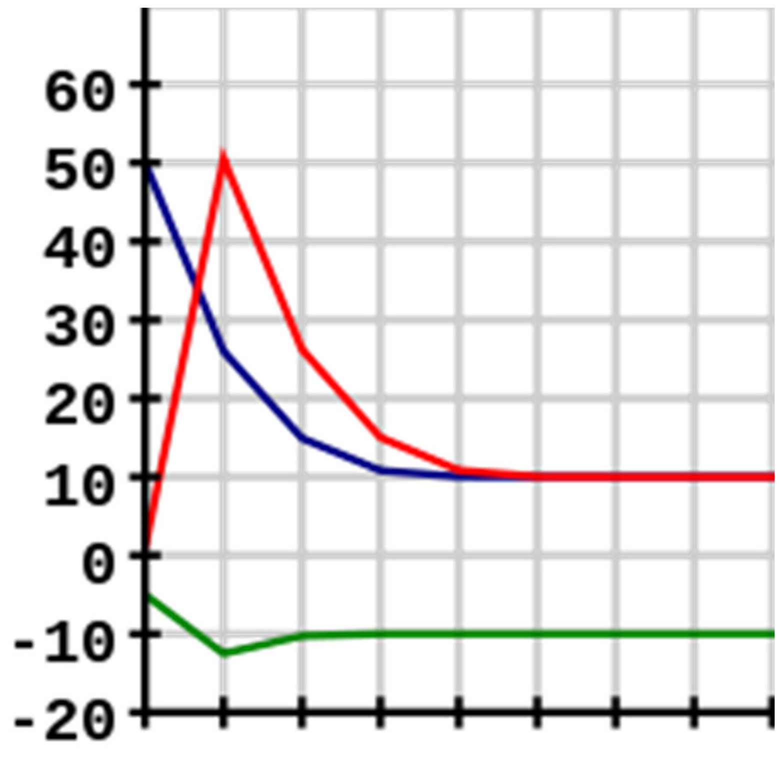
Finding $\sqrt{S}$ is the same as solving the equation:

$$f(x) = x^2 - S = 0$$

Babylonian method:

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{S}{x_n}\right) = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^2 - S}{2x_n}$$

This is a special case of Newton's method



S=100. starting values x_0 = 50, x_0 = 1, and x_0 = −5.

History

- **1690, Joseph Raphson:**

  finding roots of polynomials

- **1669, Isaac Newton:**

  finding roots of polynomials

- **1740, Thomas Simpson:**

  solving general nonlinear equation

  generalization to systems of two equations

  solving optimization problems (gradient = zero)

- **1879, Arthur Cayley:**

  generalizing the Newton's method to finding complex roots of polynomials

Newton Method for Finding a Root

Goal: $\phi : \mathbb{R} \to \mathbb{R}$

$$\phi(x^*) = 0$$

$$x^* = ?$$

Linear Approximation (1st order Taylor approx**):**

$$\underbrace{\phi(\underbrace{x + \Delta x}_{x^*})}_{\phi(x^*) = 0} = \phi(x) + \phi'(x)\Delta x + \underbrace{o(|\Delta x|)}_{\text{NEGLIGABLE}}$$

Therefore,

$$0 \approx \phi(x) + \phi'(x)\Delta x$$

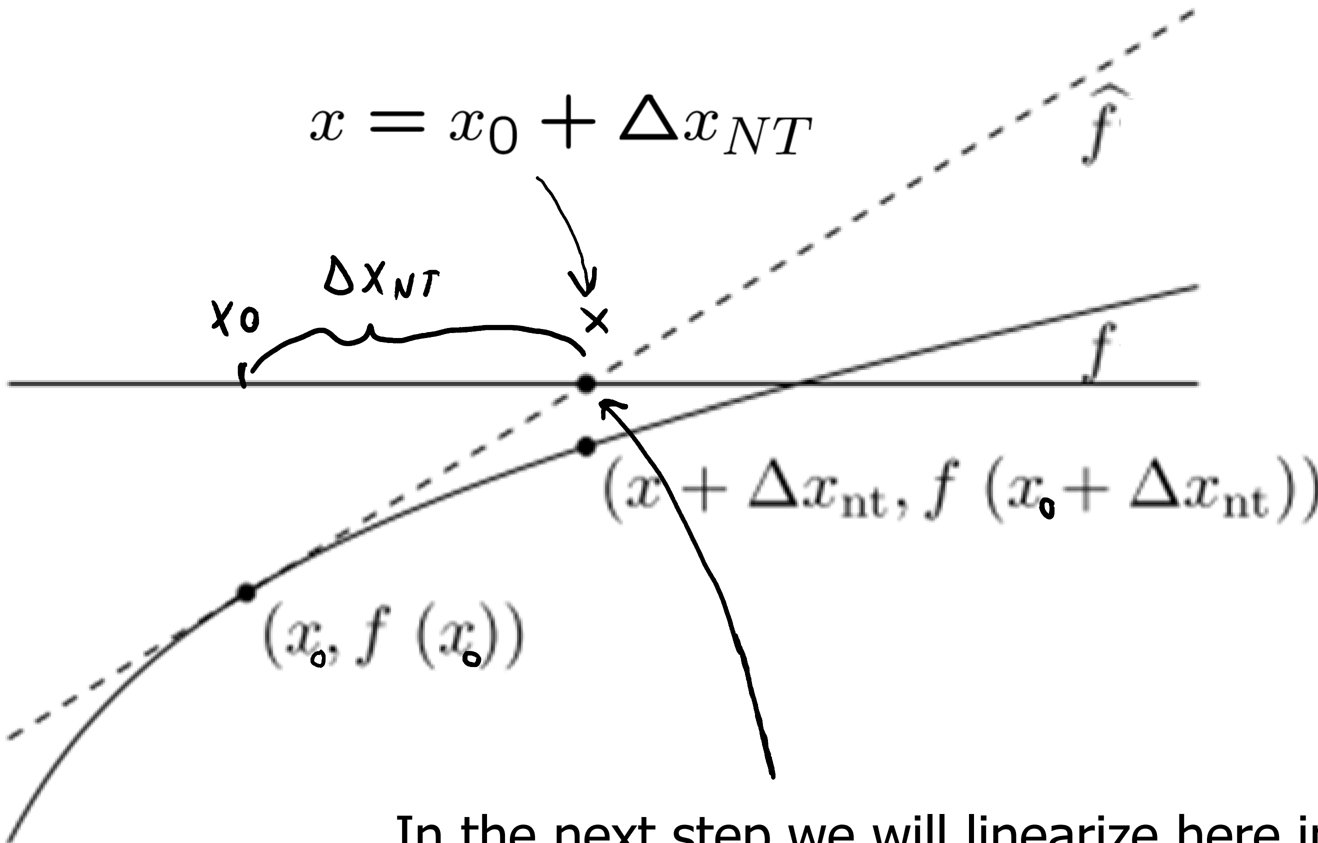
$$x^* - x = \Delta x = -\frac{\phi(x)}{\phi'(x)}$$

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

Illustration of Newton's method

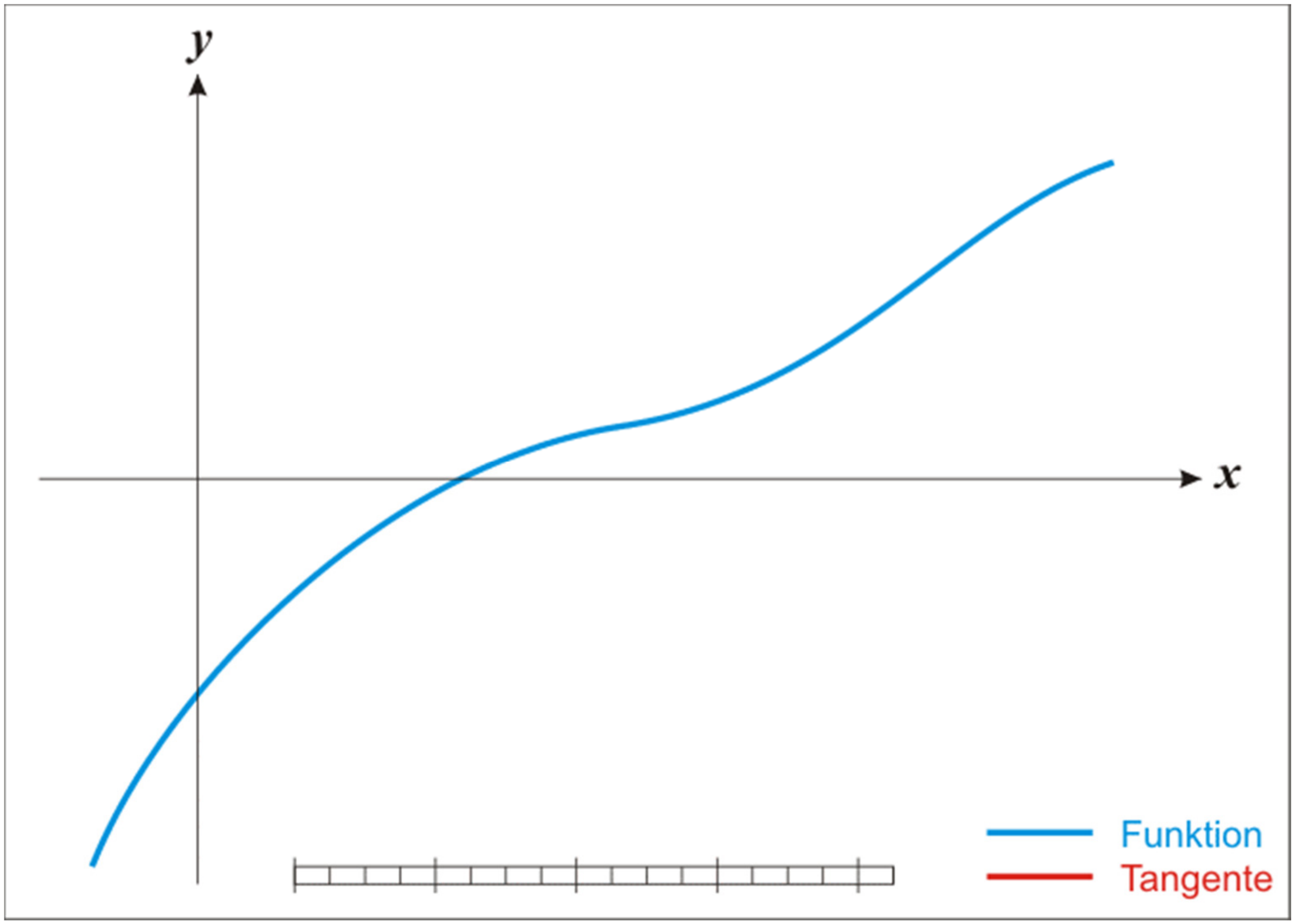
Goal: finding a root

$$\hat{f}(x) = f(x_0) + f'(x_0)(x - x_0)$$



In the next step we will linearize here in *x*

Example: Finding a Root



http://en.wikipedia.org/wiki/Newton%27s_method

Newton Method for Finding a Root

This can be generalized to multivariate functions

$$F : \mathbb{R}^n \to \mathbb{R}^m$$

$$0_m = F(x^*) = F(x + \Delta x) = F(x) + \underbrace{\nabla F(x)}_{\mathbb{R}^{m \times n}} \underbrace{\Delta x}_{\mathbb{R}^n} + \underset{\substack{\uparrow \\ \text{NEGLECT}}}{o(|\Delta x|)}$$

Therefore,

$$0_m = F(x) + \nabla F(x) \Delta x$$

$$\Delta x = -[\nabla F(x)]^{-1} F(x)$$

[Pseudo inverse if there is no inverse]

$\Delta x = x_{k+1} - x_k$, and thus

$$\underset{\mathbb{R}^n}{x_{k+1}} = \underset{\mathbb{R}^n}{x_k} - \underset{\mathbb{R}^{n \times m}}{[\nabla F(x_k)]^{-1}} \underset{\mathbb{R}^m}{F(x_k)}$$

Newton method: Start from x_0 and iterate.

Newton method for minimization

Newton method for minimization

$f : \mathbb{R}^n \to \mathbb{R}$, f is differentiable.

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

We need to find the roots of $\nabla f(x) = 0_n$

$$\nabla f : \mathbb{R}^n \to \mathbb{R}^n$$

Newton system: $\nabla f(x) + \nabla^2 f(x) \Delta x = 0_n$

Newton step: $\Delta x = x_{k+1} - x_k = -[\nabla^2 f(x)]^{-1} \nabla f(x)$

Iterate until convergence, or max number of iterations exceeded

(divergence, loops, division by zero might happen…)

How good is the Newton method?

Descent direction

Lemma [Descent direction**]**

If $\nabla^2 f \succ 0$, then Newton step is a descent direction.

Proof:

We know that if a vector has negative inner product with the gradient vector, then that direction is a descent direction

$$\text{Newton step: } \Delta x = x_{k+1} - x_k = -[\nabla^2 f(x)]^{-1} \nabla f(x)$$

$$\Rightarrow \nabla f(x)^T \Delta x = -\nabla f(x)^T [\nabla^2 f(x)]^{-1} \nabla f(x) < 0$$

Newton method properties

- Quadratic convergence in the neighborhood of a strict local minimum [under some conditions].
- It can break down if $f''(x_k)$ is degenerate.
  [no inverse]
- It can diverge.
- It can be trapped in a loop.
- It can converge to a loop…

Motivation with Quadratic Approximation

Motivation with Quadratic Approximation

$f : \mathbb{R}^n \to \mathbb{R}$, f is differentiable.

Second order Taylor approximation:

Let $\phi(x) = f(x_k) + \nabla^T f(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k)$

Assume that

$\nabla^2 f(x_k) \succ 0$ [i.e. ϕ has strict global minimum]

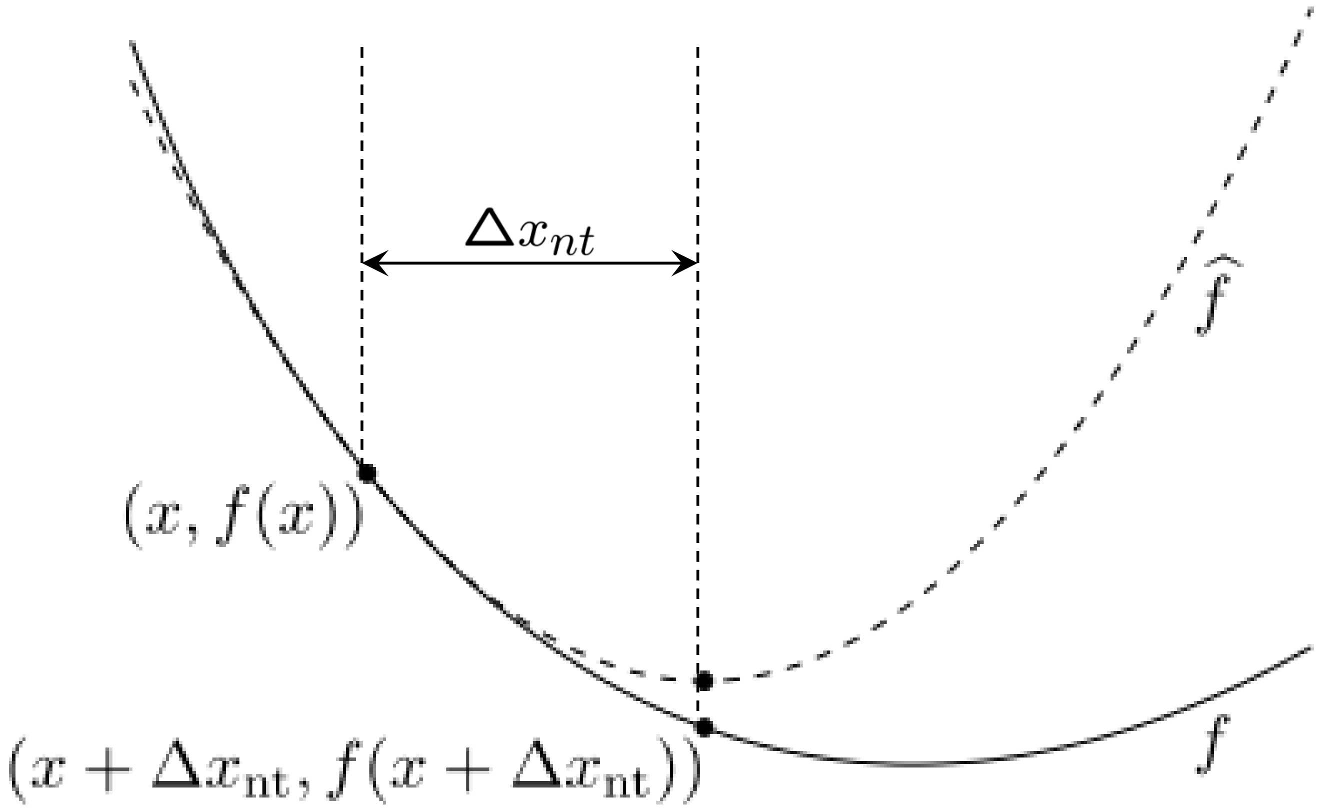
Now, if x_{k+1} is the global minimum of the quadratic function ϕ, then

$$0_n = \nabla\phi(x_{k+1}) = \nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k)$$

Newton step:

$$\Delta x = x_{k+1} - x_k = -[\nabla^2 f(x)]^{-1} \nabla f(x)$$

Motivation with Quadratic Approximation



Quadratic approaximation is good, when x is close to x^*

$$\widehat{f}(z) = f(x) + \nabla^T f(x)(z - x) + \frac{1}{2}(z - x)^T \nabla^2 f(x)(z - x)$$

Convergence rate (f: R$\rightarrow$ R case)

Rates

A sequence $\{s_i\}$ exhibits linear convergence if $\lim_{i\to\infty} s_i = \bar{s}$, and

$$\lim_{i\to\infty} \frac{|s_{i+1} - \bar{s}|}{|s_i - \bar{s}|} = \delta < 1$$

Example: $s_i = cq^i$, $0 < q < 1$

$$\frac{|s_{i+1} - \bar{s}|}{|s_i - \bar{s}|} = \frac{cq^{i+1}}{cq^i} = q < 1$$

Superlinear rate: $\delta = 0$ Example: $s_i = \dfrac{c}{i!}$

$$\frac{|s_{i+1} - \bar{s}|}{|s_i - \bar{s}|} = \frac{c\,i!}{c(i+1)!} = \frac{1}{i+1} \to 0$$

Sublinear rate: $\delta = 1$ Example: $s_i = \frac{c}{i^a}$, $a > 0$

$$\frac{|s_{i+1} - \bar{s}|}{|s_i - \bar{s}|} = \frac{c\,i^a}{c(i+1)^a} = \left(\frac{i}{i+1}\right)^a \to 1$$

Quadratic rate:

$$\lim_{i\to\infty} \frac{|s_{i+1} - \bar{s}|}{|s_i - \bar{s}|^2} < \infty$$

Example: $s_i = q^{2^i}$, $0 < q < 1$

Finding a root: Convergence rate

Goal: Find x^* s.t. $f(x^*) = 0$, where $f : \mathbb{R} \to \mathbb{R}$

Assumption: f has continuous second derviative in x^*

Taylor theorem: For a ξ_n between x_n and x^*, we have

$$0 = f(x^*) = f(x_n) + \nabla f(x_n)(x^* - x_n) + \frac{1}{2}\nabla^2 f(\xi_n)(x - x_n)^2$$

Therefore, assuming $\exists[\nabla f(x_n)]^{-1}$

$$0 = [\nabla f(x_n)]^{-1} f(x_n) + (x^* - x_n) + \frac{1}{2}[\nabla f(x_n)]^{-1}\nabla^2 f(\xi_n)(x - x_n)^2$$

$$\underbrace{[\nabla f(x_n)]^{-1} f(x_n) + (x^* - x_n)}_{\underbrace{x^* - x_{n+1}}_{\varepsilon_{n+1}}} = -\frac{1}{2}[\nabla f(x_n)]^{-1}\nabla^2 f(\xi_n)\underbrace{(x - x_n)^2}_{\varepsilon_n^2}$$

$$\Rightarrow \epsilon_{n+1} = -\tfrac{1}{2}[\nabla f(x_n)]^{-1}\nabla^2 f(\xi_n)\epsilon_n^2$$

Finding a root: Convergence rate

We have seen that

$$\epsilon_{n+1} = -\frac{1}{2}\frac{\nabla^2 f(\xi_n)}{\nabla f(x_n)}\epsilon_n^2$$

Assume that $M = \sup_x \frac{1}{2}\frac{|\nabla^2 f(\mathbf{x})|}{|\nabla f(x)|} < \infty$

$$\Rightarrow |\epsilon_{n+1}| \leq M\epsilon_n^2$$

Assume that $|\epsilon_0| = |x - x_0| < 1$

$\Rightarrow$ Quadratic convergence

Problematic cases

Finding a root: chaotic behavior

Let f(x)=x^3-2x^2-11x+12

Goal: find the roots, (-3, 1, 4), using Newton's method

2.35287527 converges to 4;

2.35284172 converges to −3;

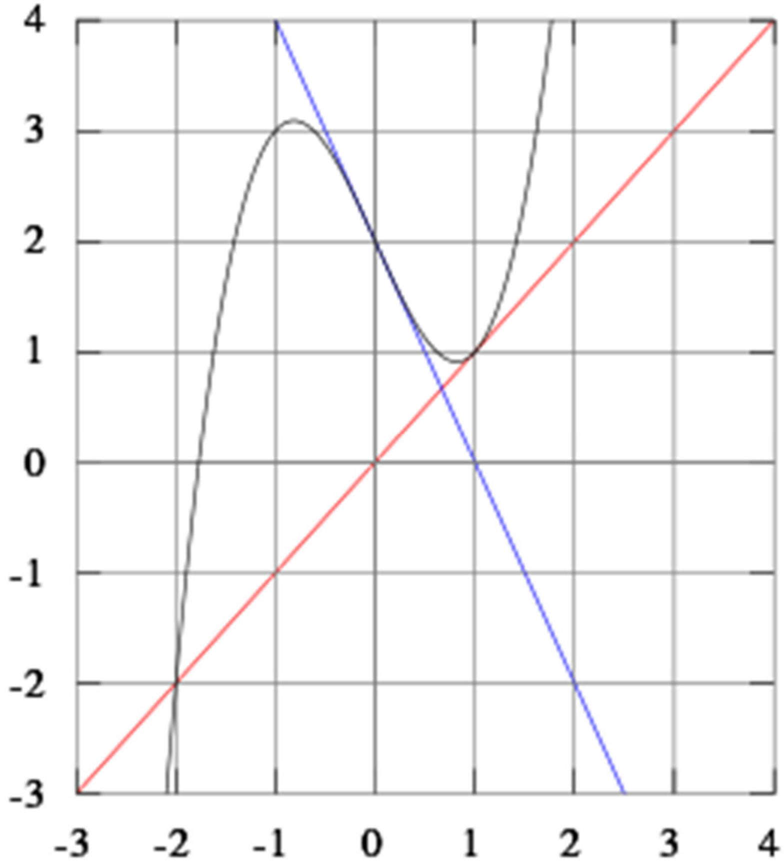
2.35283735 converges to 4;

2.352836327 converges to −3;

2.352836323 converges to 1.

Finding a root: Cycle

$f(x) = x^3 - 2x + 2$ **Goal**: find its roots



$x_0 = 0$ $X_1 = 1$, $X_2 = 0$, $X_3 = 1$, $X_4 = 0 \ldots \Rightarrow$ 2-CYCLE

Stating point is important!

Finding a root: divergence everywhere (except in the root)

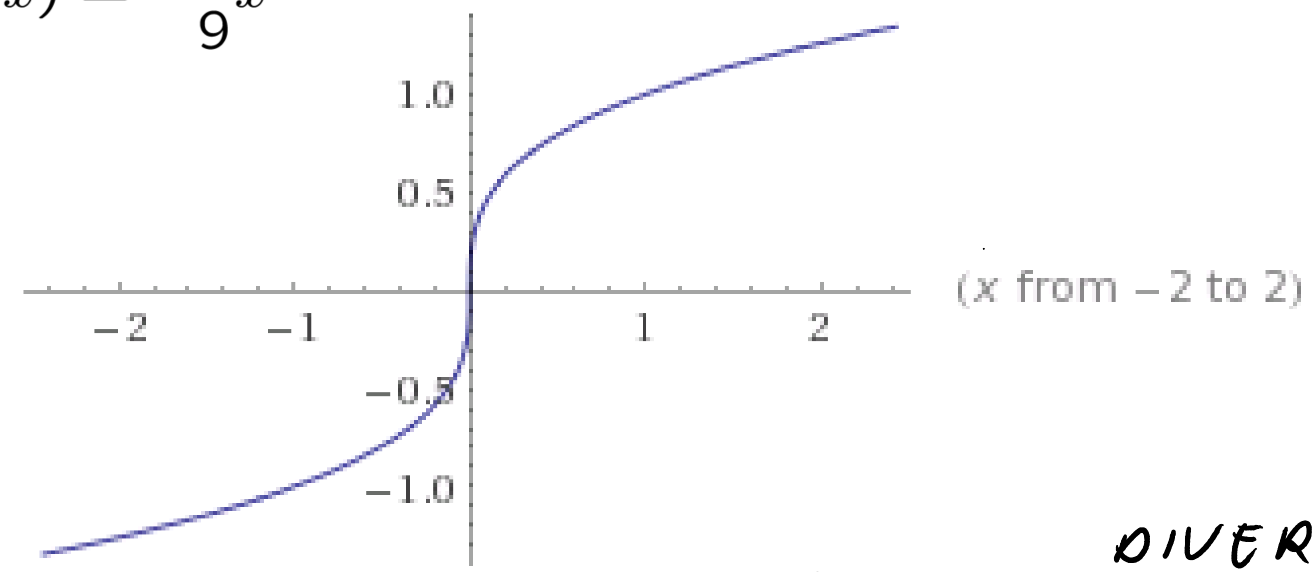
Newton's method might never converge (except in the root)!

$$f(x) = \sqrt[3]{x}$$

$$\nabla f(x) = \frac{1}{3}x^{-2/3}$$

$$\nabla^2 f(x) = -\frac{2}{9}x^{-5/3}$$

$$\lim_{x \to 0} \frac{1}{2}\frac{|\nabla^2 f(\mathbf{x})|}{|\nabla f(x)|} = \lim_{x \to 0} \frac{c}{|x|} = \infty$$



DIVERGENCE!

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^{\frac{1}{3}}}{\frac{1}{3}\, x_n^{\frac{1}{3}-1}} = x_n - 3\, x_n = -2\, x_n$$

Finding a root: Linear convergence only

If the first derivative is zero at the root, then convergence might be only linear (not quadratic)

$$f(x) = x^2$$

$$\nabla f(x) = 2x$$

$$\nabla^2 f(x) = 2$$

$$\lim_{x \to 0} \frac{1}{2} \frac{|\nabla^2 f(\mathbf{x})|}{|\nabla f(x)|} = \lim_{x \to 0} \frac{1}{|x|} = \infty$$

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - x_n^2/(2x_n) = x_n/2$$

Linear convergence only!

Difficulties in minimization

$$f(x) = 7x - \ln(x)$$

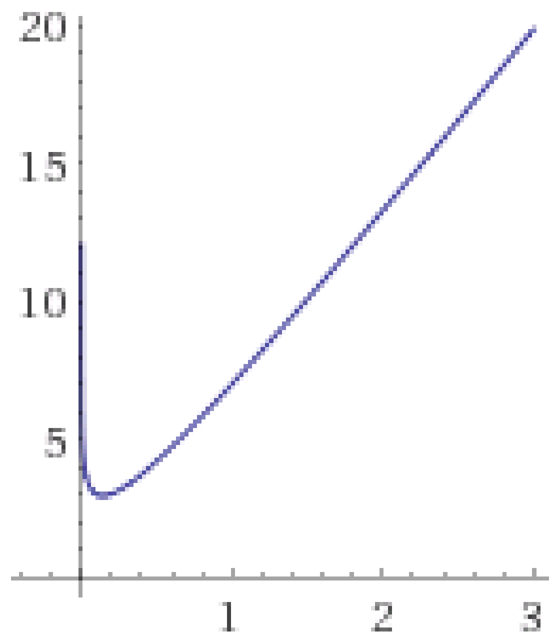
$$x^* = \frac{1}{7} = 0.142857143$$

$$f'(x) = 7 - \frac{1}{x}$$

$$f''(x) = \frac{1}{x^2}$$

$$x^{k+1} = x^k + (x^k - 7(x^k)^2) = 2x^k - 7(x^k)^2$$

k	x^k	x^k	x^k	x^k
0	1.0	0	0.1	0.01
1	-5.0	0	0.13	0.0193
2	-185.0	0	0.1417	0.03599257
3	$-239,945.0$	0	0.14284777	0.062916884
4	-4.0302×10^{11}	0	0.142857142	0.098124028
5	-1.1370×10^{24}	0	0.142857143	0.128849782
6	-9.0486×10^{48}	0	0.142857143	0.1414837
7	-5.7314×10^{98}	0	0.142857143	0.142843938
8	$-\infty$	0	0.142857143	0.142857142
9	$-\infty$	0	0.142857143	0.142857143
10	$-\infty$	0	0.142857143	0.142857143



range of quadratic convergence

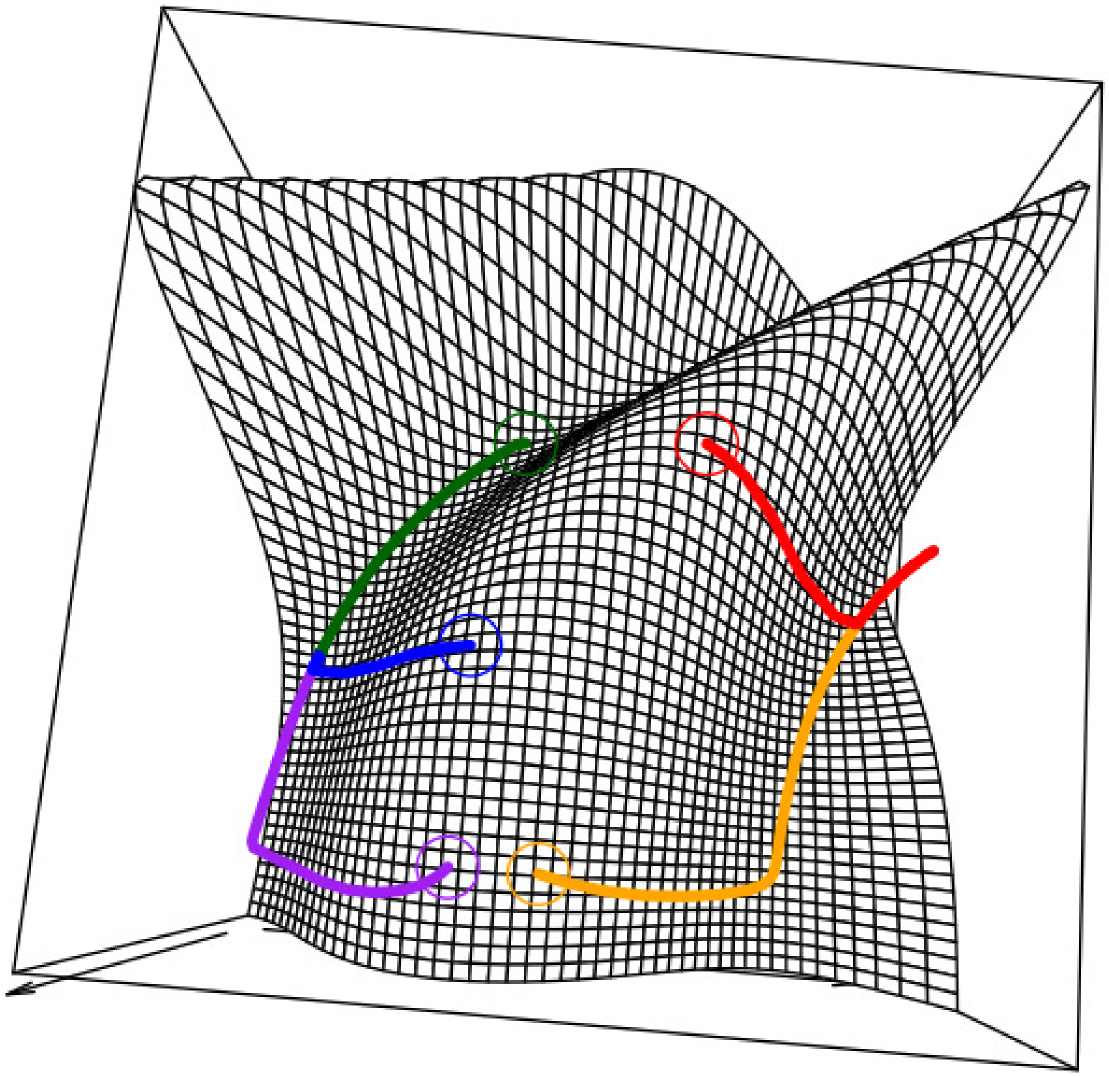
$x \in (0.0\ ,\ 0.2857143)$

Generalizations

- **Newton method in Banach spaces**
  - Newton method on the Banach space of functions
  - We need Frechet derivatives
- **Newton method on curved manifolds**
  - E.g. on space of otrthonormal matrices
- **Newton method on complex numbers**

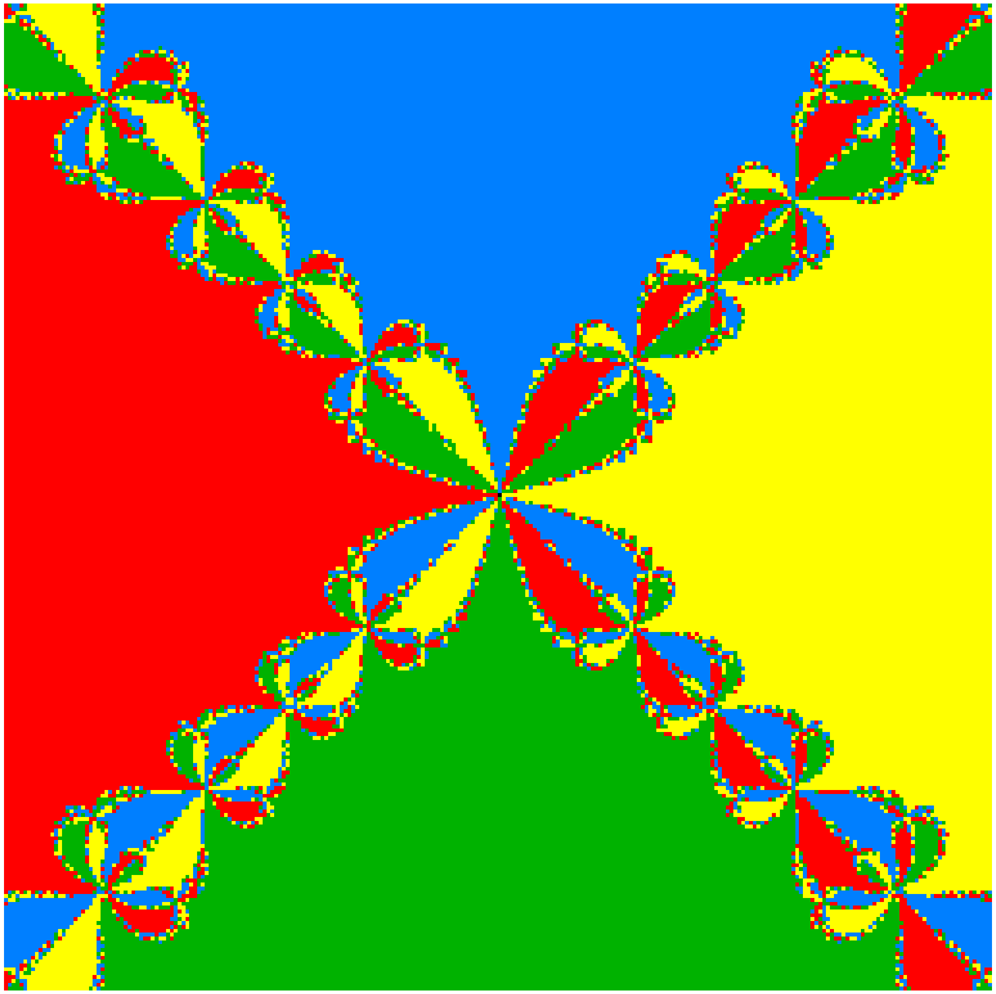
Newton Fractals

Gradient descent



Compex functions

$f(z) = z^4 - 1$, Roots: -1, +1, -i, +i

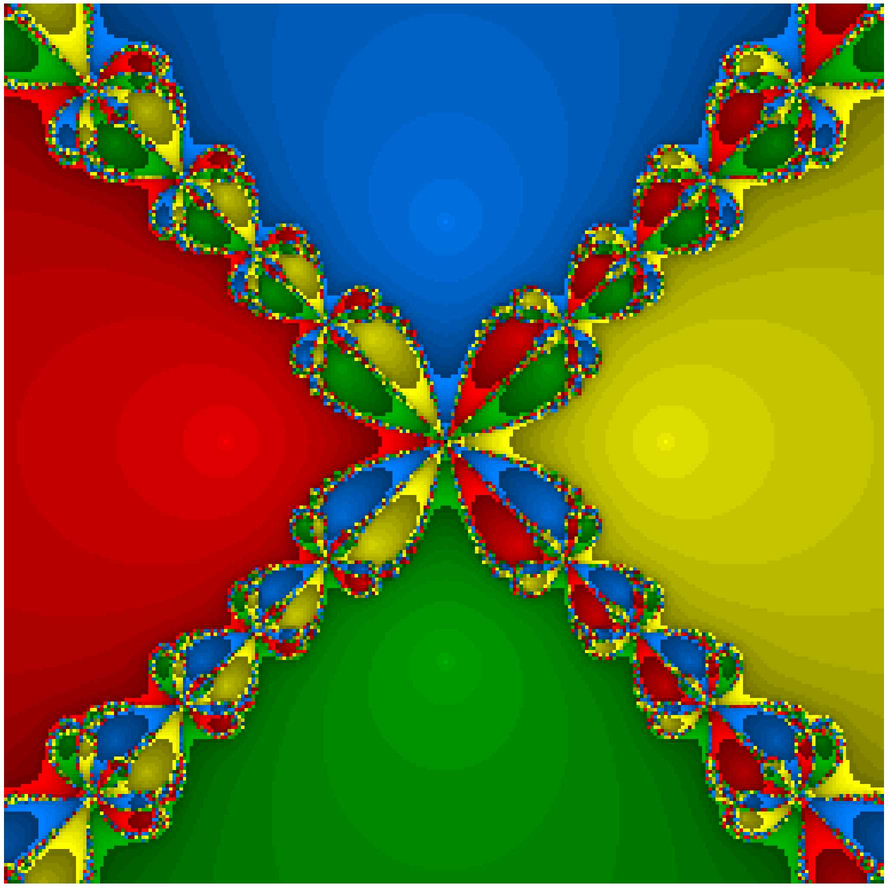


color the starting point according to *which* root it ended up

http://www.chiark.greenend.org.uk/~sgtatham/newton/

Basins of attraction

$f(z)= z^4-1$, Roots: -1, +1, -i, +i

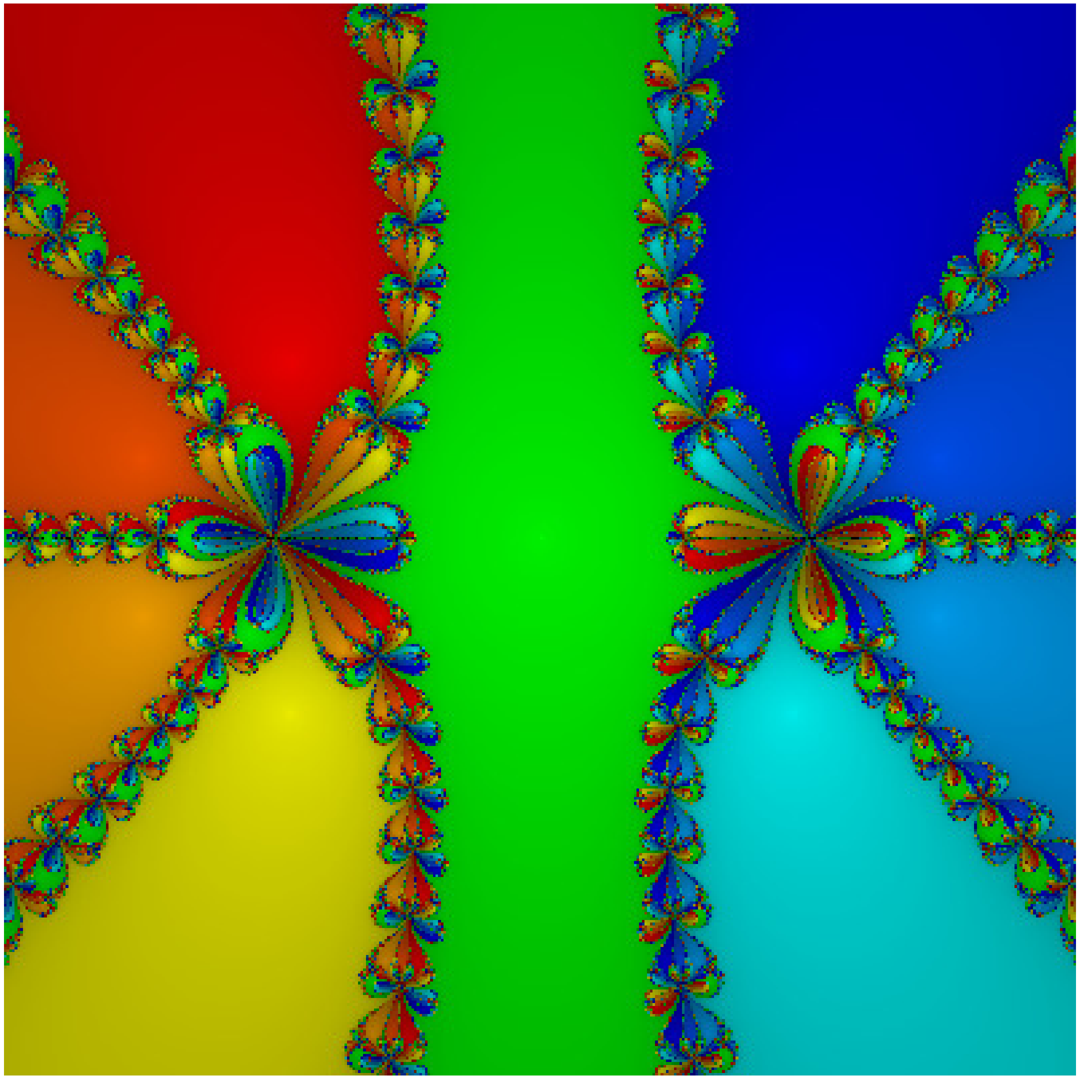


Shading according to how many iterations it needed till convergence

http://www.chiark.greenend.org.uk/~sgtatham/newton/

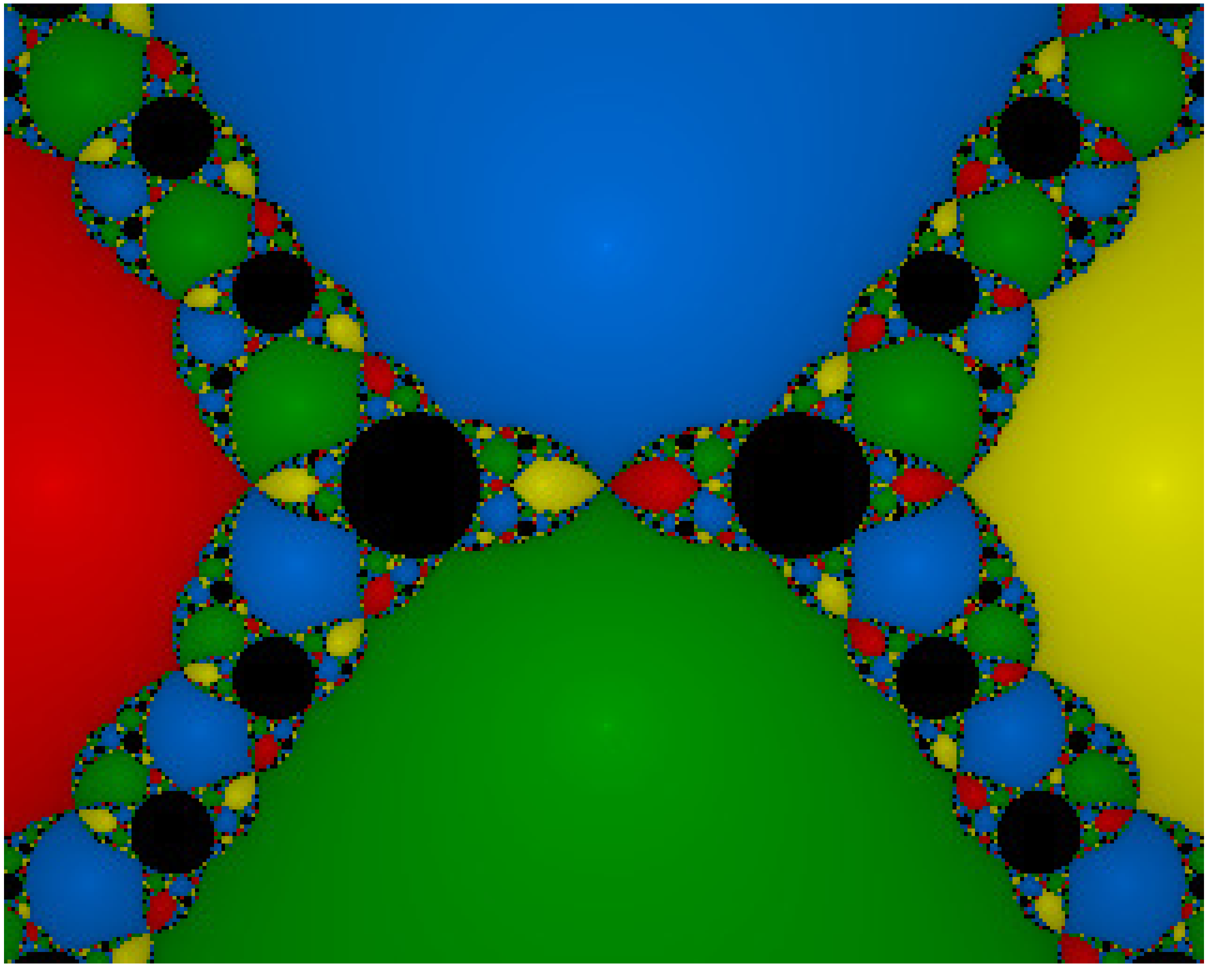
Basins of attraction

$f'(z)=(z-1)^4(z+1)^4$



No convergence

polynomial f, having the roots at +i, -i, -2.3 and +2.3



Black circles: no convergence, attracting cycle with period 2

Avoiding divergence

Damped Newton method

In order to avoid possible divergence
do line search with back tracking

$$x_{k+1} = x_k - h_k[f''(x_k)]^{-1}f'(x_k)$$

INITIAL STAGE: LINE SEARCH FOR $h_k > 0$

FINAL STAGE: $h_k = 1$ [NEWTON STEP]

We already know that the Newton direction is descent direction

Classes of differentiable functions

Classes of Differentiable functions

- Any continuous (but otherwise ugly, nondifferentiable) function can be approximated arbitrarily well with smooth (k times differentiable) function

- Assuming smoothness only is not enough to talk about rates…

- We need stronger assumptions on the derivatives. [e.g. their magnitudes behaves nicely]

The $C_L^{k,p}(Q)$ Class

Definition

$$C_L^{k,p}(Q) = \{f : Q \to \mathbb{R}$$

f is k-times continuously differentiable on $Q \subseteq \mathbb{R}^n$

$k \geq p$

$\|f^{(p)}(x) - f^{(p)}(y)\| \leq L\|x - y\|, \ \forall x, y \in Q$

$\}$ [Lipschitz continuous pth order derivative]

Notation

$$C^k(Q) = \{f : Q \to \mathbb{R}$$

f is k-times continuously differentiable on $Q \subseteq \mathbb{R}^n$ $\}$

Trivial Lemmas

Lemma [Linear combinations**]**

$$\left.\begin{array}{c} f_1 \in C^{k,p}_{L_1}(Q) \\ f_2 \in C^{k,p}_{L_2}(Q) \\ \alpha, \beta \in \mathbb{R} \\ L_3 = |\alpha| L_1 + |\beta| L_2 \end{array}\right\} \Rightarrow \alpha f_1 + \beta f_2 \in C^{k,p}_{L_3}(Q)$$

Lemma [Class hierarchy**]**

If $q \geq r$, then $C^{q,p}_L(Q) \subseteq C^{r,p}_L(Q)$

E.G. $C^{2,1}_L(Q) \subseteq C^{1,1}_L(Q)$

Relation between 1st and 2nd derivatives

Lemma Let f be twice differentiable on $\mathbb{R}^n$

$$f'(x+\alpha(y-x)) - f'(x) = \int_0^\alpha f''(x+\tau(y-x))(y-x)\,d\tau$$

Proof Let $\phi(\tau) = f'(x+\tau(y-x))$. Now we have that

$$\phi(0) = f'(x)$$
$$\phi(\alpha) = f'(x+\alpha(y-x))$$
$$\phi'(\tau) = f''(x+\tau(y-x))(y-x)$$

Therefore,

$$\begin{aligned} f'(x+\alpha(y-x)) - f'(x) &= \phi(\alpha) - \phi(0) \\ &= \int_0^\alpha \phi'(\tau)\,d\tau \\ &= \int_0^\alpha f''(x+\tau(y-x))(y-x)\,d\tau \end{aligned}$$

Q.E.D

Special case,

$$f'(y) - f'(x) = \int_0^1 f''(x+\tau(y-x))(y-x)\,d\tau$$

$C_L^{2,1}(R^n)$ and the norm of f''

Lemma [$C_L^{2,1}$ and the norm of f'']

Let f be twice differentiable on $\mathbb{R}^n$

Then $\|f''(x)\|_{op} \leq L \, \forall x \in \mathbb{R}^n \Leftrightarrow f \in C_L^{2,1}(\mathbb{R}^n)$

Proof

$\Rightarrow$

$$f'(y) - f'(x) = \int_0^1 f''(x + \tau(y-x))(y-x)\, d\tau$$

$$\Rightarrow \quad \|f'(y) - f'(x)\| \leq \int_0^1 \|f''(x + \tau(y-x))(y-x)\|\, d\tau$$

$$\leq \int_0^1 \|f''(x + \tau(y-x))\|_{op}\, d\tau \, \|(y-x)\|$$

$$\leq \|(y-x)\| \int_0^1 L\, d\tau = L\|(y-x)\|$$

$$\Rightarrow f \in C_L^{2,1}(\mathbb{R}^n)$$

Q.E.D

$C_L^{2,1}(R^n)$ and the norm of f''

$$\|f''(x)\|_{op} \le L \, \forall x \in \mathbb{R}^n \Leftrightarrow f \in C_L^{2,1}(\mathbb{R}^n)$$

Proving the other direction $\Leftarrow$

$$f \in C_L^{2,1}(\mathbb{R}^n) \Rightarrow$$

With $s = y - x$, we have seen that

$$f'(x + \alpha s) - f'(x) = \int_0^\alpha f''(x + \tau s)\, d\tau s$$

$$\Rightarrow \quad \| \int_0^\alpha f''(x + \tau s)\, d\tau s\| = \|f'(x + \alpha s) - f'(x)\| \le L\|\alpha s\|$$

$$\Rightarrow \quad \frac{1}{\alpha}\| \int_0^\alpha f''(x + \tau s)\, d\tau s\| \le L\|s\| \qquad f \in C_L^{2,1}(\mathbb{R}^n) \quad \forall s \in \mathbb{R}^n \quad \forall \alpha > 0$$

$\alpha \to 0$ $\|f''(x) \cdot s\|$

$$\Rightarrow \quad \frac{\|f''(x)s\|}{\|s\|} \le L$$

$$\Rightarrow \quad \|f''(x)\|_{op} \le L$$

Q.E.D

Examples

- $f(x) = \alpha + \langle a, x \rangle \in C_0^{1,1}(\mathbb{R}^n)$
- $f(x) = \alpha + \langle a, x \rangle + \frac{1}{2}\langle Ax, x \rangle \in C_L^{1,1}(\mathbb{R}^n)$
  WITH $L = \|A\|$
- $f(x) = \sqrt{1+x^2} \in C_1^{1,1}(\mathbb{R})$

Error of 1st order Taylor approx. in $C_L^{1,1}$

Lemma [1st orderTaylor approximation in $C_L^{1,1}$**]**

$$\left.\begin{array}{l} f \in C_L^{1,1}(\mathbb{R}^n) \\ x, y \in \mathbb{R}^n \end{array}\right\} \Rightarrow |f(y) - f(x) - \langle f'(x), y - x\rangle| \le \frac{L}{2}\|y - x\|^2$$

Proof

$$f(y) = f(x) + \int_0^1 \langle f'(x + \tau(y - x)), y - x\rangle \, d\tau$$

$$\Rightarrow |f(y) - f(x) - \langle f'(x), y - x\rangle| = \left|\int_0^1 \langle f'(x + \tau(y - x)), y - x\rangle \, d\tau - \langle f'(x), y - x\rangle\right|$$

$$= \left|\int_0^1 \langle f'(x + \tau(y - x)) - f'(x), y - x\rangle \, d\tau\right|$$

$$\le \int_0^1 |\langle f'(x + \tau(y - x)) - f'(x), y - x\rangle| \, d\tau$$

Error of 1st order Taylor approx. in $C_L^{1,1}$

$$\Rightarrow |f(y) - f(x) - \langle f'(x), y - x\rangle| \le \int_0^1 |\langle f'(x + \tau(y - x)) - f'(x), y - x\rangle| \, d\tau$$

$$\le \int_0^1 \|f'(x + \tau(y - x)) - f'(x)\| \, \|y - x\| \, d\tau$$

$\overset{\wedge}{\le} L\|\tau(y-x)\|$ $\quad f \in C_L^{1,1}(\mathbb{R}^n)$

$$\le \int_0^1 L\tau\|y - x\|^2 \, d\tau = \frac{L}{2}\|y - x\|^2$$

Q.E.D

Sandwiching with quadratic functions in $C_L^{1,1}$

We have proved:

$$f \in C_L^{1,1}(\mathbb{R}^n) \;\Rightarrow\; |f(y) - f(x) - \langle f'(x), y - x\rangle| \le \frac{L}{2}\|y - x\|^2$$
$$x, y \in \mathbb{R}^n$$

Corollary [Sandwiching $C_L^{1,1}$ functions with quadratic functions**]**

$$f \in C_L^{1,1}(\mathbb{R}^n) \;\Rightarrow\; f(x_0) + \langle f'(x_0), x - x_0\rangle - \frac{L}{2}\|x - x_0\|^2 \le f(x)$$
$$f(x) \le f(x_0) + \langle f'(x_0), x - x_0\rangle + \frac{L}{2}\|x - x_0\|^2$$

Function *f* can be lower and upper bounded with quadratic functions

$C_L^{2,2}(R^n)$ Class

Lemma [Properties of $C_L^{2,2}$ Class**]**

$$\left.\begin{array}{l} f \in C_L^{2,2}(\mathbb{R}^n) \\ x, y \in \mathbb{R}^n \end{array}\right\} \Rightarrow$$

$$\|f''(x) - f''(y)\|_{op} \leq L\|y - x\| \quad (*1)$$

$$\|f'(y) - f'(x) - f''(x)(y - x)\| \leq \tfrac{L}{2}\|y - x\|^2 \quad (*2)$$

[Error of the 1st order approximation of f’]

$$\|f(y) - f(x) - f'(x)^T(y - x) - \frac{1}{2}(y - x)^T f''(x)(y - x)\| \leq \frac{L}{6}\|y - x\|^3 \quad (*3)$$

[Error of the 2nd order approximation of f]

Proof (*1) Definition

(*2) Same as previous lemma [f’ instead of f]

(*3) Similar [Homework]

Sandwiching f''(y) in $C_L^{2,2}(R^n)$

By definition

$$\left.\begin{array}{l} f \in C_L^{2,2}(\mathbb{R}^n) \\ x, y \in \mathbb{R}^n \end{array}\right\} \Longrightarrow \|f''(x) - f''(y)\|_{op} \leq L\|y - x\| \qquad (*1)$$

Corollary [Sandwiching *f''(y)* matrix**]**

$$\left.\begin{array}{l} f \in C_L^{2,2}(\mathbb{R}^n) \\ \|x - y\| = r \end{array}\right\} \Longrightarrow f''(x) - Lr\mathbf{I}_n \preceq f''(y) \preceq f''(x) + Lr\mathbf{I}_n$$

Proof

$$f \in C_L^{2,2}(\mathbb{R}^n) \Longrightarrow \|\overbrace{f''(x) - f''(y)}^{G}\|_{op} \leq L\|y - x\| = Lr$$

$$\Longrightarrow |\lambda_i(G)| \leq Lr \;\; \forall i = 1, \ldots, n$$

$$\Longrightarrow \left\{\begin{array}{l} f''(x) - f''(y) = G \preceq Lr\mathbf{I}_n \\ f''(y) - f''(x) = -G \preceq Lr\mathbf{I}_n \end{array}\right.$$

Q.E.D

Convergence rate of Newton’s method

Convergence rate of Newton's method

Assumptions

$f \in C_L^{2,2}(\mathbb{R}^n)$

$\exists$ local minimum x^* of f with pos def Hessian in x^*:

$$f''(x^*) \succeq l\mathbf{I}_n \text{ for some } l > 0$$

x_0 is close enough to x^* [Local convergence only]

Newton step: $x_{k+1} = x_k - [f''(x_k)]^{-1} f'(x_k)$

Convergence rate of Newton's method

Newton step:

$$x_{k+1} - x^* = x_k - x^* - [f''(x_k)]^{-1} f'(x_k)$$

We already know:

$$f'(x_k) = f'(x_k) - f'(x^*) = \int_0^1 f''(x^* + \tau(x_k - x^*))(x_k - x^*)\, d\tau$$

Therefore,

$$x_{k+1} - x^* = x_k - x^* - [f''(x_k)]^{-1} \int_0^1 f''(x^* + \tau(x_k - x^*))(x_k - x^*)\, d\tau$$

Trivial identity: $x_k - x^* = [f''(x_k)]^{-1} \int_0^1 f''(x_k)(x_k - x^*)\, d\tau$

$$\Rightarrow\ x_{k+1} - x^* = [f''(x_k)]^{-1} G_k (x_k - x^*)$$

where $G_k = \int_0^1 [f''(x_k) - f''(x^* + \tau(x_k - x^*))]\, d\tau$

Convergence rate of Newton's method

$$G_k = \int_0^1 [f''(x_k) - f''(x^* + \tau(x_k - x^*))]\, d\tau$$

$$\Rightarrow \|G_k\|_{op} \le \int_0^1 \|f''(x_k) - f''(x^* + \tau(x_k - x^*))\|_{op}\, d\tau$$

$$\le \int_0^1 L\|x_k - x^* - \tau(x_k - x^*)\|\, d\tau$$

$$f \in C_L^{2,2}(\mathbb{R}^n)$$

$$= \int_0^1 L(1-\tau)\|x_k - x^*\|\, d\tau = \int_0^1 L(1-\tau) r_k\, d\tau \le \frac{L r_k}{2}$$

$$\Rightarrow \|G_k\|_{op} \le \frac{L r_k}{2}$$

Convergence rate of Newton's method

We have already proved:

$$\left.\begin{array}{l} f \in C_L^{2,2}(\mathbb{R}^n) \\ \|x-y\| = r \end{array}\right\} \Longrightarrow f''(x) - Lr\mathbf{I}_n \preceq f''(y) \preceq f''(x) + Lr\mathbf{I}_n$$

Therefore,

$$f''(x_k) \succeq \underbrace{f''(x^*)}_{\succeq\, l\mathbf{I}_n \text{ ASSUMPTION}} - Lr_k\mathbf{I}_n \succeq l\mathbf{I}_n - Lr\mathbf{I}_n = (l - Lr_k)\mathbf{I}_n$$

and thus,

If $l - Lr_k > 0$, then $\begin{cases} f''(x_k) \text{ is positive definite} \\ \|[f''(x_k)]^{-1}\|_{op} \leq \frac{1}{l-Lr_k} \end{cases}$

Convergence rate of Newton's method

We already know:

$$r_{k+1} = \|x_{k+1} - x^*\| = \|[f''(x_k)]^{-1} G_k (x_k - x^*)\|$$

$$\le \underbrace{\|[f''(x_k)]^{-1}\|_{op}}_{\le (l - Lr_k)^{-1}} \; \underbrace{\|G_k\|_{op}}_{\le \frac{r_k}{2} L} \; \underbrace{\|(x_k - x^*)\|}_{= r_k}$$

$$\Rightarrow \; r_{k+1} \le \frac{L r_k^2}{2(l - L r_k)}$$

Convergence rate of Newton's method

$$\Rightarrow r_{k+1} \leq \frac{Lr_k^2}{2(l - Lr_k)}$$

Now, we have that

$$\left.\begin{array}{r} \text{If } l > Lr_k \\ 2l > 3Lr_k \end{array}\right\} \Rightarrow r_{k+1} \leq \frac{Lr_k^2}{2(l - Lr_k)} = \frac{Lr_k^2}{2l - 2Lr_k}$$

$$< \frac{Lr_k^2}{3Lr_k - 2Lr_k} = r_k$$

The error doesn't increase!

Convergence rate of Newton's method

We have proved the following theorem

Theorem [Rate of Newton's method**]**

$$\left.\begin{array}{l}\text{Let } f \text{ satisfy the above asumptions} \\ \bar{r} \doteq \|x_0 - x^*\| \leq \frac{2l}{3L}\end{array}\right\} \Rightarrow$$

$$\Rightarrow \begin{cases} \|x_k - x^*\| \leq \bar{r} \ \forall k \\ \|x_{k+1} - x^*\| \leq \frac{L\|x_k - x^*\|^2}{2(l - L\|x_k - x^*\|)} \leq \begin{cases} c\|x_k - x^*\|^2 \\ \|x_k - x^*\| \end{cases} \end{cases}$$

$\Rightarrow$ **Quadratic rate!**

Summary

Newton method

- Finding a root
- Unconstrained minimization
  - Motivation with quadratic approximation
  - Rate of Newton’s method
- Newton fractals

Classes of differentiable functions