3-rd Order Newton Method and its Global Performance

Boris Polyak and **Sergey Nazin**

Institute of Control Sciences RAS, Moscow, Russia

{*boris, snazin*}*@ipu.ru*

Lab. tutorial seminar; Moscow, Apr. 1, 2008, 11h00.

Main paper

Yu.E. Nesterov, B.T. Polyak.
Cubic Regularization of Newton Method and its Global Performance.
Math. Program., Ser. A, 2006, **108**(1), 117-205.

Available from Lab.7 web-site (see *Selected papers*):

http://www.ipu.ru/s_004/files/lab7/rus/sel-papers.htm

Outline

Introduction to Newton-like Methods for Optimization

Cubic Regularization of Newton Method

Global Properties and Performance

Simulation Examples

Unconstrained Optimization

Let $f : \mathbb{R}^n \to \mathbb{R}$. Find: $\min f(x),\ x \in \mathbb{R}^n$;

or, equivalently, solve: $f'(x) = 0$.

Popular iterative schemes:

Gradient methods: $x_{k+1} = x_k - \gamma_k f'(x_k)$,

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

This seminar: cubic regularization of Newton method.

Introduction to Newton-like Methods

Classical Newton method: solve $P(x) = 0,\ x \in \mathbb{R}^1$.

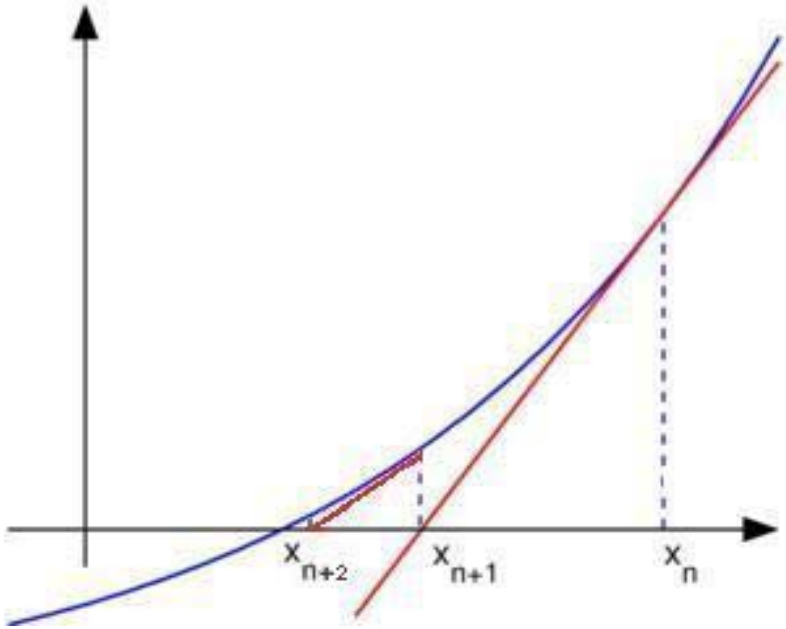
$P(x+h) \approx P(x) + h^T P'(x) = 0$ – linear approximation

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

Local convergence

Quadratic rate of convergence !

$$\|x_k - x^*\| \le C q^{2^k}, \quad q < 1.$$



Historical reference

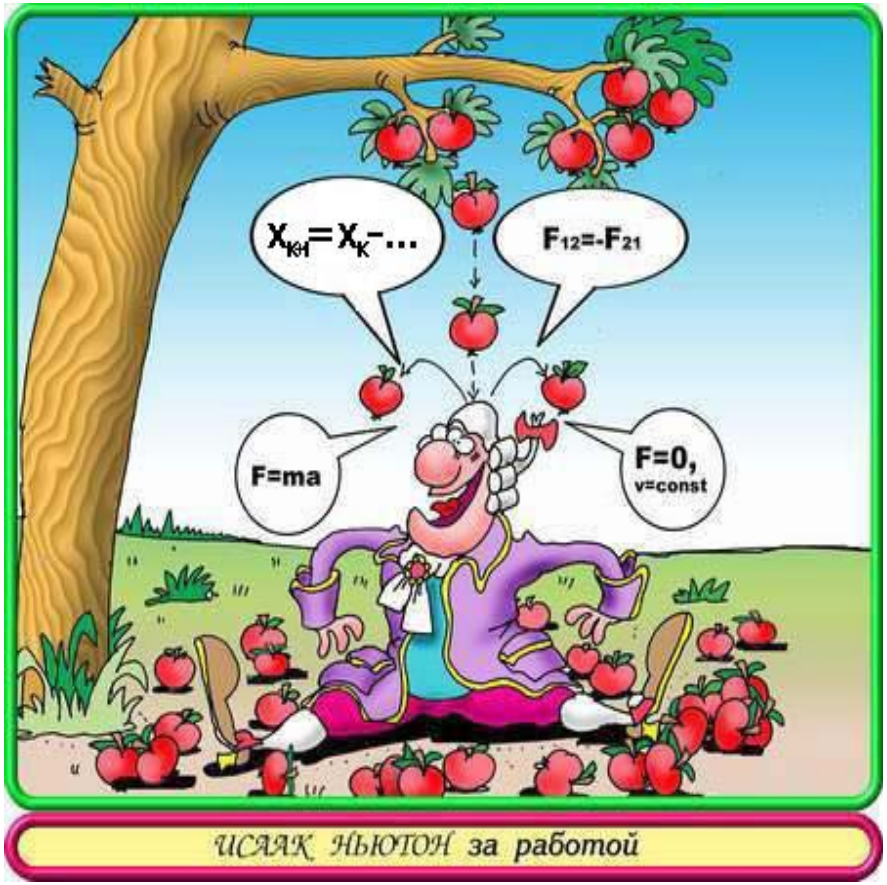
Newton 1669

(published later)

Raphson 1690

. . .

Nesterov, Polyak 2003, 2006



Ortega J.M., Rheinboldt W.C. *Iterative solution of nonlinear equations in several variables.* Academic Press, NY, 1970.

Deulfhard P. *Newton methods for nonlinear problems.* Springer, 2004.

Ypma T.J. Historical development of the Newton-Raphson method. *SIAM Review*, 1995, 531-551.

Main Drawbacks

$$x_{k+1} = x_k - [f''(x_k)]^{-1} f'(x_k) \quad - \text{ Newton scheme}$$

Local convergence

Convergence to saddle point or to local maximum

Possible non-monotone decrease

Hessian $f''(x)$ can become degenerated !!!

Global behavior is unclear

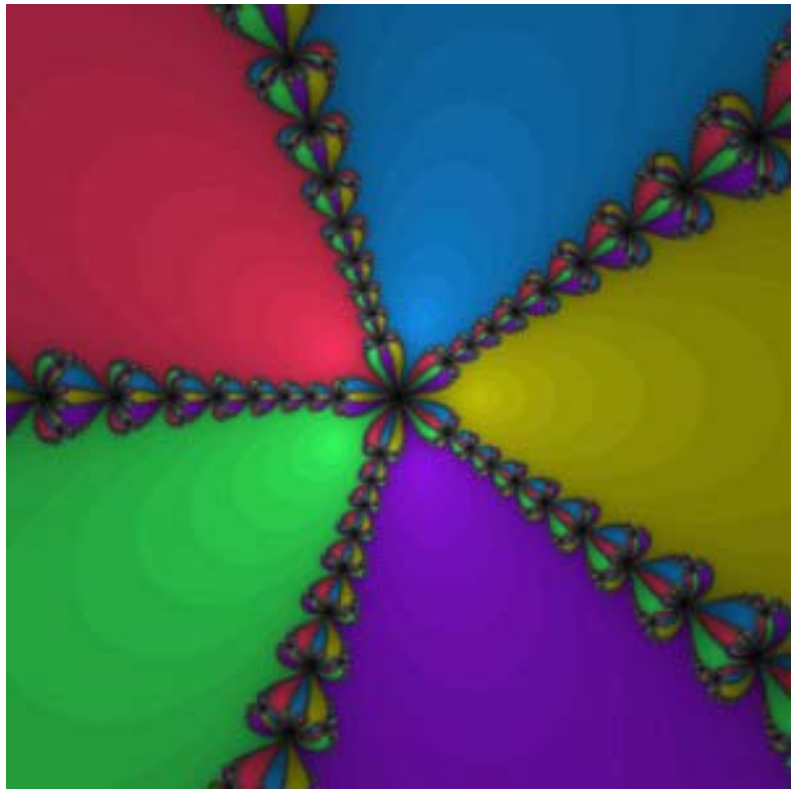
Difficulties on the complex plane

$$p(z) = z^5 - 1,\ z \in \mathbb{C}$$

$$p(z) = 0$$

$z_1 = 1,$ + 4 complex roots

$$z_{k+1} = z_k - [p'(z_k)]^{-1}\, p(z_k)$$



basins of attraction

$$\mathcal{A}_m = \{z_0 \in \mathbb{C} : z_k \to z_m\}$$

Modifications

Damped Newton method

$$x_{k+1} = x_k - \gamma_k \left[f''(x_k)\right]^{-1} f'(x_k)$$

Levenberg-Marquardt regularization

$$x_{k+1} = x_k - \left[f''(x_k) + \alpha\, I\right]^{-1} f'(x_k)$$

Trust-region approach

$\Delta(x_k) = \{x : \quad \|x - x_k\| \leq \varepsilon_k\}$ where $f''(x)$ is well-defined

Selfconcordant functions

$f(x),\ f \in C^3(\mathbb{R}^n)$ is called selfconcordant, if

$f(x)$ is convex,

$|(\nabla^3 f(x)\, h, h, h)| \leq 2\,(\nabla^2 f(x)h, h), \quad \forall\, x, h.$

Damped Newton method

$$x_{k+1} = x_k - \gamma_k \left[\nabla^2 f(x_k)\right]^{-1} f'(x_k)$$

globally converges to local minimum.

Unconstrained minimization

Newton Method

$$f(x) \to \min, \quad x \in \mathbb{R}^n$$

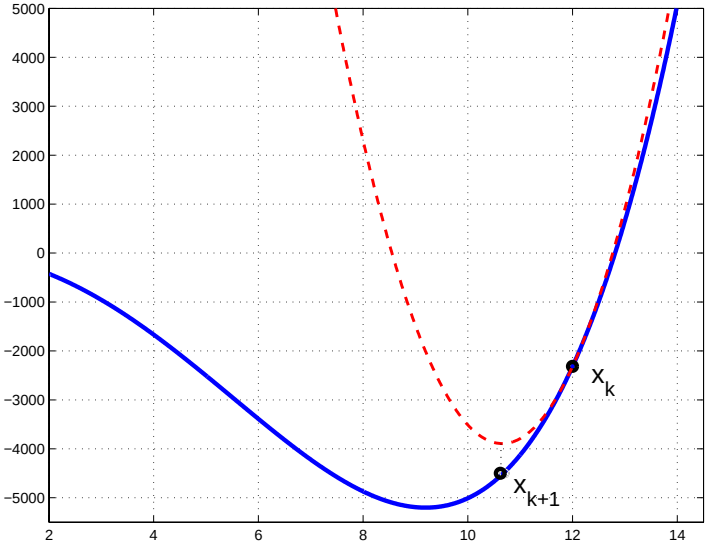
Quadratic approximation:

$$f(x+h) \approx f_x(h) = f(x) + h^T f'(x) + \frac{1}{2} h^T f''(x) h$$

$$x_{k+1} = \arg\min_{h \in \mathbb{R}^n} f_{x_k}(h)$$

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

Quadratic rate of convergence !



Third order Newton method

$$f(x) \to \min, \quad x \in \mathbb{R}^n$$

$$\|f''(x) - f''(y)\| \leq L\|x - y\|, \quad \forall\, x, y \in \mathbb{R}^n$$

(Hessian is Lipschitz continuous)

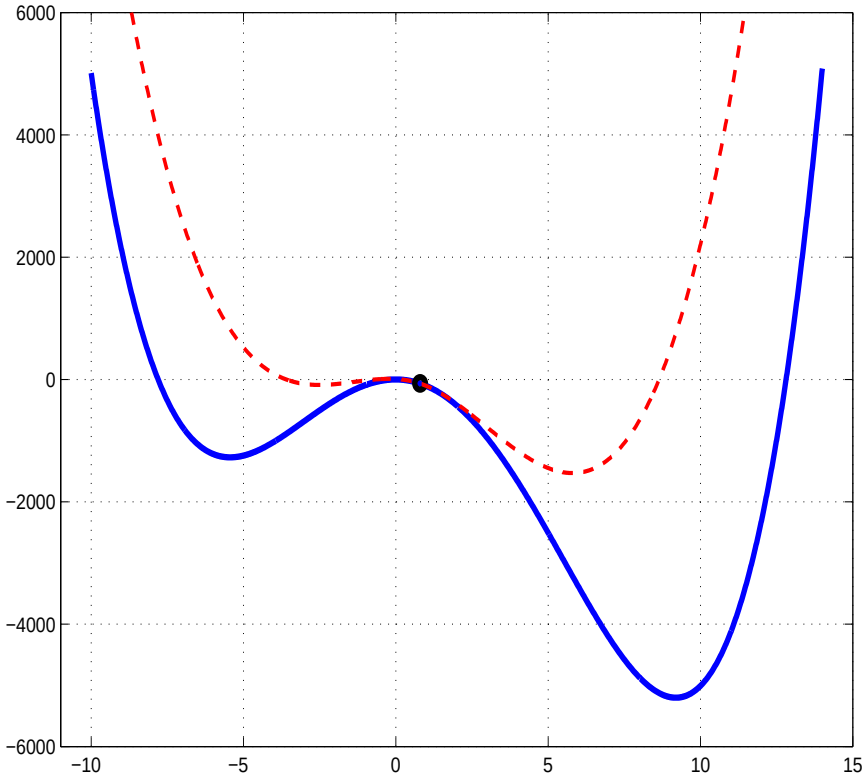
Then: $\forall\, y \in \mathbb{R}^n \qquad f(y) \leq \widehat{f}_x(y) =$

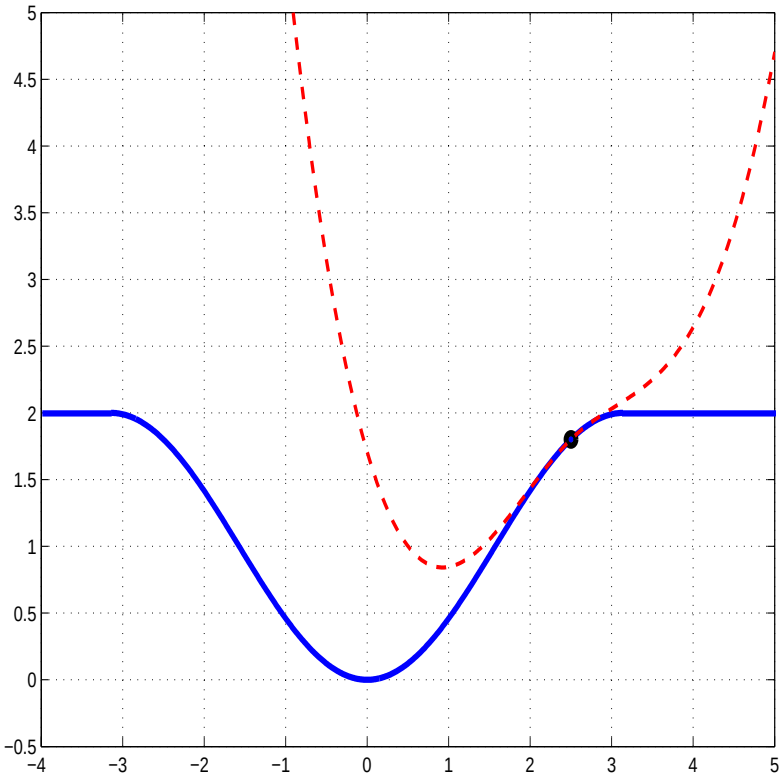
$$= f(x) + (y-x)^T f'(x) + \frac{1}{2}(y-x)^T f''(x)(y-x) + \frac{L}{6}\|y-x\|^3.$$

$$x_{k+1} \in \arg\min_{y \in \mathbb{R}^n} \widehat{f}_{x_k}(y)$$

$$\min_{y \in \mathbb{R}^n} \widehat{f}_x(y)$$

non-convex problem and it can have local minima.





Advantages

Method globally converges to a local minimum

No problem with Hessian

Monotone decrease $f(x_{k+1}) \leq f(x_k)$

Global efficiency on specific classes of functions

Computation of $x_{k+1} \in \arg\min_{y \in \mathbb{R}^n} \widehat{f}_{x_k}(y)$

can be reduced to 1D problem !

1. $\widehat{f}_x(y) \quad \Rightarrow \quad \widetilde{f}(h) = g^T h + \frac{1}{2} h^T H h + \frac{L}{6}\|h\|^3 \ \rightarrow \min_{h \in \mathbb{R}^n}$

2. $\xi_1(h) = g^T h + \frac{1}{2} h^T H h, \ \ \xi_2 = \|h\|^2$ – quadratic transformation

$$Q = \{\xi = (\xi_1, \xi_2)^T : \ \xi_1 = \xi_1(h), \ \xi_2 = \xi_2(h), \ h \in \mathbb{R}^n\} \subset \mathbb{R}^2$$

$$\varphi(\xi) = \xi_1 + \frac{L}{6}(\xi_2)^{3/2}$$

Then:

$$\min_{h\in\mathbb{R}^n}\left[g^T h+\frac{1}{2}h^T Hh+\frac{L}{6}\|h\|^3\right]=\min_{\xi\in Q}\left[\xi_1+\frac{L}{6}(\xi_2)^{3/2}\right]$$

Unconstrained minimization $\Rightarrow$ Minimization with constrains

(Q is closed and convex 2D set)

3. Duality:

$$\min_{\xi\in Q}\varphi(\xi)=\sup_{\tau\in\mathcal{D}}\left[-\frac{1}{2}\,g^T\left(H+\frac{L}{2}\,\tau I\right)^{-1}g-\frac{L}{12}\,\tau^3\right]$$

where $\mathcal{D}=\left\{\tau\in\mathbb{R}_+:\ H+\frac{L}{2}\,\tau I>0\right\}$.

The dual solution can be found from 1D equation

$$\tau = \left\| \left(H + \frac{L}{2}\tau I \right)^{-1} g \right\|, \quad \tau \geq \frac{2}{L}(-\lambda_{\min}(H))_+$$

$$h(\tau) = -\left(H + \frac{L}{2}\tau I \right)^{-1} g$$

Global efficiency on specific problem classes

- Star-convex functions

$$f(\alpha x_* + (1-\alpha)x) \leq \alpha f(x_*) + (1-\alpha) f(x), \;\; \forall x \in \mathcal{X}, \; \forall \alpha \in [0,1].$$

- Gradient-dominated functions

$$f(x) - f(x_*) \leq C \| f'(x) \|^p, \;\; C > 0.$$

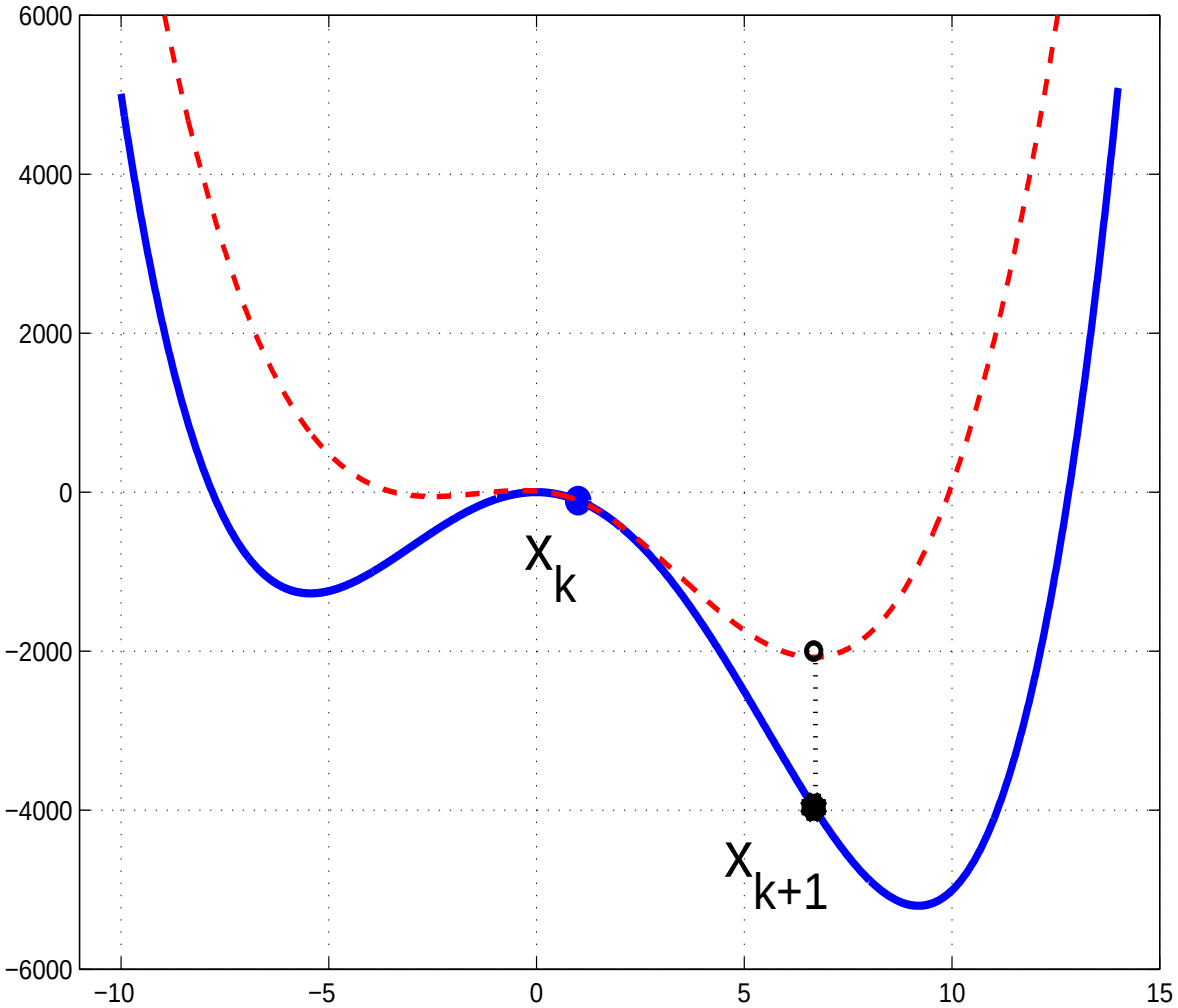
- Nonlinear transformation of convex functions

$$f(x) = \phi(u(x)), \quad u(x) \text{ is convex.}$$

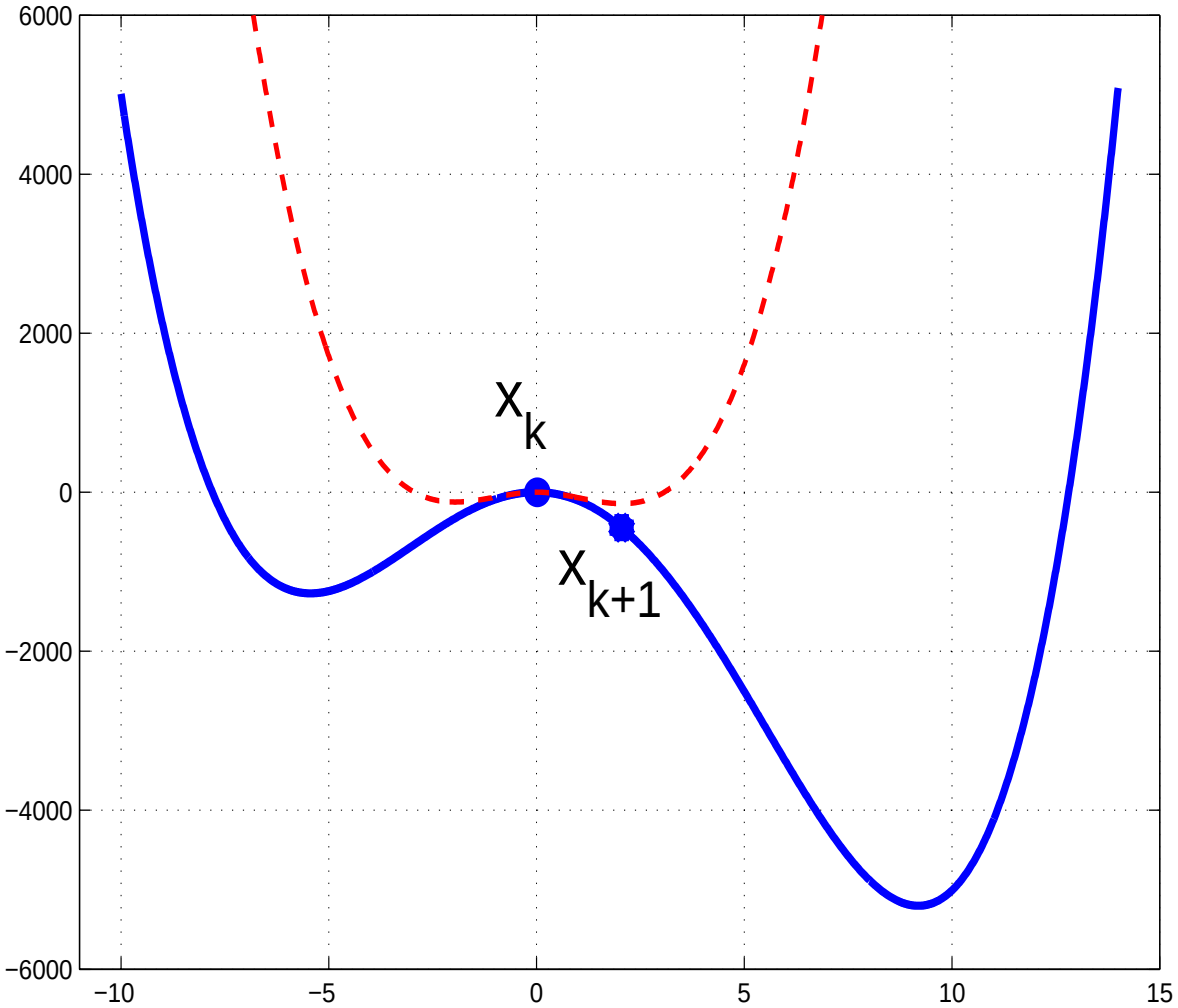
$$\Longrightarrow \quad \text{Super-linear convergence}$$

i.e.: $f(x_{k+1}) - f(x_*) \leq C \, (f(x_k) - f(x_*))^{4/3}$.

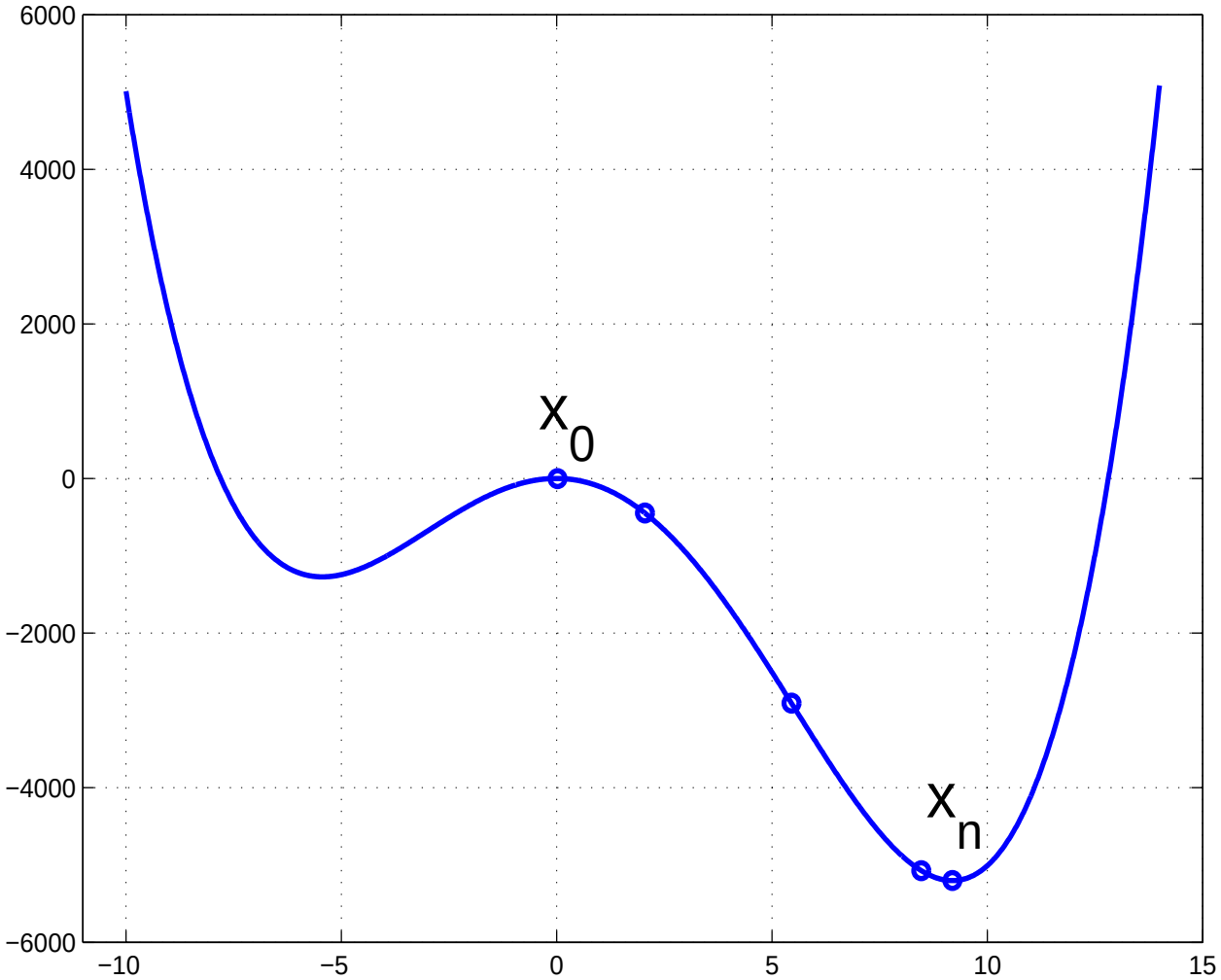
Examples



Examples



Examples



Simulations: Minimization of Rosenbrok function

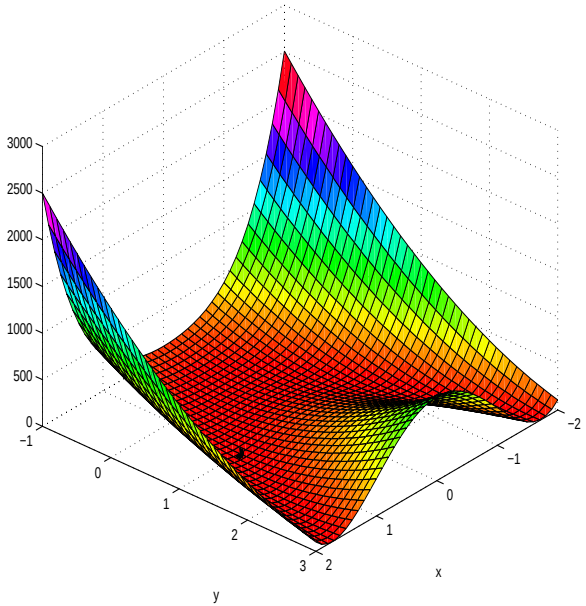
$$f(x_1, x_2) = (x_1 - 1)^2 + 100\,(x_2 - x_1^2)^2$$

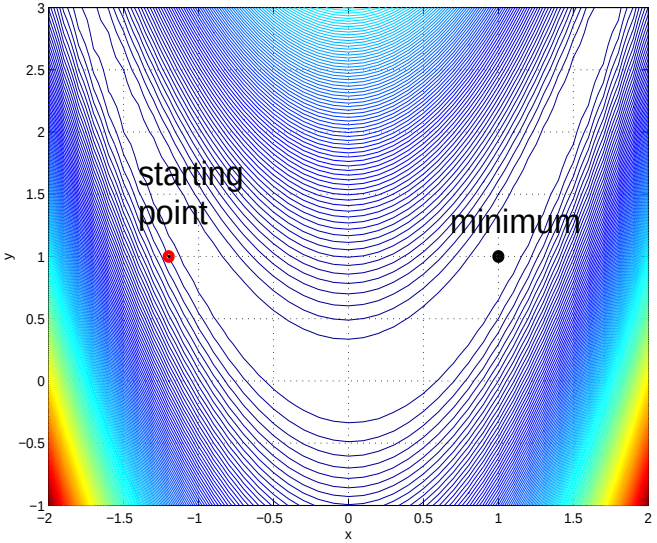
$x^0 = (-1.2,\ 1)^T$

Gradient: $\sim 2 \cdot 10^5$ it.,
$t \approx 5$ sec., $\Delta = 10^{-8}$.

2D Newton: diverges.

3D Newton: ≈ 25 it.,
$t \approx 0.02$ sec., $\Delta = 10^{-8}$.





Conclusion 1

Gradient method:

Slow rate of convergence

Global convergence to local min
Monotone decrease
Computationally cheap

2D Newton method:

Computational costs
Only local convergence
Problems with Hessian
Non-monotone decrease

Quadratic rate of convergence

Conclusion 2

3D Newton method:

Computationally expensive

Non fast rate of convergence

Method globally converges to a local minimum

No problem with Hessian

Monotone decrease $f(x_{k+1}) \leq f(x_k)$

Global efficiency on specific classes of functions