Lecture 6: Second-order methods: Newton's method for unconstrained optimization

Coralia Cartis, Mathematical Institute, University of Oxford

C6.2/B2: Continuous Optimization

Other search directions in Generic Linesearch Methods (GLMs)

Let B^k symmetric, positive definite matrix [$B^k \succ 0$]. Let s^k be defined by

$$B^k s^k = -\nabla f(x^k). \qquad (*)$$

Other search directions in Generic Linesearch Methods (GLMs)

Let B^k symmetric, positive definite matrix [$B^k \succ 0$]. Let s^k be defined by

$$B^k s^k = -\nabla f(x^k). \qquad (*)$$

- $\Longrightarrow$ s^k descent direction:
  $\nabla f(x^k)^T s^k = -\nabla f(x^k)^T (B^k)^{-1} \nabla f(x^k) < 0$ whenever $\nabla f(x^k) \neq 0$ as B^k pos. def. implies $(B^k)^{-1}$ pos. def.

Other search directions in Generic Linesearch Methods (GLMs)

Let B^k symmetric, positive definite matrix [$B^k \succ 0$]. Let s^k be defined by

$$B^k s^k = -\nabla f(x^k). \qquad (*)$$

- $\Longrightarrow$ s^k descent direction:
  $\nabla f(x^k)^T s^k = -\nabla f(x^k)^T (B^k)^{-1} \nabla f(x^k) < 0$ whenever $\nabla f(x^k) \neq 0$ as B^k pos. def. implies $(B^k)^{-1}$ pos. def.
- $\Longrightarrow$ s^k uniquely solves

  minimize$_{s \in \mathbb{R}^n}$ $m_k(s) = f(x^k) + \nabla f(x^k)^T s + \frac{1}{2} s^T B^k s$.

  $m_k(s)$ is a convex quadratic function in s:
  $\nabla m_k(s^k) = \nabla f(x^k) + B^k s^k = 0$ and $\nabla^2 m_k(s) = B^k$.

Other search directions in Generic Linesearch Methods (GLMs)

Let B^k symmetric, positive definite matrix [$B^k \succ 0$]. Let s^k be defined by

$$B^k s^k = -\nabla f(x^k). \qquad (*)$$

- $\Longrightarrow$ s^k descent direction:
  $\nabla f(x^k)^T s^k = -\nabla f(x^k)^T (B^k)^{-1} \nabla f(x^k) < 0$ whenever $\nabla f(x^k) \neq 0$ as B^k pos. def. implies $(B^k)^{-1}$ pos. def.
- $\Longrightarrow$ s^k uniquely solves

  $$\text{minimize}_{s \in \mathbb{R}^n} \; m_k(s) = f(x^k) + \nabla f(x^k)^T s + \tfrac{1}{2} s^T B^k s.$$

  $m_k(s)$ is a convex quadratic function in s:
  $\nabla m_k(s^k) = \nabla f(x^k) + B^k s^k = 0$ and $\nabla^2 m_k(s) = B^k$.

- $(*)$ is a scaled steepest descent direction;
- For some B^k, resulting GLMs can be made scale-invariant, and faster than steepest descent asymptotically

How to choose B^k ?...[Newton, modified Newton, quasi-Newton; to follow.]

Linesearch Newton's method

Let $f \in C^2(\mathbb{R}^n)$ and $B^k := \nabla^2 f(x^k)$ in GLM.

Linesearch-Newton (also called Damped Newton's) method for minimization:

Choose $\epsilon > 0$ and $x^0 \in \mathbb{R}^n$.

While $\|\nabla f(x^k)\| > \epsilon$, REPEAT:

- solve the linear system $\nabla^2 f(x^k) s^k = -\nabla f(x^k)$.
- set $x^{k+1} = x^k + \alpha^k s^k$, where $\alpha^k \in (0, 1]$; $k := k + 1$. END.

Linesearch Newton's method

Let $f \in C^2(\mathbb{R}^n)$ and $B^k := \nabla^2 f(x^k)$ in GLM.

Linesearch-Newton (also called Damped Newton's) method for minimization:

Choose $\epsilon > 0$ and $x^0 \in \mathbb{R}^n$.

While $\|\nabla f(x^k)\| > \epsilon$, REPEAT:

- solve the linear system $\nabla^2 f(x^k) s^k = -\nabla f(x^k)$.
- set $x^{k+1} = x^k + \alpha^k s^k$, where $\alpha^k \in (0, 1]$; $k := k + 1$. END.

- Needs $\nabla^2 f(x^k)$ to be positive definite so that s^k descent. Then α^k can be computed by exact linesearch, bArmijo, etc.
- Whenever $\nabla^2 f(x^k)$ is positive definite, s^k minimizes the second-order Taylor approximation of f around x^k (recall stp. descent minimizes first-order Taylor).

Linesearch Newton's method

Let $f \in C^2(\mathbb{R}^n)$ and $B^k := \nabla^2 f(x^k)$ in GLM.

Linesearch-Newton (also called Damped Newton's) method for minimization:

Choose $\epsilon > 0$ and $x^0 \in \mathbb{R}^n$.

While $\|\nabla f(x^k)\| > \epsilon$, REPEAT:

- solve the linear system $\nabla^2 f(x^k) s^k = -\nabla f(x^k)$.
- set $x^{k+1} = x^k + \alpha^k s^k$, where $\alpha^k \in (0, 1]$; $k := k + 1$. END.

- Needs $\nabla^2 f(x^k)$ to be positive definite so that s^k descent. Then α^k can be computed by exact linesearch, bArmijo, etc.
- Whenever $\nabla^2 f(x^k)$ is positive definite, s^k minimizes the second-order Taylor approximation of f around x^k (recall stp. descent minimizes first-order Taylor).

Some terminology:

Newton direction: $s^k = -(\nabla^2 f(x^k))^{-1} \nabla f(x^k)$.

(Pure) Newton's method: Newton's method without linesearch sets $x^{k+1} = x^k + s^k$ where s^k is the Newton direction for all k.

Connection to Newton's method for root-finding

x^* stationary point of $f \iff \nabla f(x^*) = 0$.

Let $r(x) := \nabla f(x) = 0$ $n \times n$ system of nonlinear equations

Connection to Newton's method for root-finding

x^* stationary point of $f \iff \nabla f(x^*) = 0$.

Let $r(x) := \nabla f(x) = 0$ $\quad n \times n$ system of nonlinear equations

$\longrightarrow$ apply Newton's method for root-finding to $\nabla f(x) = 0$:

Let x^{k+1} s. t. $r(x^k) + J(x^k)(x^{k+1} - x^k) = 0$, where $J(x^k)$ is the Jacobian (matrix) of $r(x)$ at $x = x^k$, i.e., $J(x^k)_{ij} = \left(\frac{\partial r_i}{\partial x_j}\right)(x^k)$.

Connection to Newton's method for root-finding

x^* stationary point of $f \iff \nabla f(x^*) = 0$.

Let $r(x) := \nabla f(x) = 0$ $\quad n \times n$ system of nonlinear equations

$\longrightarrow$ apply Newton's method for root-finding to $\nabla f(x) = 0$:

Let x^{k+1} s. t. $r(x^k) + J(x^k)(x^{k+1} - x^k) = 0$, where $J(x^k)$ is the Jacobian (matrix) of $r(x)$ at $x = x^k$, i.e., $J(x^k)_{ij} = \left(\frac{\partial r_i}{\partial x_j}\right)(x^k)$.

$$J(x^k) \text{ nonsingular} \implies x^{k+1} = x^k - (J(x^k))^{-1} r(x^k).$$

Connection to Newton's method for root-finding

x^* stationary point of $f \iff \nabla f(x^*) = 0$.

Let $r(x) := \nabla f(x) = 0$ $\quad n \times n$ system of nonlinear equations

$\longrightarrow$ apply Newton's method for root-finding to $\nabla f(x) = 0$:

Let x^{k+1} s. t. $r(x^k) + J(x^k)(x^{k+1} - x^k) = 0$, where $J(x^k)$ is the Jacobian (matrix) of $r(x)$ at $x = x^k$, i.e., $J(x^k)_{ij} = \left(\frac{\partial r_i}{\partial x_j}\right)(x^k)$.

$$J(x^k) \text{ nonsingular} \implies x^{k+1} = x^k - (J(x^k))^{-1} r(x^k).$$

- The Jacobian of ∇f at x is the Hessian matrix $\nabla^2 f(x)$

$$\Downarrow \nabla^2 f(x^k) \text{ nonsingular}$$

$$\textbf{(Pure) Newton iterate : } x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k).$$

Advantages of Newton's method for optimization

- Fast (i.e., quadratic) local rate of convergence.

Theorem 7 (local convergence of (pure) Newton's method):

- let $f \in \mathcal{C}^2(\mathbb{R}^n)$, $\nabla f(x^*) = 0$ with $\nabla^2 f(x^*)$ nonsingular;
- $\nabla^2 f$ locally Lipschitz continuous at x^*.

If x^{k_0} is sufficiently close to x^*, for some $k_0 \geq 0$,

$\Longrightarrow$ x^k is well-defined for all $k \geq k_0$;

$x^k \to x^*$ as $k \to \infty$, at quadratic rate. $\square$

Advantages of Newton's method for optimization

- Fast (i.e., quadratic) local rate of convergence.

Theorem 7 (local convergence of (pure) Newton's method):

- let $f \in \mathcal{C}^2(\mathbb{R}^n)$, $\nabla f(x^*) = 0$ with $\nabla^2 f(x^*)$ nonsingular;
- $\nabla^2 f$ locally Lipschitz continuous at x^*.

If x^{k_0} is sufficiently close to x^*, for some $k_0 \geq 0$,

$\Longrightarrow$ x^k is well-defined for all $k \geq k_0$;

$x^k \to x^*$ as $k \to \infty$, at quadratic rate. $\square$

- In the conditions of Th 7: $\nabla f(x^k) \to 0$ quadratically as well.
- "x^{k_0} sufficiently close to x^*"= there exists $\mathcal{N}(x^*, \delta)$ such that $x^{k_0} \in \mathcal{N}$. In general, $\mathcal{N}$ not known beforehand (depends on unknown x^* and problem-dependent constants).

Advantages of Newton's method for optimization

Sketch of Proof for Theorem 7:

Taylor expansion of ∇f around x [vector form]:

$$\nabla f(x^*) = \nabla f(x) + \nabla^2 f(x)(x^* - x) + \mathcal{O}(\|x^* - x\|^2),$$

where x is sufficiently close to x^* and $\mathcal{O}(\cdot)$ depends on the Lipschitz constant of $\nabla^2 f(x^*)$.

Advantages of Newton's method for optimization

Sketch of Proof for Theorem 7:

Taylor expansion of ∇f around x [vector form]:

$$\nabla f(x^*) = \nabla f(x) + \nabla^2 f(x)(x^* - x) + \mathcal{O}(\|x^* - x\|^2),$$

where x is sufficiently close to x^* and $\mathcal{O}(\cdot)$ depends on the Lipschitz constant of $\nabla^2 f(x^*)$. Using $\nabla f(x^*) = 0$ and $x := x^k$, whenever x^k suff. close to x^*, we have

$$0 = \nabla f(x^k) + \nabla^2 f(x^k)(x^* - x^k) + \mathcal{O}(\|x^* - x^k\|^2). \quad (**)$$

Advantages of Newton's method for optimization

Sketch of Proof for Theorem 7:

Taylor expansion of ∇f around x [vector form]:

$$\nabla f(x^*) = \nabla f(x) + \nabla^2 f(x)(x^* - x) + \mathcal{O}(\|x^* - x\|^2),$$

where x is sufficiently close to x^* and $\mathcal{O}(\cdot)$ depends on the Lipschitz constant of $\nabla^2 f(x^*)$. Using $\nabla f(x^*) = 0$ and $x := x^k$, whenever x^k suff. close to x^*, we have

$$0 = \nabla f(x^k) + \nabla^2 f(x^k)(x^* - x^k) + \mathcal{O}(\|x^* - x^k\|^2). \quad (**)$$

$\nabla^2 f(x^*)$ nonsingular $\Longrightarrow$ $\nabla^2 f(x^k)$ nonsingular whenever x^k suff. close to x^*. Then (**) implies

$x^k - x^* = [\nabla^2 f(x^k)]^{-1} \nabla f(x^k) + \mathcal{O}(\|x^* - x^k\|^2)$.

Advantages of Newton's method for optimization

Sketch of Proof for Theorem 7:

Taylor expansion of ∇f around x [vector form]:

$$\nabla f(x^*) = \nabla f(x) + \nabla^2 f(x)(x^* - x) + \mathcal{O}(\|x^* - x\|^2),$$

where x is sufficiently close to x^* and $\mathcal{O}(\cdot)$ depends on the Lipschitz constant of $\nabla^2 f(x^*)$. Using $\nabla f(x^*) = 0$ and $x := x^k$, whenever x^k suff. close to x^*, we have

$$0 = \nabla f(x^k) + \nabla^2 f(x^k)(x^* - x^k) + \mathcal{O}(\|x^* - x^k\|^2). \ (**)$$

$\nabla^2 f(x^*)$ nonsingular $\Longrightarrow$ $\nabla^2 f(x^k)$ nonsingular whenever x^k suff. close to x^*. Then (**) implies
$x^k - x^* = [\nabla^2 f(x^k)]^{-1} \nabla f(x^k) + \mathcal{O}(\|x^* - x^k\|^2)$. Letting s^k be the Newton direction, and $x^{k+1} = x^k + s^k$, we deduce that, $x^k - x^* = x^k - x^{k+1} + \mathcal{O}(\|x^* - x^k\|^2)$, and so $x^{k+1} - x^* = \mathcal{O}(\|x^k - x^*\|^2)$. $\square$

Local convergence for linesearch-Newton's method

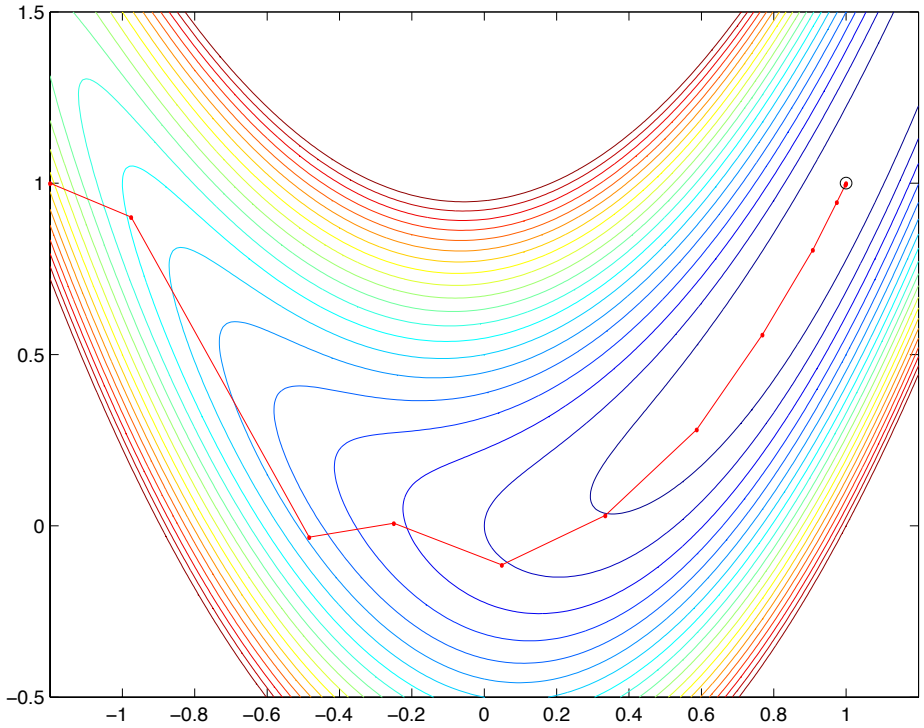
Theorem 8 Let $f \in \mathcal{C}^2(\mathbb{R}^n)$ and $\nabla^2 f$ be Lipschitz continuous and positive definite at the iterates.

Apply Newton's method with bArmijo linesearch and the choices $\beta \leq 0.5$ and $\alpha_{(0)} = 1$. Assume the iterates $x^k \to x^*$ as $k \to \infty$, where $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succ 0$.

Then $\alpha^k = 1$ for all k sufficiently large, and the rate of convergence of x^k to x^* is quadratic (asymptotically).

Local convergence for Newton with bArmijo linesearch

$f(x_1, x_2) = 10(x_2 - x_1^2)^2 + (x_1 - 1)^2; \quad x^* = (1, 1).$



Newton with bArmijo linesearch applied to the Rosenbrock function f.

- $\beta < 0.5$ and $\alpha_{(0)} = 1$ in bArmijo; $\alpha^k = 1$ for suff. large k.

Advantages of Newton's method for optimization

- Newton's method (with or without linesearch) is scale invariant with respect to linear transformations of variables.

Let $A \in \mathbb{R}^{n \times n}$ nonsingular matrix and $y = Ax$
(A is constant, independent of x and y); let $B = A^{-1}$.

Let $\overline{f}(y) := f(x(y)) = f(By)$; minimize $\overline{f}$ wrt y.

$\Longrightarrow \quad \nabla \overline{f}(y) = B^T \nabla f(x)$ and $\nabla^2 \overline{f}(y) = B^T \nabla^2 f(x) B$.

Newton direction at y:
$$\begin{aligned} s_y &= -[B^T \nabla^2 f(x) B]^{-1} B^T \nabla f(x) \\ &= -B^{-1} [\nabla^2 f(x)]^{-1} B^{-T} B^T \nabla f(x) \\ &= -B^{-1} [\nabla^2 f(x)]^{-1} \nabla f(x) \\ &= A s_x. \end{aligned}$$

$\Longrightarrow \quad y + \alpha s_y = A(x + \alpha s_x)$.

Thus $y + \alpha s_y \approx y^* \quad \Longrightarrow \quad x + \alpha s_x \approx x^*$, where $y^* = Ax^*$.

Disadvantages of Newton's method for optimization

- Newton's method with/without linesearch: the Newton direction s^k is not well-defined if $\nabla^2 f(x^k)$ singular; s^k may not be descent if $\nabla^2 f(x^k)$ is not positive definite.

Disadvantages of Newton's method for optimization

- Newton's method with/without linesearch: the Newton direction s^k is not well-defined if $\nabla^2 f(x^k)$ singular; s^k may not be descent if $\nabla^2 f(x^k)$ is not positive definite.
- Newton's method ('pure', without linesearch): iterates can get attracted to local maxima or saddle points of f if sufficiently close to them (in the conditions of local convergence Theorem 7, $\nabla^2 f(x^*)$ only required to be nonsingular).

Disadvantages of Newton's method for optimization

- Newton's method with/without linesearch: the Newton direction s^k is not well-defined if $\nabla^2 f(x^k)$ singular; s^k may not be descent if $\nabla^2 f(x^k)$ is not positive definite.
- Newton's method ('pure', without linesearch): iterates can get attracted to local maxima or saddle points of f if sufficiently close to them (in the conditions of local convergence Theorem 7, $\nabla^2 f(x^*)$ only required to be nonsingular).
- Newton's method ('pure', without linesearch): iterates may fail to converge at all if x^0 'too far' from solution (outside neighbourhood of local convergence, failure may occur). Thus linesearch is needed to make Newton's method globally convergent.

Disadvantages of Newton's method for optimization

Example of failure of (pure) Newton's method to converge globally.

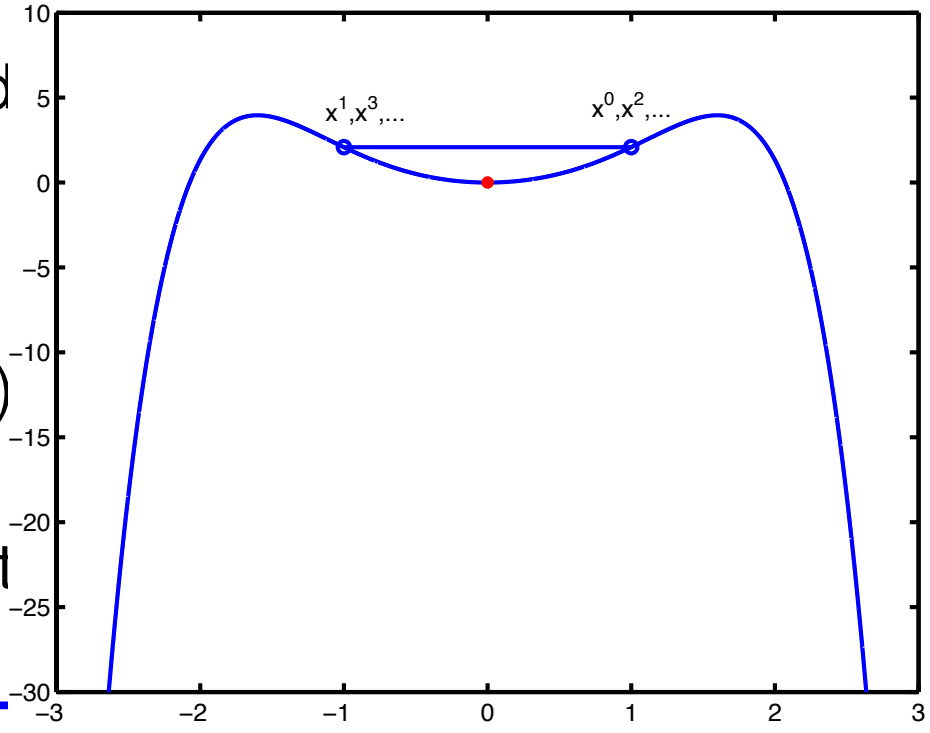
$$f : \mathbb{R} \to \mathbb{R}, \quad f(x) = -\frac{x^6}{6} + \frac{x^4}{4} + 2x^2.$$

$x^* = 0$ local minimizer; $x = \pm\sqrt{(1+\sqrt{17})/2} \approx \pm 1.6$ global max.

Newton's method applied to f, with $x^0 = 1$;
$\Rightarrow$ $x^{2k} = 1$ and
$x^{2k+1} = -1$, for all k.

-1 and 1 are not (even) stationary points of f.

Note that s^k descent but we have gone 'too far'.



Newton's method applied to f.