Modern Optimization Techniques

2. Unconstrained Optimization / 2.3. Newton's Method

Lars Schmidt-Thieme

Information Systems and Machine Learning Lab (ISMLL)
Institute for Computer Science
University of Hildesheim, Germany

Syllabus

Mon.	1.11.	(1)	0. Overview
			1. Theory
Mon.	8.11.	(2)	1. Convex Sets and Functions
			2. Unconstrained Optimization
Mon.	15.11.	(3)	2.1 Gradient Descent
Mon.	22.11.	(4)	2.2 Stochastic Gradient Descent
Mon.	29.11.	(5)	2.3 Newton's Method
Mon.	6.12.	(6)	2.4 Quasi-Newton Methods
Mon.	13.12.	(7)	2.5 Subgradient Methods
Mon.	20.12.	(8)	2.6 Coordinate Descent
		—	— *Christmas Break* —
			3. Equality Constrained Optimization
Mon.	3.1.	(9)	3.1 Duality
Mon.	10.1.	(10)	3.2 Methods
			4. Inequality Constrained Optimization
Mon.	17.1.	(11)	4.1 Primal Methods
Mon.	24.1.	(12)	4.2 Barrier and Penalty Methods
Mon.	31.1.	(13)	4.3 Cutting Plane Methods
Mon.	7.2.	(14)	Q & A

Outline

1. Newton's Method

2. Convergence

Outline

1. Newton's Method

2. Convergence

An idea using second order approximations

Be $f : X \to \mathbb{R}, X \subseteq \mathbb{R}^N$ open and f convex:

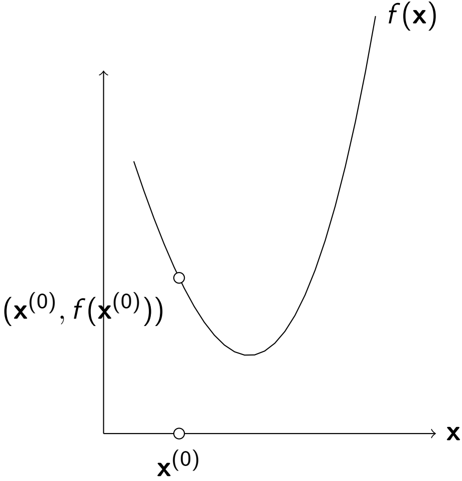
$$\arg\min_{x \in X} \quad f(\mathbf{x})$$

- Let $\mathbf{x}^{(k)}$ the last iterate
- Compute a quadratic approximation $\hat{f}$ of f around $\mathbf{x}^{(k)}$
- Find the minimum of the quadratic approximation $\hat{f}$ and take it as next iterate:

$$\mathbf{x}^{(k+1)} := \arg\min_{x \in X} \hat{f}(\mathbf{x})$$

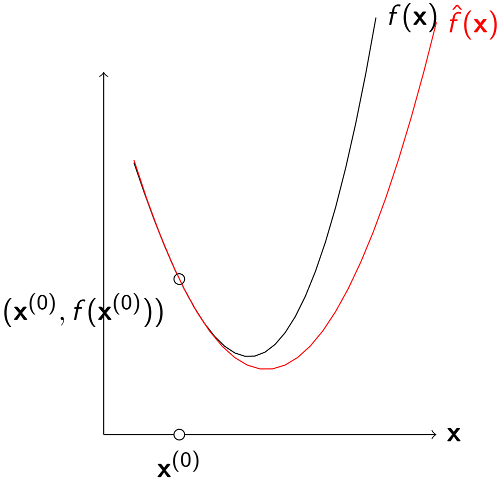
An idea using second order approximations

$$f(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-3)^2 + \frac{1}{10}\mathbf{x}^3$$



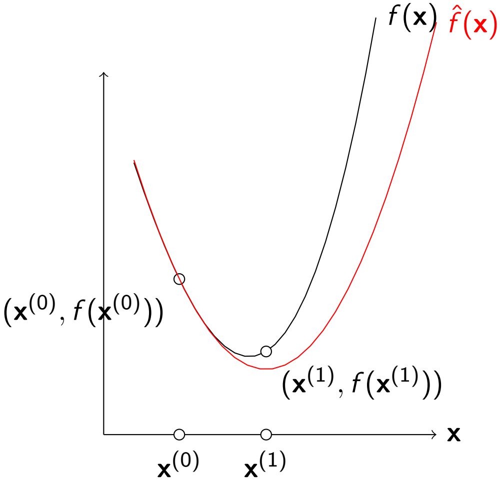
An idea using second order approximations

$$f(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-3)^2 + \frac{1}{10}\mathbf{x}^3$$



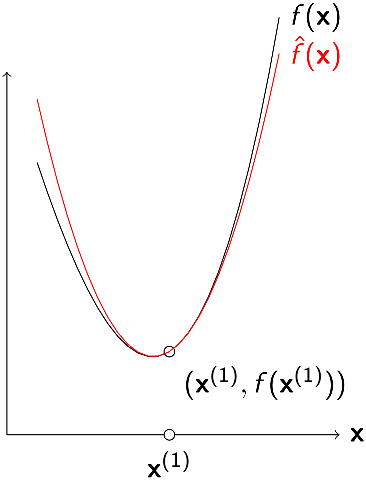
An idea using second order approximations

$$f(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-3)^2 + \frac{1}{10}\mathbf{x}^3$$



An idea using second order approximations

$$f(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-3)^2 + \frac{1}{10}\mathbf{x}^3$$



Taylor Approximation

Be $f : X \to \mathbb{R}, X \subseteq \mathbb{R}^N$ an infinitely differentiable function,
$\mathbf{a} \in X$ any point.

f can be represented by its **Taylor expansion**:

$$f(\mathbf{x}) = \sum_{k=0}^{\infty} \frac{\nabla^k f(\mathbf{a})}{k!}(\mathbf{x}-\mathbf{a})^k$$
$$= f(\mathbf{a}) + \frac{\nabla f(\mathbf{a})}{1!}(\mathbf{x}-\mathbf{a}) + \frac{\nabla^2 f(\mathbf{a})}{2!}(\mathbf{x}-\mathbf{a})^2 + \frac{\nabla^3 f(\mathbf{a})}{3!}(\mathbf{x}-\mathbf{a})^3 + \cdots$$

For x close enough to a and K large enough,
f can be approximated by its **truncated Taylor expansion**:

$$f(\mathbf{x}) \approx \sum_{k=0}^{K} \frac{\nabla^k f(\mathbf{a})}{k!}(\mathbf{x}-\mathbf{a})^k$$

Note: For $N > 1$, $\nabla^k f(x)$ is a tensor of order k and $\nabla^k f(x)(x-a)^k$ a tensor product.

Second Order Approximation

Let us take the second order approximation of a twice differentiable function $f : X \to \mathbb{R}, X \subseteq \mathbb{R}^N$ at a point $\mathbf{x}$:

$$\hat{f}(\mathbf{y}) := f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{1}{2}(\mathbf{y} - \mathbf{x})^T \nabla^2 f(\mathbf{x})(\mathbf{y} - \mathbf{x})$$

We want to find the point $x^{\text{next}} := \arg\min_y \hat{f}(y)$:

Q: How can we find the point x^{next}?

Second Order Approximation

Let us take the second order approximation of a twice differentiable function $f : X \to \mathbb{R}, X \subseteq \mathbb{R}^N$ at a point $\mathbf{x}$:

$$\hat{f}(\mathbf{y}) := f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{1}{2}(\mathbf{y}-\mathbf{x})^T \nabla^2 f(\mathbf{x})(\mathbf{y}-\mathbf{x})$$

We want to find the point $x^{\text{next}} := \arg\min_y \hat{f}(y)$:

$$\nabla_{\mathbf{y}} \hat{f}(\mathbf{y}) = \nabla f(\mathbf{x}) + \nabla^2 f(\mathbf{x})(\mathbf{y}-\mathbf{x}) \stackrel{!}{=} 0$$
$$\rightsquigarrow \quad \mathbf{y} = \mathbf{x} - \nabla^2 f(\mathbf{x})^{-1} \nabla f(\mathbf{x})$$

Newton's Step

- Newton's method is a descent method

- It uses the descent direction

  $$\Delta \mathbf{x} := -\nabla^2 f(\mathbf{x})^{-1} \nabla f(\mathbf{x})$$

  called **Newton step**.
  - the negative gradient
  - twisted by the local curvature (Hessian)

- Newton's step is affine invariant,
  while the gradient step is not.

Newton's Step / Proof

(i) Show that the Gradient step is not affine invariant.

for $g(y) := f(Ay)$ with a pos.def. matrix A

$$\nabla_y g(y) = A^T \nabla_x f(Ay) \stackrel{?}{=} A^{-1} \nabla_x f(x), \quad \text{for } x := Ay$$

No, as in general $A^T \neq A^{-1}$.

(ii) Show that Newton's step is affine invariant.

$$\begin{aligned}
\nabla_y^2 g(y) &= A^T \nabla_x^2 f(Ay) A \\
\Delta y &= (\nabla_y^2 g(y))^{-1} \nabla_y g(y) \\
&= A^{-1} \nabla_x^2 f(Ay)^{-1} (A^T)^{-1} A^T \nabla_x f(Ay) \\
&= A^{-1} \nabla_x^2 f(Ay)^{-1} \nabla_x f(Ay) \\
&= A^{-1} \nabla_x^2 f(x)^{-1} \nabla_x f(x), \quad \text{for } x := Ay \\
&= A^{-1} \Delta x
\end{aligned}$$

Newton's Stepsize

- For quadratic objective functions f:
  - Newton's method will find the minimum in a single step
    - with stepsize 1

    (**pure Newton**)

- For general objective functions:
  - a possibly smaller stepsize has to be used
    (**damped Newton**)

  - any stepsize controller is applicable

Newton Decrement

$$\lambda(x) := (\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x))^{\frac{1}{2}}$$

is called **newton decrement**.

Basic properties:

(i)

$$\lambda(x) = (\Delta x^T \nabla^2 f(x) \Delta x)^{\frac{1}{2}}$$

(ii)

$$\lambda(x)^2 = -\nabla f(x)^T \Delta x$$

(iii)

$$f(x) - \inf_y \hat{f}(y) = f(x) - \hat{f}(x + \Delta x) = \frac{1}{2}\lambda(x)^2$$

(iv) The Newton decrement is affine invariant.

Newton Decrement / Proofs

ad (i), (ii) insert the definition of $\Delta x = -\nabla^2 f(x)^{-1} \nabla f(x)$

ad (iii)

$$f(x) - \hat{f}(x + \Delta x) = f(x) - f(x) \underbrace{-\nabla f(x)^T \Delta x}_{\stackrel{ii}{=}\lambda(x)^2} - \frac{1}{2} \underbrace{\Delta x^T \nabla^2 f(x)}_{\stackrel{i}{=}\lambda(x)^2}$$

ad (iv) for $g(y) := f(Ay)$ with a pos.def. matrix A

$$\begin{aligned}
\nabla_y g(y) &= A^T \nabla_x f(Ay), \quad \nabla_y^2 g(y) = A^T \nabla_x^2 f(Ay) A \\
\lambda_g(y) &= \nabla_x f(Ay)^T A A^{-1} \nabla_x^2 f(Ay)^{-1} (A^T)^{-1} A^T \nabla_x f(Ay)^T \\
&= \nabla_x f(Ay)^T \nabla_x^2 f(Ay)^{-1} \nabla_x f(Ay)^T \\
&= \lambda_f(x) \text{ at } x := Ay
\end{aligned}$$

Newton's Method

min-newton$(f, \nabla f, \nabla^2 f, x^{(0)}, \mu, \epsilon, K)$:
  for $k := 1, \ldots, K$:
    $\Delta x^{(k-1)} := -\nabla^2 f(x^{(k-1)})^{-1} \nabla f(x^{(k-1)})$
    if $-\nabla f(x^{(k-1)})^T \Delta x^{(k-1)} < \epsilon$:
      return $x^{(k-1)}$
    $\mu^{(k-1)} := \mu(f, x^{(k-1)}, \Delta x^{(k-1)})$
    $x^{(k)} := x^{(k-1)} + \mu^{(k-1)} \Delta x^{(k-1)}$
  return "not converged"

where

- f objective function
- ∇f, $\nabla^2 f$ gradient and Hessian of objective function f
- $x^{(0)}$ starting value
- μ step length controller
- ϵ convergence threshold for Newton's decrement
- K maximal number of iterations

Considerations

- works extremely well for a lot of problems.
- requires f to be twice differentiable.
- computing, storing and inverting the Hessian limits scalability for high dimensional problems.
  - as the Hessian has N^2 elements.

Newton's method / Example

For $\mathbf{x} \in \mathbb{R}$

$$\min_{\mathbf{x}} \quad (2\mathbf{x} - 4)^4$$

Q: What is the Newton update step?

Newton's method / Example

For $\mathbf{x} \in \mathbb{R}$

$$\min_{\mathbf{x}} \quad (2\mathbf{x} - 4)^4$$

Algorithm:

- $\nabla f(\mathbf{x}) = 8\ (2\mathbf{x} - 4)^3$
- $\nabla^2 f(\mathbf{x}) = 48\ (2\mathbf{x} - 4)^2$
- Step:

$$\begin{aligned} \Delta\mathbf{x} &= -\nabla^2 f(\mathbf{x})^{-1} \nabla f(\mathbf{x}) \\ &= -\frac{1}{6}(2\mathbf{x} - 4) = -\frac{1}{3}x + \frac{2}{3} \end{aligned}$$

- Update:

$$\begin{aligned} x^{(k+1)} &= x^{(k)} + \mu^{(k)} \Delta x^{(k)}, \quad \text{using } \mu^{(k)} := 1 \\ &= x^{(k)} - \frac{1}{3}x^{(k)} + \frac{2}{3} = \frac{2}{3}(x^{(k)} + 1) \end{aligned}$$

Newton's method / Example

$$
\begin{aligned}
x^{(0)} &:= 10 \\
x^{(1)} &= \frac{2}{3}(10.0 + 1) = 7.33333 \\
x^{(2)} &= \frac{2}{3}(7.33333 + 1) = 5.55556 \\
x^{(3)} &= \frac{2}{3}(5.55556 + 1) = 4.37037 \\
x^{(4)} &= \frac{2}{3}(4.37037 + 1) = 3.58025 \\
x^{(5)} &= \frac{2}{3}(3.58025 + 1) = 3.0535 \\
x^{(6)} &= \frac{2}{3}(3.0535 + 1) = 2.70233 \\
x^{(7)} &= \frac{2}{3}(2.70233 + 1) = 2.46822 \\
x^{(8)} &= \frac{2}{3}(2.46822 + 1) = 2.31215 \\
x^{(9)} &= \frac{2}{3}(2.31215 + 1) = 2.2081 \\
x^{(10)} &= \frac{2}{3}(2.2081 + 1) = 2.13873
\end{aligned}
$$

Outline

Strongly Convex Functions / Basic Facts (Review)

(def) the eigenvalues of the Hessian are uniformly bounded from below:

$$\nabla^2 f(x) \succeq mI, \quad \exists m \in \mathbb{R}^+ \; \forall x \in \operatorname{dom} f$$

(i) f is above a parabola:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} ||y - x||_2^2$$

$$p^* \geq f(x) - \frac{1}{2m} ||\nabla f(x)||_2^2$$

(ii) if f is closed and S one of its sublevel sets, then

a) the eigenvalues of the Hessian are also uniformly bounded from above on S:

$$\nabla^2 f(x) \preceq MI, \quad \exists M \in \mathbb{R}^+ \; \forall x \in S$$

b) f is below a parabola ("sandwiched between two parabolas"):

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{M}{2} ||y - x||_2^2, \quad x, y \in S$$

$$p^* \leq f(x) - \frac{1}{2M} ||\nabla f(x)||_2^2$$

Newton Decrement / Strongly Convex Functions

If f is strongly convex ($\nabla^2 f(x) \succeq mI, m \in \mathbb{R}^+$), then
(i)

$$m||\Delta x||_2^2 \leq \lambda(x)^2 \leq M||\Delta x||_2^2$$

(ii)

$$\frac{1}{M}||\nabla f(x)||_2^2 \leq \lambda(x)^2 \leq \frac{1}{m}||\nabla f(x)||_2^2$$

where $\nabla^2 f(x) \preceq MI, M \in \mathbb{R}^+$.

Newton Decrement / Strongly Convex Functions / Proofs

ad (i)

$$\lambda(x)^2 = \Delta x^T \nabla^2 f(x) \Delta x \geq m||\Delta x||_2^2$$
$$\lambda(x)^2 = \Delta x^T \nabla^2 f(x) \Delta x \leq M||\Delta x||_2^2$$

ad (ii) The inverse of $\nabla^2 f(x)$ has inverse eigenvalues, thus

$$\nabla^2 f(x)^{-1} \preceq \frac{1}{m} I$$
$$\nabla^2 f(x)^{-1} \succeq \frac{1}{M} I$$

Then proceed as (i).

Convergence / Assumptions

Until the end of this section, assume

I. f is strongly convex (m, M),

II. $\nabla^2 f(x)$ is Lipschitz-continuous:
$||\nabla^2 f(y) - \nabla^2 f(x)||_2 \leq L||y - x||_2, \quad L \in \mathbb{R}^+$ and

III. backtracking stepsize control is used $(\alpha \leq \frac{1}{2}, \beta)$

Convergence / Damped Phase

Theorem (Convergence of Newton's Algorithm / Damped Phase)

Far away from the minimum,

(i) *backtracking may select stepsizes* $t < 1$ *(be damped) and*

(ii) f *is reduced by at least a constant each step.*

$$\text{for } ||\nabla f(x)||_2 \geq \eta : \; f(x^{next}) - f(x) \leq -\gamma$$
$$\text{with } \gamma := \alpha\beta\frac{m}{M^2}\eta^2$$

Convergence / Damped Phase / Proof

$$
\begin{aligned}
f(x + t\Delta x) &\underset{\text{s.c. ii}}{\leq} f(x) + t\nabla f(x)^T \Delta x + \frac{M}{2} ||\Delta x||_2^2 t^2 \\
&\underset{\text{dec. ii}}{\leq} f(x) - t\lambda(x)^2 + \frac{M}{2m} t^2 \lambda(x)^2
\end{aligned} \tag{1}
$$

$\hat{t} := m/M$ satisfies exit condition of backtracking:

$$
\begin{aligned}
f(x + \hat{t}\Delta x) &\underset{(1)}{\leq} f(x) - \frac{m}{M}\lambda(x)^2 + \frac{m}{2M}\lambda(x)^2 \\
&= f(x) - \frac{m}{2M}\lambda(x)^2 \\
&\underset{\alpha \leq \frac{1}{2}}{\leq} f(x) - \alpha \hat{t} \lambda(x)^2
\end{aligned}
$$

and thus stepsize

$$
t \geq \beta \frac{m}{M} \tag{2}
$$

Convergence / Damped Phase / Proof (2/2)

$$
\begin{aligned}
f(x^{\text{next}}) - f(x) &\leq -\alpha t \lambda(x)^2 \\
&\underset{(2)}{\leq} -\alpha\beta \frac{m}{M} \lambda(x)^2 \\
&\underset{\text{dec s.c. ii}}{\leq} -\alpha\beta \frac{m}{M^2} ||\nabla f(x)||_2^2 \\
&\underset{||\nabla f(x)||_2 \geq \eta}{\leq} -\alpha\beta \frac{m}{M^2} \eta^2 = -\gamma
\end{aligned}
$$

Convergence / Pure Phase

Theorem (Convergence of Newton's Algorithm / Pure Phase)

Close to the minimum,

(i) *backtracking always selects stepsize* $t = 1$ *and*

(ii) $\nabla f(x)$ *is shrunken quadratically.*

$$\textit{for } ||\nabla f(x)||_2 < \eta : \ ||\nabla f(x^{next})||_2 \leq \frac{L}{2m^2}(||\nabla f(x)||_2)^2$$

$$\textit{with } \eta \leq 3(1-2\alpha)\frac{m^2}{L}$$

(iii) *it stays close to the mimimum.*

$$\textit{for } ||\nabla f(x)||_2 < \eta : \ ||\nabla f(x^{next})||_2 < \eta$$

$$\textit{with } \eta := \min\{1, 3(1-2\alpha)\}\frac{m^2}{L}$$

Convergence / Pure Phase / Proof (1/5)

(a) If the second derivative of a function is bound linearly, then the function is bound by a third order polynomial:

$$g''(t) \leq a + bt \quad \Longrightarrow \quad g(t) \leq g(0) + g'(0)t + \frac{1}{2}at^2 + \frac{1}{6}bt^3$$

seen by simply integrating

$$\begin{aligned} g''(t) &\leq a + bt && \mid \int_0^1 (\ldots)dt \\ g'(t) &\leq g'(0) + at + \frac{1}{2}bt^2 && \mid \int_0^1 (\ldots)dt \\ g(t) &\leq g(0) + g'(0)t + \frac{1}{2}at^2 + \frac{1}{6}bt^3 \end{aligned}$$

Convergence / Pure Phase / Proof (2/5)

(b) polynomial upper bound of the objective in search direction:

$$\begin{aligned}
\tilde{f}(t) &:= f(x + t\Delta x) & \tilde{f}(0) &= f(x)\\
\tilde{f}'(t) &= \Delta x^T \nabla f(x + t\Delta x) & \tilde{f}'(x) &= -\lambda(x)^2\\
\tilde{f}''(t) &= \Delta x^T \nabla^2 f(x + t\Delta x)\Delta x & \tilde{f}''(0) &\underset{\text{dec ii}}{=} \lambda(x)^2
\end{aligned}$$

the second derivative is linearly bounded:

$$\begin{aligned}
\tilde{f}''(t) &\leq \tilde{f}''(0) + |\tilde{f}''(t) - \tilde{f}''(0)|\\
&= \tilde{f}''(0) + |\Delta x^T \nabla^2 f(x + t\Delta x)\Delta x|\\
&\leq \tilde{f}''(0) + ||\nabla^2 f(x + t\Delta x)||\, ||\Delta x||^2\\
&\underset{II}{\leq} \tilde{f}''(0) + tL||\Delta x||^3 \underset{\text{dec s.c. i}}{\leq} \tilde{f}''(0) + L\frac{1}{m^{\frac{3}{2}}}\lambda(x)^3 t
\end{aligned}$$

$$\underset{a}{\leadsto} \quad \tilde{f}(t) \leq f(x) - \lambda(x)^2 t + \frac{1}{2}\lambda(x)^2 t^2 + \frac{L}{6m^{\frac{3}{2}}}\lambda(x)^3 t^3$$

Note: Remember: $g(t) \leq g(0) + g'(0)t + \frac{1}{2}at^2 + \frac{1}{6}bt^3$.

Convergence / Pure Phase / Proof (3/5)

(i) show backtracking accepts stepsize $t = 1$, if $\eta \leq 3(1-2\alpha)\frac{m^2}{L}$

$$
\begin{aligned}
f(x+\Delta x) = \tilde{f}(1) &\underset{b}{\leq} f(x) - \lambda(x)^2 + \frac{1}{2}\lambda(x)^2 + \frac{L}{6m^{\frac{3}{2}}}\lambda(x)^3 \\
&= f(x) - \lambda(x)^2(\frac{1}{2} - \frac{L}{6m^{\frac{3}{2}}}\lambda(x)) \\
&\underset{\text{dec s.c ii}}{\leq} f(x) - \lambda(x)^2(\frac{1}{2} - \frac{L}{6m^2}||\nabla f(x)||) \\
&\underset{\text{close to min.}}{\leq} f(x) - \lambda(x)^2(\frac{1}{2} - \frac{L}{6m^2}3(1-2\alpha)\frac{m^2}{L}) \\
&= f(x) - \lambda(x)^2\alpha
\end{aligned}
$$

i.e., stepsize $t = 1$ fulfils the exit condition.

Note: close to minimum: $||\nabla f(x)|| \leq \eta \leq 3(1-2\alpha)\frac{m^2}{L}$.

Convergence / Pure Phase / Proof (4/5)

(ii) show decrease in $\nabla f(x^{\text{next}})$:

$$
\begin{aligned}
||\nabla f(x^{\text{next}})||_2 &\underset{t=1}{=} ||\nabla f(x+\Delta x)||_2 \\
&\underset{\text{def } \Delta x}{=} ||\nabla f(x+\Delta x) - \nabla f(x) - \nabla^2 f(x)\Delta x||_2 \\
&\underset{(*)}{=} ||\int_0^1 (\nabla^2 f(x+t\Delta x) - \nabla^2 f(x))\Delta x \; dt||_2 \\
&\leq \int_0^1 ||(\nabla^2 f(x+t\Delta x) - \nabla^2 f(x))||_2 dt \; ||\Delta x||_2 \\
&\underset{\text{II}}{\leq} \int_0^1 Lt||\Delta x||_2 dt ||\Delta x||_2 = \frac{1}{2} L ||\Delta x||_2^2 \\
&\underset{\text{def } \Delta x}{=} \frac{1}{2} L ||\nabla^2 f(x)^{-1} \nabla f(x)||_2^2 \\
&\underset{\text{dec s.c. ii}}{\leq} \frac{L}{2m^2} ||\nabla f(x)||_2^2
\end{aligned}
$$

where (*) $\nabla f(x+\Delta x) = \nabla^2 f(x)\Delta x + \int_0^1 \nabla^2 f(x+t\Delta x)\Delta x \; dt$

Convergence / Pure Phase / Proof (5/5)

(iii) show that Newton stays close to the minimum:

$$||\nabla f(x^{\text{next}})||_2 \underset{ii}{\leq} \frac{L}{2m^2}||\nabla f(x)||_2^2 \leq \frac{L}{2m^2}\eta^2 \underset{\text{def }\eta}{\leq} \frac{1}{2}\eta < \eta$$

Convergence

Theorem (Convergence of Newton's Algorithm)

If

(i) f *is strongly convex* (m, M)*,*

(ii) $\nabla^2 f(x)$ *is Lipschitz-continuous:*
$||\nabla^2 f(y) - \nabla^2 f(x)||_2 \leq L||y - x||_2, \quad L \in \mathbb{R}^+$ *and*

(iii) *backtracking stepsize control is used* $(\alpha \leq \frac{1}{2}, \beta)$

then

$$f(x^{(k)}) - p^* \leq \frac{2m^3}{L^2} \left(\frac{1}{2}\right)^{2^{k-l+1}}, \quad k \geq l$$

$$l := \lceil \frac{f(x^{(0)}) - p^*}{\gamma} \rceil, \quad \gamma := \alpha\beta \frac{m}{M^2}\eta^2, \quad \eta := \min\{1, 3(1 - 2\alpha)\}\frac{m^2}{L}$$

(quadratic convergence)

Convergence / Proof

- If initially we are far away from the minimum, latest after l steps we must be close (damped phase ii) and then

$$\frac{L}{2m^2}||\nabla f(x^{(l)})||_2 \leq \frac{L}{2m^2}\eta \leq \frac{L}{2m^2}\frac{m^2}{L} \leq \frac{1}{2} \tag{1}$$

- In the pure phase $k > l$ we have (pure phase ii)

$$\begin{aligned}
\frac{L}{2m^2}||\nabla f(x^{(k)})||_2 &\leq (\frac{L}{2m^2}||\nabla f(x^{(k-1)})||_2)^2 \underset{\text{rec}}{\leq} (\frac{L}{2m^2}||\nabla f(x^{(l)})||_2)^{2^{k-l}} \\
&\underset{(1)}{\leq} (\frac{1}{2})^{2^{k-l}} \quad \rightsquigarrow \quad ||\nabla f(x^{(k)})||_2 \leq \frac{2m^2}{L}(\frac{1}{2})^{2^{k-l}}
\end{aligned} \tag{2}$$

$$\begin{aligned}
f(x^{(k)}) - p^* &\underset{\text{s.c. i}}{\leq} \frac{1}{2m}||\nabla f(x^{(k)})||_2^2 \underset{(2)}{\leq} \frac{1}{2m}\left(\frac{2m^2}{L}(\frac{1}{2})^{2^{k-l}}\right)^2 \\
&= \frac{2m^3}{L^2}(\frac{1}{2})^{2^{k-l+1}}
\end{aligned}$$

Summary

- Newton's method approximates the objective function by means of a quadratic truncated Taylor expansion around last iterate $x^{(k)}$.

$$\hat{f}(x) = f_0 + g_0^T(x - x_0) + \frac{1}{2}(x - x_0)^T H_0 \, (x - x_0)$$

  - requires current position $x_0 := x^{(k)}$, function value $f_0 := f(x^{(k)})$, gradient $g_0 := \nabla f(x^{(k)})$ and Hessian $H_0 := \nabla^2 f(x^{(k)})$

- Newton's method is a descent method where the descent direction called Newton step Δx is computed as solution of a linear system of equations:

$$H_0 \Delta x = -g_0$$

  - Newton step is affine invariant.

Summary (2/2)

- Newton's method works very well for many problems.
  - requires objective to be twice differentiable.
  - but often too slow for high-dimensional problems (with many variables)
    - as Hessian has size N^2 and solving for the Newton step is $O(N^3)$
- Convergence of Newton's method decomposes in two phases:
  - damped phase:
    - far away from the minimum
    - requires step length control
    - f reduced by at least a constant per step
  - pure phase:
    - close to the minimum
    - always stepsize 1 can be chosen
    - f-distance to minimum shrinks double exponentially in the number of steps
      ($(\frac{1}{2})^{2^k}$; quadratic convergence).

Further Readings

- Newton's method including convergence proof
  - Boyd and Vandenberghe, 2004, ch. 9.5

Acknowledgement: Thanks to John Rothman for pointing out several typos in an earlier version of these slides.

References

Boyd, Stephen and Lieven Vandenberghe (2004). *Convex Optimization*. Cambridge University Press.