Chapter 7: Gradient-based Methods

Chao Wang

SUSTech

$\max Z = -2x_1 - 3x_2 - x_3$

s.t. $x_1 + 4x_2 + 2x_3 - x_4 = 8$

$3x_1 + 2x_2 - x_5 = 6$

$x_1, x_2, x_3, x_4, x_5 \geq 0.$

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	b
x_6	1	4	2	-1	0	1	0	8
x_7	3	2	0	0	-1	0	1	6
Z	-2	-3	-1	0	0	M	M	0

⇒

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	b.
x_6	1*	4	2	-1	0	1	0	8
x_7	3	2	0	0	-1	0	1	6
Z	-2-4M	-3-6M	-1-2M	M	M	0	0	-14M.

x_1 作为入基：

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	b
x_1	1	4	2	-1	0	1	0	8
x_7	0	-10	-6	3	-1	-3	1	-18
Z	0	10M+5	6M+3	-3M-2	M	4M+2	0	18M+16

Outline

1. Gradient Descent
   - Framework
   - Step size strategies
   - Analysis of gradient method
2. Newton's Method
   - Properties of Newton's method
   - Framework
   - Convergence analysis
   - Quasi-Newton Method

梯度下降法

Gradient Descent

Unconstrained minimization

- Unconstrained problem: minimize $f(x)$
  - f convex, twice continuously differentiable
  - we assume optimal value $p^\star = \inf_x f(x)$ is attained (and finite)
- Unconstrained minimization methods
  - produce sequence of points $x^{(k)} \in \operatorname{dom} f$, $k = 0, 1, \ldots$, with

$$f\left(x^{(k)}\right) \to p^\star$$

  - can be interpreted as iterative methods for solving optimality condition

$$\nabla f\left(x^*\right) = 0$$

Descent methods

- $x^{(k+1)} = x^{(k)} + t^{(k)}\Delta x^{(k)}$ with $f\left(x^{(k+1)}\right) < f\left(x^{(k)}\right)$
- Other notations: $x^+ = x + t\Delta x, \quad x := x + t\Delta x$
- Δx is the step, or search direction; t is the step size, or step length
- By first-order Taylor expansion:

$$\begin{aligned} f\left(x^{(k)} + t\Delta x\right) &= f\left(x^{(k)}\right) + t\nabla f\left(x^{(k)}\right)^T \Delta x + o(t) \\ &= f(x^{(k)}) + t\left(\nabla f\left(x^{(k)}\right)^T \Delta x + \frac{o(t)}{t}\right) \end{aligned}$$

- Suppose that t is small so that $\frac{o(t)}{t}$ is negligible. If $\nabla f\left(x^{(k)}\right)^T \Delta x < 0$, we have for a range of sufficiently small stepsizes,

$$f\left(x^{(k)} + t\Delta x\right) \approx f(x) + t\nabla f\left(x^{(k)}\right)^T \Delta x < f\left(x^{(k)}\right)$$

- Given a differential function f and $x \in \mathbb{R}^N$, we called d as a descent direction if d satisfies $\nabla f(x)^T d < 0$.

Gradient descent method

$-\|\nabla f(x)\|\|d\| \le |\nabla f(x)^T d| \le \|\nabla f(x)\|\|d\|$

"=": $d = -\nabla f(x)$

minimize

- Uses $\Delta x = -\nabla f(x)$
- Known as Gradient Descent

$$x^{(k+1)} = x^{(k)} - t^{(k)} \nabla f\left(x^{(k)}\right)$$

- Stepsize Choices
  - exact line search: $t^{(k)} = \text{argmin}_t f\left(x^{(k)} - t\nabla f\left(x^{(k)}\right)\right)$
  - fixed: $t^{(k)}$ constant
  - backtracking line search (most practical)

Gradient descent interpretation

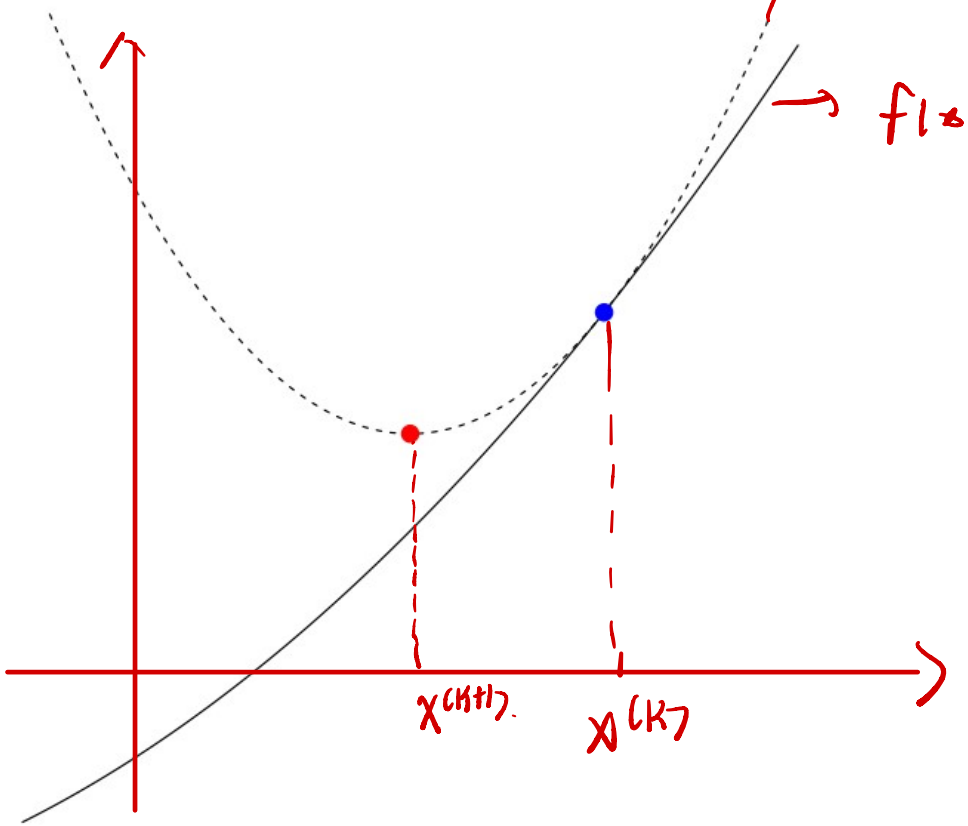
- Quadratic approximation, replacing usual Hessian $\nabla^2 f(x)$ by $\frac{1}{t} I$
- At each iteration, the 2nd-order Taylor expansion becomes

$$f(y) \approx f(x) + \nabla f(x)^T (y - x) + \frac{1}{2t} \|y - x\|_2^2$$

- This is a linear approximation to f and a proximity term to x with weight $\frac{1}{2t}$
- Choose next point $y = x^+$ to minimize quadratic approximation:

$$x^+ = x - t \nabla f(x)$$

Gradient descent interpretation

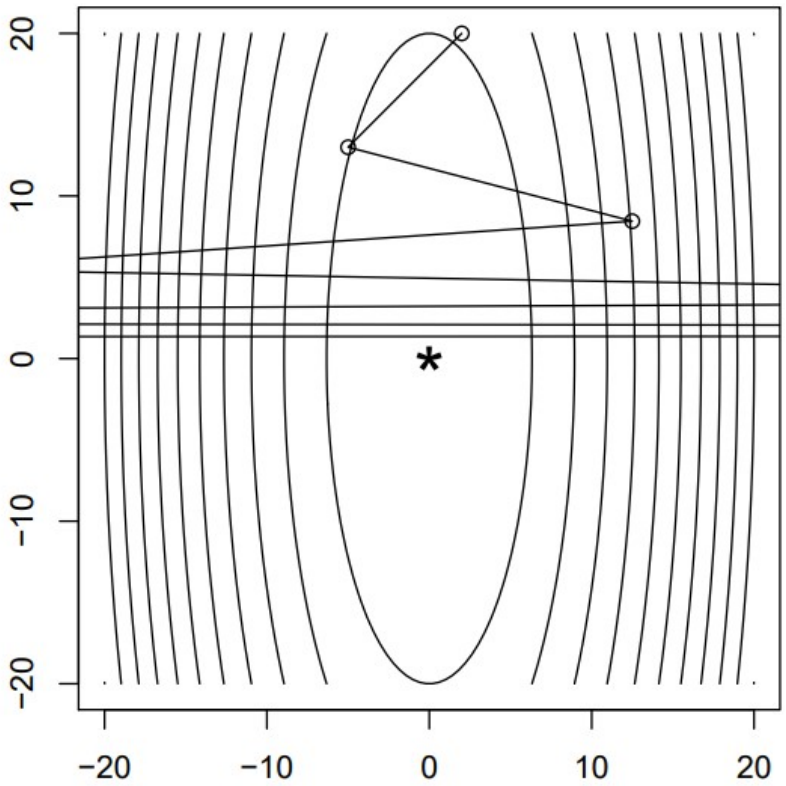


Blue point is x, red point is

$$x^+ = \underset{y}{\operatorname{argmin}} \left\{ f(x) + \nabla f(x)^T (y - x) + \frac{1}{2t} \|y - x\|_2^2 \right\}.$$

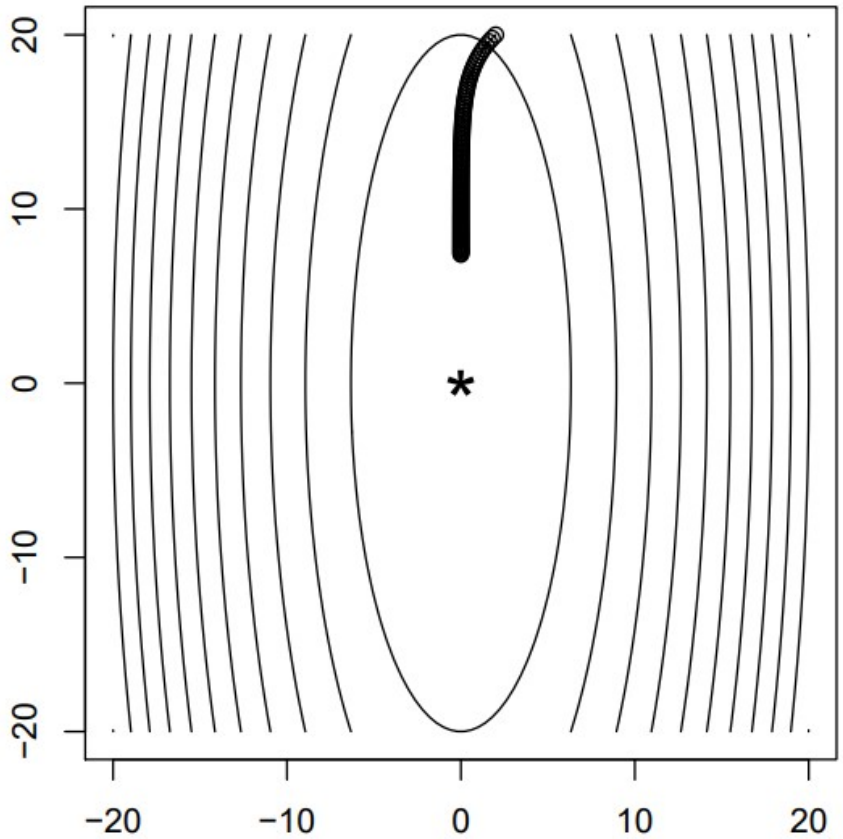
Fixed step size

Simply take $t_k = t$ for all $k = 1, 2, 3, \ldots$, can diverge if t is too big.
Consider $f(x) = \left(10x_1^2 + x_2^2\right)/2$, gradient descent after 8 steps:



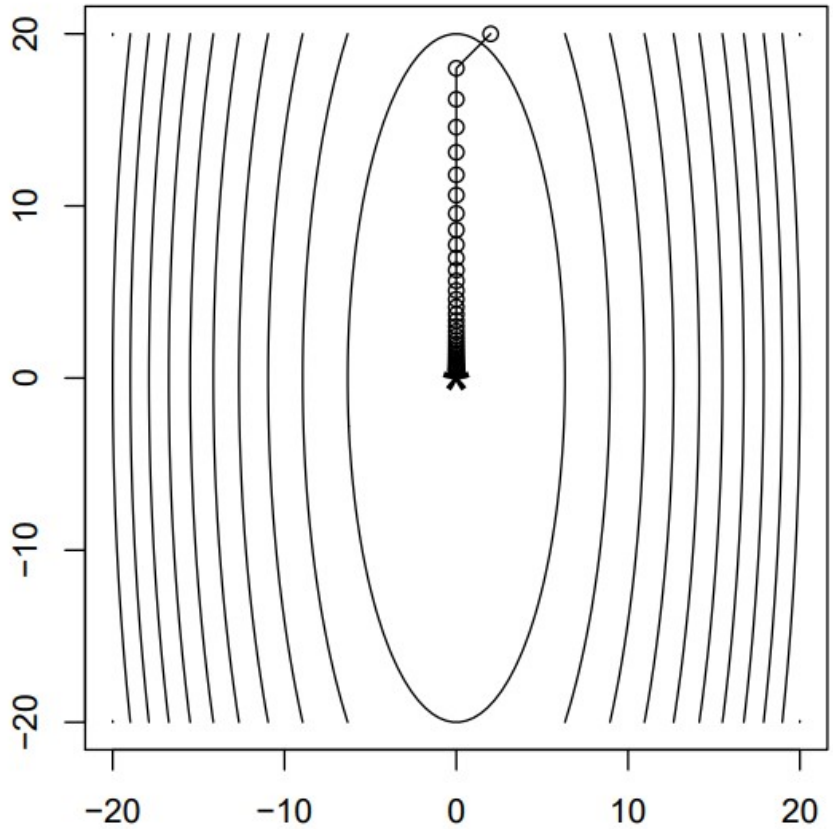
Fixed step size

Can be slow if t is too small. Same example, gradient descent after 100 steps:



Fixed step size

Converges nicely when t is "just right". Same example, 40 steps:



合理控制步长:

t太大: 振荡

t太小: 收敛太慢.

Exact line search

We could also choose step to do the best we can along direction of negative gradient, called exact line search:

$$t = \underset{s \geq 0}{\operatorname{argmin}} f(x - s\nabla f(x))$$

Usually not possible to do this minimization exactly
Approximations to exact line search are typically not as efficient as backtracking, and it's typically not worth it

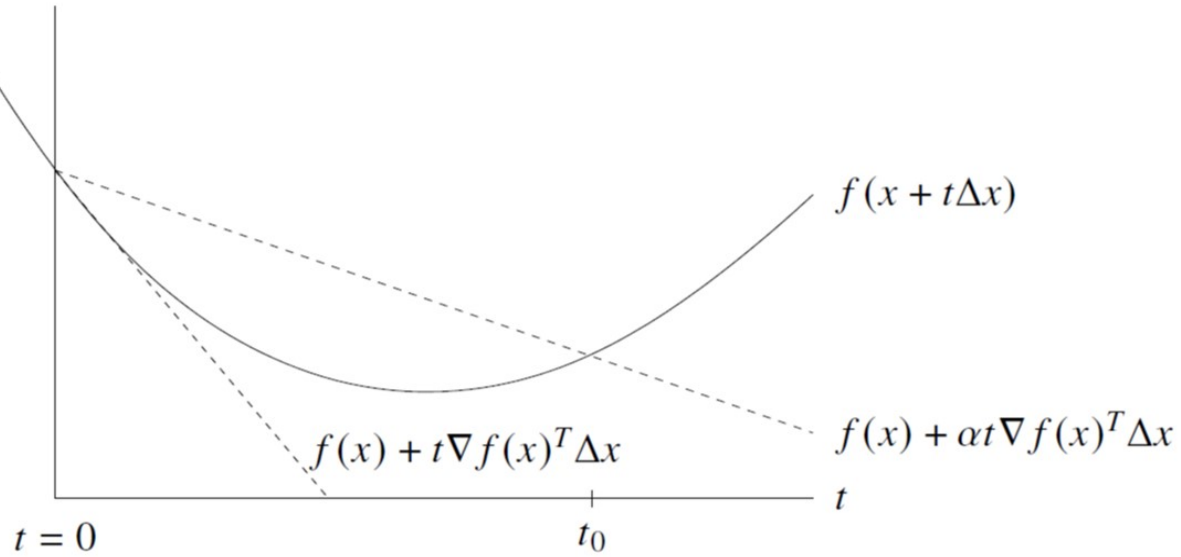
$$\frac{1}{t}(y - x) = -\nabla f(x)^T$$
$$y = x - t\nabla f(x).$$

Backtracking line search

- Starting at $t = 1$, repeat $t := \beta t$ until

$$f(x + t\Delta x) < f(x) + \alpha t \nabla f(x)^T \Delta x$$

- $0 < \beta < 1, 0 < \alpha \leq 0.5$.
- Graphical interpretation: backtrack until $t \leq t_0$



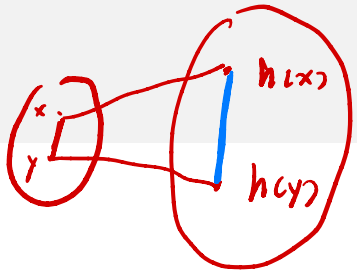
Analysis of gradient method

$$x^{(k+1)} = x^{(k)} - t^{(k)} \nabla f\left(x^{(k)}\right), k = 0, 1, 2, \ldots$$

with fixed step size or backtracking line search assumptions:

- f is convex and differentiable with dom $f = \mathbb{R}^n$
- $\nabla f(x)$ is Lipschitz continuous with parameter $L > 0$
- optimal value $f^* = \inf_x f(x)$ is finite and attained at x^*

Lipschitz continuity



- A function h is called Lipschitz continuous with Lipschitz constant L, if

$$|h(x) - h(y)| \leq L\|x - y\|, \forall x, y \in \text{dom}\, h.$$

- If $\nabla f(x)$ is Lipschitz-continuous with parameter $L > 0$, then we have a quadratic upper bound

$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$ let $L := \sup_x h(x)$

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|x - y\|_2^2, x, y \in \text{dom}\, f$$

- If $\text{dom}\, f = \mathbb{R}^n$ and f has a minimizer x^*, then

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

Strongly convexity

- f is strongly convex with parameter $\mu > 0$ if $f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex
- First-order condition

strong convex:

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|x - y\|_2^2, \forall x, y \in \text{dom}\, f$$

$f(y) \leq f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|x-y\|_2^2$ (∇f has Lipschitz - continuous).

- Second-order condition

$$\nabla^2 f(x) \succeq \mu I, \forall x \in \text{dom}\, f$$

- If $\text{dom}\, f = \mathbb{R}^n$, then f has a minimizer x^*, and

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

Convergence analysis

Theorem (analysis for fixed stepsize)

Assume that f convex and differentiable, with $\mathrm{dom}(f) = \mathbb{R}^n$, and additionally that ∇f is Lipschitz continuous with constant $L > 0$,

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \text{ for any } x, y.$$

Gradient descent with fixed step size $t \leq 1/L$ satisfies

双重约束.

$$f\left(x^{(k)}\right) - f^* \leq \frac{\|x^{(0)} - x^*\|_2^2}{2tk} \leq \varepsilon.$$

and same result holds for backtracking, with t replaced by β/L

Comment: We say gradient descent has convergence rate $O(1/k)$. That is, it finds ϵ-suboptimal point in $O(1/\epsilon)$ iterations

Convergence analysis

Faster convergence rate with the additional assumption of strong convexity

Theorem (analysis for exact line search)

Assume that f convex and differentiable, with $\mathrm{dom}(f) = \mathbb{R}^n$, and additionally that ∇f is Lipschitz continuous with constant $L > 0$ and f is stronlgy convex with parameter μ. Gradient descent with exact line search satisfies $f\left(x^{(k)}\right) - f^* \leq \left(1 - \frac{\mu}{L}\right)^{(k)} \left(f\left(x^0\right) - f^*\right)$

Comment:

- number of iterations to reach $f\left(x^{(k)}\right) - f^* \leq \epsilon$ is

$$\frac{\log\left(\left(f\left(x^0\right) - f^*\right)/\epsilon\right)}{\log(1-\mu/L)^{-1}} \approx \frac{L}{\mu} \log\left(\frac{f\left(x^0\right) - f^*}{\epsilon}\right)$$

- roughly proportional to condition number L/μ when it is large

Proof

Recall from quadratic upper bound

$$f(x - t\nabla f(x)) \leq f(x) - t\left(1 - \frac{Lt}{2}\right)\|\nabla f(x)\|_2^2$$

use $t^+ = \text{argmin}_t\, f(x - t\nabla f(x))$ and $x^+ = x - t^+\nabla f(x)$ to obtain

$$f\left(x^+\right) \leq f\left(x - \frac{1}{L}\nabla f(x)\right) \leq f(x) - \frac{1}{2L}\|\nabla f(x)\|_2^2$$

subtract f^* from both sides

$$f\left(x^+\right) - f^* \leq f(x) - f^* - \frac{1}{2L}\|\nabla f(x)\|_2^2$$

now use strong convexity: $f(x) - f^* \leq \frac{1}{2\mu}\|\nabla f(x)\|_2^2$

$$f\left(x^+\right) - f^* \leq \left(1 - \frac{\mu}{L}\right)\left(f(x) - f^*\right)$$

therefore

$$f\left(x^{(k)}\right) - f^* \leq \left(1 - \frac{\mu}{L}\right)^{(k)}\left(f\left(x^0\right) - f^*\right)$$

Example

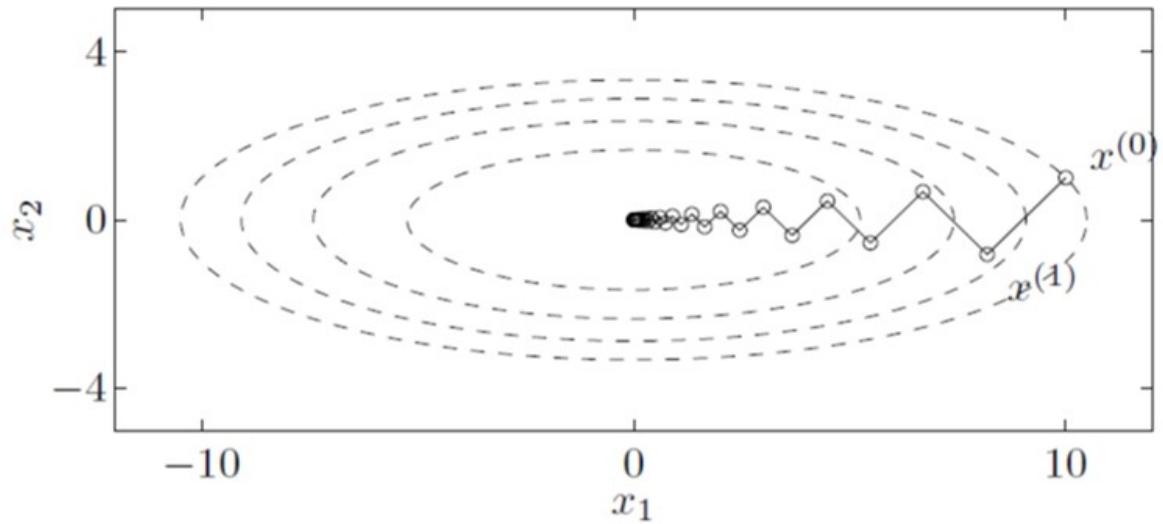
quadratic example

$$f(x) = (1/2)\left(x_1^2 + \gamma x_2^2\right)(\gamma > 1)$$

with exact line search, starting at $x^0 = (\gamma, 1)$

$$x_1^{(k)} = \gamma\left(\frac{\gamma-1}{\gamma+1}\right)^{(k)}, \quad x_2^{(k)} = \left(-\frac{\gamma-1}{\gamma+1}\right)^{(k)}$$

if $\gamma = 10^4, k = 100$, then $\left(\frac{\gamma-1}{\gamma+1}\right)^{(k)} = 0.98$. gradient method can be very slow, and very much dependent on scaling.



Convergence rate

- sublinear rate: $r_k \leq c/k^p$
- linear rate: $r_k \leq c(1-q)^{(k)}$
- quadratic rate: $r_{k+1} \leq cr_k^2$
  r_k can be $f\left(x^{(k)}\right) - f^*$, $\left\|x^{(k)} - x^*\right\|_2$, or $\left\|\nabla f\left(x^{(k)}\right)\right\|_2$; c is some constant

Newton's Method

Newton's method

- Use $\Delta x_{\mathrm{nt}} = -\nabla^2 f(x)^{-1} \nabla f(x)$
- (Pure) Newton's method

$$x^{(k+1)} = x^{(k)} - \nabla^2 f\left(x^{(k)}\right)^{-1} \nabla f\left(x^{(k)}\right)$$

- Damped Newton method

$$x^{(k+1)} = x^{(k)} - t_k \nabla^2 f\left(x^{(k)}\right)^{-1} \nabla f\left(x^{(k)}\right)$$

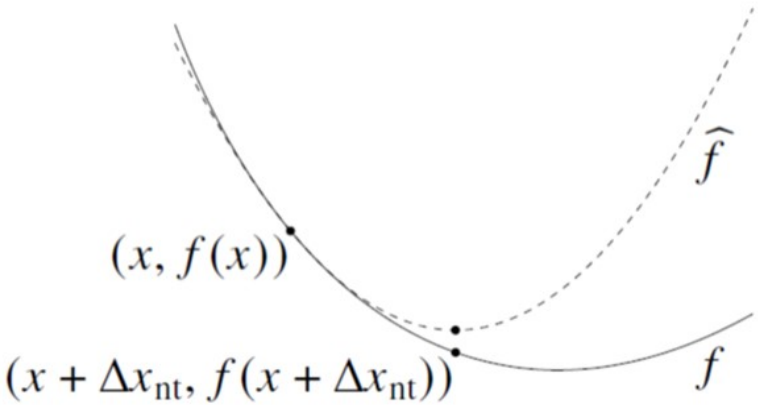
- Interpretations:
  - $x + \Delta x_{\mathrm{nt}}$ minimizes second order approximation

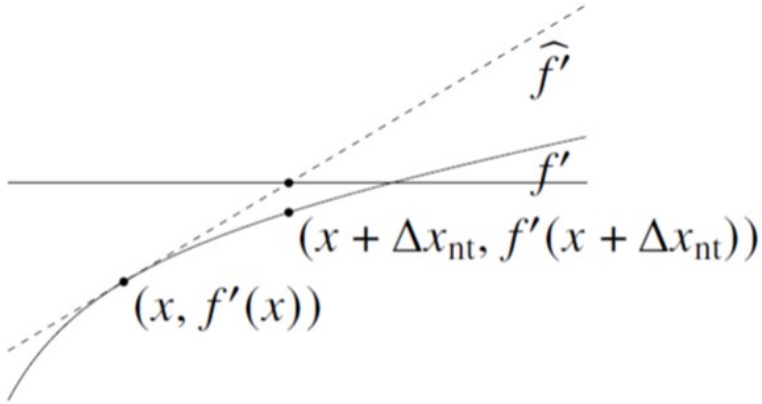
$$\widehat{f}(x+v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v$$

  - $x + \Delta x_{\mathrm{nt}}$ solves linearized optimality condition

$$\nabla f(x+v) \approx \nabla \widehat{f}(x+v) = \nabla f(x) + \nabla^2 f(x) v = 0$$

Interpretations





Affine invariance of Newton's method

- Important property Newton's method: affine invariance.
- Given f, nonsingular $A \in \mathbb{R}^{n \times n}$. Let $x = Ay$, and $g(y) = f(Ay)$.
- Newton steps on g are

$$\begin{aligned} y^+ &= y - \left(\nabla^2 g(y)\right)^{-1} \nabla g(y) \\ &= y - \left(A^T \nabla^2 f(Ay) A\right)^{-1} A^T \nabla f(Ay) \\ &= y - A^{-1} \left(\nabla^2 f(Ay)\right)^{-1} \nabla f(Ay) \end{aligned}$$

- Hence

$$Ay^+ = Ay - \left(\nabla^2 f(Ay)\right)^{-1} \nabla f(Ay)$$

  i.e.,

$$x^+ = x - \left(\nabla^2 f(x)\right)^{-1} \nabla f(x)$$

- So progress is independent of problem scaling. This is not true of gradient descent.

Newton decrement

- At a point x, we define the Newton decrement as

$$\lambda(x) = \left(\nabla f(x)^T \left(\nabla^2 f(x) \right)^{-1} \nabla f(x) \right)^{1/2}$$

- This relates to the difference between $f(x)$ and the minimum of its quadratic approximation:

$$\begin{aligned} & f(x) - \min_y \left(f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} (y - x)^T \nabla^2 f(x) (y - x) \right) \\ &= f(x) - \left(f(x) - \frac{1}{2} \nabla f(x)^T \left(\nabla^2 f(x) \right)^{-1} \nabla f(x) \right) \\ &= \frac{1}{2} \lambda(x)^2 \end{aligned}$$

- Directional derivative in the Newton direction: $\nabla f(x)^T \Delta x_{\text{nt}} = -\lambda(x)^2$

Newton's method

- Given: a starting point $x \in \mathrm{dom}\, f$, tolerance $\epsilon > 0$
- Repeat:
  1. compute the Newton step and decrement

$$\Delta x_{\mathrm{nt}} := -\nabla^2 f(x)^{-1} \nabla f(x); \lambda^2 := \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)$$

  2. stopping criterion: quit if $\lambda^2/2 \leq \epsilon$ 停止条件. $\|\nabla f(x)\|_2 < \varepsilon$.
  3. line search: choose step size t by backtracking line search
  4. update: $x := x + t\Delta x_{\mathrm{nt}}$

Convergence analysis

Theorem

Assumptions

- f *strongly convex on* S *with constant* m
- $\nabla^2 f$ *is Lipschitz continuous on* S*, with constant* $L > 0$ *:*

$$\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\|_2 \leq L\|x - y\|_2$$

(L *measures how well* f *can be approximated by a quadratic function)*

Conclusions *: there exist constants* $\eta \in \left(0, m^2/L\right), \gamma > 0$ *such that*

- *if* $\|\nabla f(x)\|_2 \geq \eta$*, then* $f\left(x^{(k+1)}\right) - f\left(x^{(k)}\right) \leq -\gamma$

依赖起始点.

- *if* $\|\nabla f(x)\|_2 < \eta$*, then*

$$\frac{L}{2m^2}\left\|\nabla f\left(x^{(k+1)}\right)\right\|_2 \leq \left(\frac{L}{2m^2}\left\|\nabla f\left(x^{(k)}\right)\right\|_2\right)^2$$

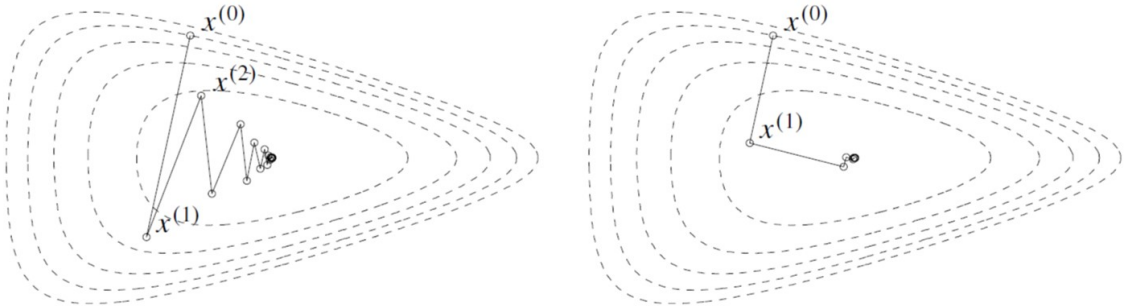
Convergence analysis

- Damped Newton phase ($\|\nabla f(x)\|_2 \geq \eta$)
  - most iterations require backtracking steps
  - function value decreases by at least γ
  - if $f^* > -\infty$, this phase ends after at most $\left(f\left(x^{(0)}\right) - f^*\right)/\gamma$ iterations
- Quadratically convergent phase ($\|\nabla f(x)\|_2 < \eta$)
  -
  - all iterations use step size $t = 1$
  - $\|\nabla f(x)\|_2$ converges to zero quadratically: if $\left\|\nabla f\left(x^{(k)}\right)\right\|_2 < \eta$, then

$$\frac{L}{2m^2}\left\|\nabla f\left(x^{l}\right)\right\|_2 \leq \left(\frac{L}{2m^2}\left\|\nabla f\left(x^{(k)}\right)\right\|_2\right)^{2^{l-k}} \leq \left(\frac{1}{2}\right)^{2^{l-k}}, l \geq k$$

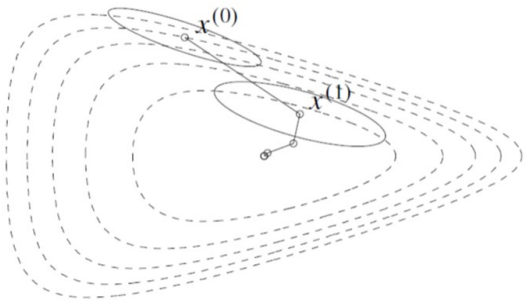
Example

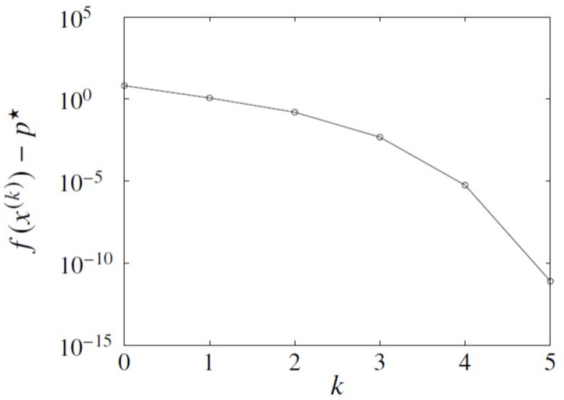
- Nonquadratic example:
  $f(x_1, x_2) = e^{x_1+3x_2-0.1} + e^{x_1-3x_2-0.1} + e^{-x_1-0.1}$
- Gradient descent:



backtracking line search exact line search

- Newton's method (backtracking parameter $\alpha = 0.1, \beta = 0.7$; converges in only 5 steps):





Pros and cons on Newton's method

- Advantages: fast convergence, affine invariance
  - affine invariant means: independent of linear changes of coordinates
  - for example: Newton iterates for $\tilde{f}(y) = f(Ty)$ with starting point $y^0 = T^{-1}x^0$ are $y^{(k)} = T^{-1}x^{(k)}$
- Disadvantages: requires second derivatives, solution to linear equation. It can be too expensive for large-scale applications

Quasi-Newton Method

- It is too expensive to compute $\nabla^2 f\left(x^{(k)}\right)$, use some B_k to replace it.
- Use a quadratic model to approximate $f(x)$ locally at $x^{(k)}$:

$$m_k(p) = f\left(x^{(k)}\right) + \nabla f\left(x^{(k)}\right)^\top p + \frac{1}{2} p^\top B_k p$$

  here B_k is an $n \times n$ symmetric positive definite matrix.
- Note that the function value and gradient of this model at $p = 0$ match $f\left(x^{(k)}\right)$ and $\nabla f\left(x^{(k)}\right)$, respectively.
- By minimizing this quadratic approximation, we obtain

$$p_k = -B_k^{-1} \nabla f\left(x^{(k)}\right)$$

  then we update the iterate via

$$x^{(k+1)} = x^{(k)} + \alpha_k p_k$$

Secant equation

- When we are at $x^{(k+1)}$, we want to construct

$$m_{k+1}(p) = f\left(x^{(k+1)}\right) + \nabla f\left(x^{(k+1)}\right)^\top p + \frac{1}{2} p^\top B_{k+1} p$$

- We want the gradient of m_{k+1} to match the gradient of f at $x^{(k)}$ and $x^{(k+1)}$.

$$\begin{aligned} \nabla m_{k+1}(0) &= \nabla f\left(x^{(k+1)}\right), \\ \nabla m_{k+1}\left(-\alpha_k p_k\right) &= \nabla f\left(x^{(k+1)}\right) - \alpha_k B_{k+1} p_k = \nabla f\left(x^{(k)}\right) \end{aligned}$$

- We have,

$$B_{k+1}\alpha_k p_k = \nabla f\left(x^{(k+1)}\right) - \nabla f\left(x^{(k)}\right)$$

- To simplify the notation, define

$$s_k = x^{(k+1)} - x^{(k)} = \alpha_k p_k, y_k = \nabla f\left(x^{(k+1)}\right) - \nabla f\left(x^{(k)}\right)$$

then we get

$$B_{k+1} s_k = y_k$$

and this is called secant equation.

DFP and BFGS

- To compute B_{k+1}, we solve

$$\begin{aligned} \min \quad & \|B - B_k\| \\ \text{s.t.}, \quad & B = B^\top, \quad Bs_k = y_k \end{aligned}$$

- This gives the following DFP updating formula (originally given by Davidon in 1959, and subsequently studied by Fletcher and Powell)

$$B_{k+1} = \left(I - \rho_k y_k s_k^\top\right) B_k \left(I - \rho_k s_k y_k^\top\right) + \rho_k y_k y_k^\top$$

with $\rho_k = 1/y_k^\top s_k$.

DFP and BFGS

- The other way to compute B_{k+1} : denote its inverse as H_{k+1}, solve the following problem

$$\begin{aligned} \min \quad & \|H - H_k\| \\ \text{s.t.}, \quad & H = H^\top, \quad Hy_k = s_k \end{aligned}$$

- This gives the following BFGS updating formula (proposed by Broyden, Fletcher, Goldfarb and Shanno, independently)

$$H_{k+1} = \left(I - \rho_k s_k y_k^\top\right) H_k \left(I - \rho_k y_k s_k^\top\right) + \rho_k s_k s_k^\top$$

or, by Sherman-Morrison-Woodbury formula:

$$B_{k+1} = B_k - \frac{B_k s_k s_k^\top B_k}{s_k^\top B_k s_k} + \frac{y_k y_k^\top}{y_k^\top s_k}$$

Convergence on BFGS method

- **Global convergence**
  if f is strongly convex, then BFGS with backtracking line search converges to the optimum for any x^0 and $B_0 \succ 0$
- **Local convergence**
  if f is strongly convex and $\nabla^2 f(x)$ is Lipschitz continuous, then local convergence is superlinear: for sufficiently large k,

$$\left\| x^{(k+1)} - x^* \right\|_2 \leq c_k \left\| x^{(k)} - x^* \right\|_2$$

  where $c_k \to 0$.