An introduction to convex methods for life science
Unconstrained minimization for smooth convex problems

Math et Sciences du Vivant – Université Paris-Saclay / Paris-Sud

Autumn semester 2017

http://julien.cremeriefamily.info

References

See Chapter 9 in

Convex Optimization,
Stephen Boyd and Lieve Lieven Vandenberghe
https://web.stanford.edu/~boyd/cvxbook/

All slides stolen (extracted/re-arranged) from Lieve Vandenberghe:

- Convex Optimization:
  http://www.seas.ucla.edu/~vandenbe/ee236b/ee236b.html
- Optimization Methods for Large-Scale Systems
  http://www.seas.ucla.edu/~vandenbe/ee236c/ee236c.html

Outline

Outline

Unconstrained minimization

$$\text{minimize} \quad f(x)$$

- f convex, twice continuously differentiable (hence $\mathbf{dom}\, f$ open)
- we assume optimal value $p^\star = \inf_x f(x)$ is attained (and finite)

unconstrained minimization methods

- produce sequence of points $x^{(k)} \in \mathbf{dom}\, f$, $k = 0, 1, \ldots$ with

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

- can be interpreted as iterative methods for solving optimality condition

$$\nabla f(x^\star) = 0$$

Initial point and sublevel set

algorithms in this chapter require a starting point $x^{(0)}$ such that

- $x^{(0)} \in \mathbf{dom}\, f$
- sublevel set $S = \{x \mid f(x) \leq f(x^{(0)})\}$ is closed

2nd condition is hard to verify, except when *all* sublevel sets are closed:

- equivalent to condition that $\mathbf{epi}\, f$ is closed
- true if $\mathbf{dom}\, f = \mathbf{R}^n$
- true if $f(x) \to \infty$ as $x \to \mathbf{bd}\,\mathbf{dom}\, f$

examples of differentiable functions with closed sublevel sets:

$$f(x) = \log(\sum_{i=1}^{m} \exp(a_i^T x + b_i)), \qquad f(x) = -\sum_{i=1}^{m} \log(b_i - a_i^T x)$$

Outline

Monotonicity of gradient

a differentiable function f is convex if and only if $\operatorname{dom} f$ is convex and

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq 0 \quad \text{for all } x, y \in \operatorname{dom} f$$

i.e., the gradient $\nabla f : \mathbf{R}^n \to \mathbf{R}^n$ is a *monotone* mapping

a differentiable function f is strictly convex if and only if $\operatorname{dom} f$ is convex and

$$(\nabla f(x) - \nabla f(y))^T (x - y) > 0 \quad \text{for all } x, y \in \operatorname{dom} f,\ x \neq y$$

i.e., the gradient $\nabla f : \mathbf{R}^n \to \mathbf{R}^n$ is a *strictly monotone* mapping

Proof

- if f is differentiable and convex, then

$$f(y) \geq f(x) + \nabla f(x)^T(y-x), \qquad f(x) \geq f(y) + \nabla f(y)^T(x-y)$$

  combining the inequalities gives $(\nabla f(x) - \nabla f(y))^T(x-y) \geq 0$

- if ∇f is monotone, then $g'(t) \geq g'(0)$ for $t \geq 0$ and $t \in \operatorname{dom} g$, where

$$g(t) = f(x + t(y-x)), \qquad g'(t) = \nabla f(x + t(y-x))^T(y-x)$$

  hence

$$\begin{array}{rcl} f(y) = g(1) = g(0) + \displaystyle\int_0^1 g'(t)\, dt & \geq & g(0) + g'(0) \\ & = & f(x) + \nabla f(x)^T(y-x) \end{array}$$

  this is the first-order condition for convexity

Lipschitz continuous gradient

the gradient of f is *Lipschitz continuous* with parameter $L > 0$ if

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \text{for all } x, y \in \operatorname{dom} f$$

- note that the definition does not assume convexity of f
- we will see that for convex f with $\operatorname{dom} f = \mathbf{R}^n$, this is equivalent to

$$\frac{L}{2}x^T x - f(x) \quad \text{is convex}$$

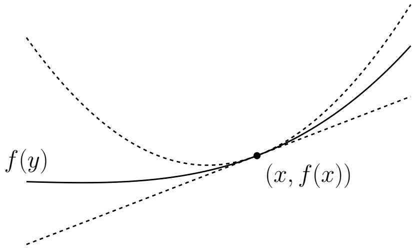
(*i.e.*, if f is twice differentiable, $\nabla^2 f(x) \preceq LI$ for all x)

Quadratic upper bound

suppose ∇f is Lipschitz continuous with parameter L and $\mathrm{dom}\, f$ is convex

- then $g(x) = (L/2)x^T x - f(x)$, with $\mathrm{dom}\, g = \mathrm{dom}\, f$, is convex
- convexity of g is equivalent to a quadratic upper bound on f:

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}\|y - x\|_2^2 \quad \text{for all } x, y \in \mathrm{dom}\, f$$



Proof

- Lipschitz continuity of ∇f and the Cauchy-Schwarz inequality imply

$$(\nabla f(x) - \nabla f(y))^T (x - y) \leq L\|x - y\|_2^2 \quad \text{for all } x, y \in \operatorname{dom} f$$

  this is monotonicity of the gradient

$$\nabla g(x) = Lx - \nabla f(x)$$

- hence, g is a convex function if its domain $\operatorname{dom} g = \operatorname{dom} f$ is convex
- the quadratic upper bound is the first-order condition for convexity of g

$$g(y) \geq g(x) + \nabla g(x)^T (y - x) \quad \text{for all } x, y \in \operatorname{dom} g$$

Strongly convex function

f is *strongly convex* with parameter $m > 0$ if

$$g(x) = f(x) - \frac{m}{2}x^T x \quad \text{is convex}$$

Jensen's inequality: Jensen's inequality for g is

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y) - \frac{m}{2}\theta(1-\theta)\|x-y\|_2^2$$

Monotonicity: monotonicity of ∇g gives

$$(\nabla f(x) - \nabla f(y))^T(x-y) \geq m\|x-y\|_2^2 \quad \text{for all } x, y \in \operatorname{dom} f$$

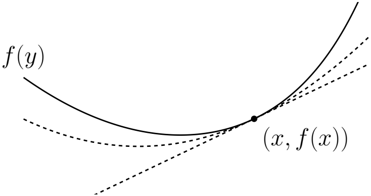
this is called *strong monotonicity (coercivity)* of ∇f

Second-order condition: $\nabla^2 f(x) \succeq mI$ for all $x \in \operatorname{dom} f$

Quadratic lower bound

from 1st order condition of convexity of g:

$$f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}\|y-x\|_2^2 \quad \text{for all } x, y \in \operatorname{dom} f$$



- implies sublevel sets of f are bounded
- if f is closed (has closed sublevel sets), it has a unique minimizer $x^\star$ and

$$\frac{m}{2}\|x-x^\star\|_2^2 \leq f(x) - f(x^\star) \leq \frac{1}{2m}\|\nabla f(x)\|_2^2 \quad \text{for all } x \in \operatorname{dom} f$$

Outline

Outline

Descent methods

$$x^{(k+1)} = x^{(k)} + t^{(k)}\Delta x^{(k)} \quad \text{with } f(x^{(k+1)}) < f(x^{(k)})$$

- other notations: $x^+ = x + t\Delta x$, $x := x + t\Delta x$
- Δx is the *step*, or *search direction*; t is the *step size*, or *step length*
- from convexity, $f(x^+) < f(x)$ implies $\nabla f(x)^T \Delta x < 0$ ($i.e.$, Δx is a *descent direction*)

General descent method.

given a starting point $x \in \mathbf{dom}\, f$.
repeat
1. Determine a descent direction Δx.
2. *Line search.* Choose a step size $t > 0$.
3. *Update.* $x := x + t\Delta x$.

until stopping criterion is satisfied.

Line search types

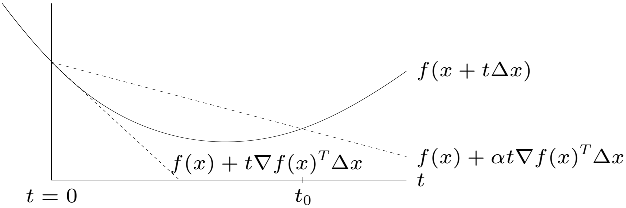
exact line search: $t = \mathrm{argmin}_{t>0} f(x + t\Delta x)$

backtracking line search (with parameters $\alpha \in (0, 1/2)$, $\beta \in (0, 1)$)

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

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

- graphical interpretation: backtrack until $t \leq t_0$



Outline

Gradient descent method

general descent method with $\Delta x = -\nabla f(x)$

---

given a starting point $x \in \textbf{dom} f$.
repeat
1. $\Delta x := -\nabla f(x)$.
2. *Line search.* Choose step size t via exact or backtracking line search.
3. *Update.* $x := x + t\Delta x$.

until stopping criterion is satisfied.

- stopping criterion usually of the form $\|\nabla f(x)\|_2 \leq \epsilon$
- convergence result: for strongly convex f,

$$f(x^{(k)}) - p^\star \leq c^k (f(x^{(0)}) - p^\star)$$

  $c \in (0,1)$ depends on m, $x^{(0)}$, line search type
- very simple, but often very slow; rarely used in practice

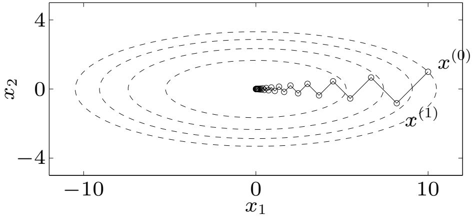
quadratic problem in $\mathbf{R}^2$

$$f(x) = (1/2)(x_1^2 + \gamma x_2^2) \qquad\qquad (\gamma > 0)$$

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

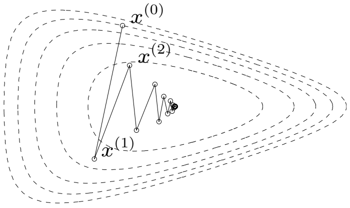
$$x_1^{(k)} = \gamma \left(\frac{\gamma - 1}{\gamma + 1}\right)^k, \qquad x_2^{(k)} = \left(-\frac{\gamma - 1}{\gamma + 1}\right)^k$$

- very slow if $\gamma \gg 1$ or $\gamma \ll 1$
- example for $\gamma = 10$:

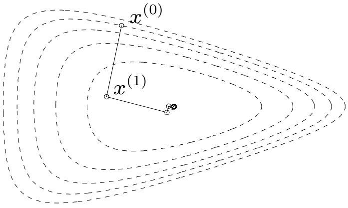


nonquadratic example

$$f(x_1, x_2) = e^{x_1+3x_2-0.1} + e^{x_1-3x_2-0.1} + e^{-x_1-0.1}$$



backtracking line search

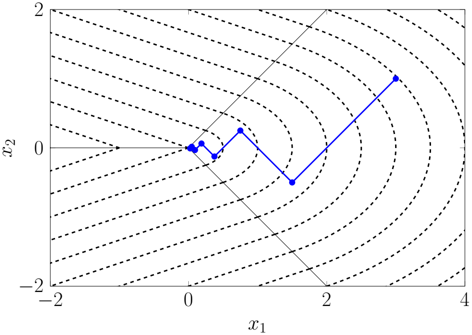


exact line search

Nondifferentiable example

$$f(x) = \sqrt{x_1^2 + \gamma x_2^2} \quad \text{for } |x_2| \le x_1, \qquad f(x) = \frac{x_1 + \gamma |x_2|}{\sqrt{1+\gamma}} \quad \text{for } |x_2| > x_1$$

with exact line search, starting point $x^{(0)} = (\gamma, 1)$, converges to non-optimal point



gradient method does not handle nondifferentiable problems

Outline

Analysis of gradient method

$$x^{(k)} = x^{(k-1)} - t_k \nabla f(x^{(k-1)}), \qquad k = 1, 2, \ldots$$

with fixed step size or backtracking line search

Assumptions

1. f is convex and differentiable with $\mathbf{dom}\, f = \mathbf{R}^n$
2. $\nabla f(x)$ is Lipschitz continuous with parameter $L > 0$
3. optimal value $f^\star = \inf_x f(x)$ is finite and attained at $x^\star$

Analysis for constant step size

- from quadratic upper bound (page 1-12) with $y = x - t\nabla f(x)$:

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

- therefore, if $x^+ = x - t\nabla f(x)$ and $0 < t \le 1/L$,

$$\begin{aligned} f(x^+) &\le f(x) - \frac{t}{2}\|\nabla f(x)\|_2^2 \qquad (1) \\ &\le f^\star + \nabla f(x)^T(x - x^\star) - \frac{t}{2}\|\nabla f(x)\|_2^2 \\ &= f^\star + \frac{1}{2t}\left(\|x - x^\star\|_2^2 - \|x - x^\star - t\nabla f(x)\|_2^2\right) \\ &= f^\star + \frac{1}{2t}\left(\|x - x^\star\|_2^2 - \|x^+ - x^\star\|_2^2\right) \end{aligned}$$

second line follows from convexity of f

- define $x = x^{(i-1)}$, $x^+ = x^{(i)}$, $t_i = t$, and add the bounds for $i = 1, \ldots, k$:

$$\begin{aligned}
\sum_{i=1}^{k} (f(x^{(i)}) - f^\star) &\leq \frac{1}{2t} \sum_{i=1}^{k} \left(\|x^{(i-1)} - x^\star\|_2^2 - \|x^{(i)} - x^\star\|_2^2 \right) \\
&= \frac{1}{2t} \left(\|x^{(0)} - x^\star\|_2^2 - \|x^{(k)} - x^\star\|_2^2 \right) \\
&\leq \frac{1}{2t} \|x^{(0)} - x^\star\|_2^2
\end{aligned}$$

- since $f(x^{(i)})$ is non-increasing (see (1))

$$f(x^{(k)}) - f^\star \leq \frac{1}{k} \sum_{i=1}^{k} (f(x^{(i)}) - f^\star) \leq \frac{1}{2kt} \|x^{(0)} - x^\star\|_2^2$$

Conclusion: number of iterations to reach $f(x^{(k)}) - f^\star \leq \epsilon$ is $O(1/\epsilon)$

Gradient method for strongly convex functions

better results exist if we add strong convexity to the assumptions on p. 1-20

Analysis for constant step size

if $x^+ = x - t\nabla f(x)$ and $0 < t \leq 2/(m+L)$:

$$\begin{aligned}
\|x^+ - x^\star\|_2^2 &= \|x - t\nabla f(x) - x^\star\|_2^2 \\
&= \|x - x^\star\|_2^2 - 2t\nabla f(x)^T(x - x^\star) + t^2\|\nabla f(x)\|_2^2 \\
&\leq (1 - t\frac{2mL}{m+L})\|x - x^\star\|_2^2 + t(t - \frac{2}{m+L})\|\nabla f(x)\|_2^2 \\
&\leq (1 - t\frac{2mL}{m+L})\|x - x^\star\|_2^2
\end{aligned}$$

(step 3 follows from result on p. 1-19)

Distance to optimum

$$\|x^{(k)} - x^\star\|_2^2 \leq c^k \|x^{(0)} - x^\star\|_2^2, \qquad c = 1 - t\frac{2mL}{m+L}$$

- implies (linear) convergence
- for $t = 2/(m+L)$, get $c = \left(\frac{\gamma-1}{\gamma+1}\right)^2$ with $\gamma = L/m$

Bound on function value (from page 1-14)

$$f(x^{(k)}) - f^\star \leq \frac{L}{2}\|x^{(k)} - x^\star\|_2^2 \leq \frac{c^k L}{2}\|x^{(0)} - x^\star\|_2^2$$

Conclusion: number of iterations to reach $f(x^{(k)}) - f^\star \leq \epsilon$ is $O(\log(1/\epsilon))$

Outline

Outline

Newton method for unconstrained minimization

$$\text{minimize} \quad f(x)$$

f convex, twice continously differentiable

Newton method

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

- advantages: fast convergence, affine invariance
- disadvantages: requires second derivatives, solution of linear equation

can be too expensive for large scale applications

Newton step

$$\Delta x_{\rm nt} = -\nabla^2 f(x)^{-1}\nabla f(x)$$

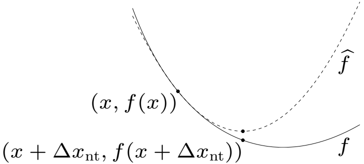
interpretations

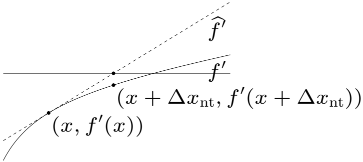
- $x + \Delta x_{\rm nt}$ minimizes second order approximation

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

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

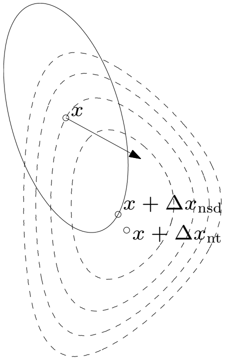
$$\nabla f(x+v) \approx \nabla \widehat{f}(x+v) = \nabla f(x) + \nabla^2 f(x) v = 0$$





- $\Delta x_{\rm nt}$ is steepest descent direction at x in local Hessian norm

$$\|u\|_{\nabla^2 f(x)} = \left(u^T \nabla^2 f(x) u\right)^{1/2}$$



dashed lines are contour lines of f; ellipse is $\{x + v \mid v^T \nabla^2 f(x) v = 1\}$

arrow shows $-\nabla f(x)$

Newton decrement

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

a measure of the proximity of x to $x^\star$

properties

- gives an estimate of $f(x) - p^\star$, using quadratic approximation $\widehat{f}$:

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

- equal to the norm of the Newton step in the quadratic Hessian norm

$$\lambda(x) = \left(\Delta x_{\mathrm{nt}}^T \nabla^2 f(x) \Delta x_{\mathrm{nt}}\right)^{1/2}$$

- directional derivative in the Newton direction: $\nabla f(x)^T \Delta x_{\mathrm{nt}} = -\lambda(x)^2$
- affine invariant (unlike $\|\nabla f(x)\|_2$)

Newton's method

---

given a starting point $x \in \mathbf{dom}\, f$, tolerance $\epsilon > 0$.
repeat

1. *Compute the Newton step and decrement.*
   $\Delta x_{\rm nt} := -\nabla^2 f(x)^{-1}\nabla f(x);\quad \lambda^2 := \nabla f(x)^T \nabla^2 f(x)^{-1}\nabla f(x)$.
2. *Stopping criterion.* **quit** if $\lambda^2/2 \le \epsilon$.
3. *Line search.* Choose step size t by backtracking line search.
4. *Update.* $x := x + t\Delta x_{\rm nt}$.

---

affine invariant, $i.e.$, independent of linear changes of coordinates:

Newton iterates for $\tilde{f}(y) = f(Ty)$ with starting point $y^{(0)} = T^{-1}x^{(0)}$ are

$$y^{(k)} = T^{-1}x^{(k)}$$

Outline

Classical convergence analysis

assumptions

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

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

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

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

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

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

damped Newton phase ($\|\nabla f(x)\|_2 \geq \eta$)

- most iterations require backtracking steps
- function value decreases by at least γ
- if $p^\star > -\infty$, this phase ends after at most $(f(x^{(0)}) - p^\star)/\gamma$ 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 $\|\nabla f(x^{(k)})\|_2 < \eta$, then

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

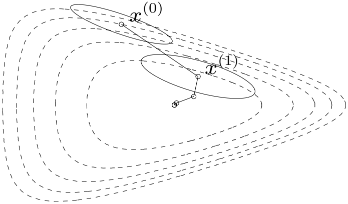
conclusion: number of iterations until $f(x) - p^\star \leq \epsilon$ is bounded above by

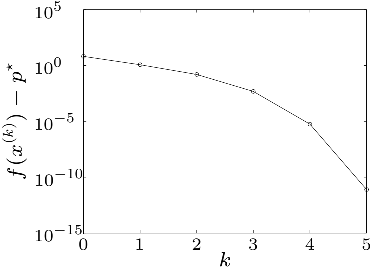
$$\frac{f(x^{(0)}) - p^\star}{\gamma} + \log_2 \log_2(\epsilon_0/\epsilon)$$

- γ, ϵ_0 are constants that depend on m, L, $x^{(0)}$
- second term is small (of the order of 6) and almost constant for practical purposes
- in practice, constants m, L (hence γ, ϵ_0) are usually unknown
- provides qualitative insight in convergence properties ($i.e.$, explains two algorithm phases)

Examples

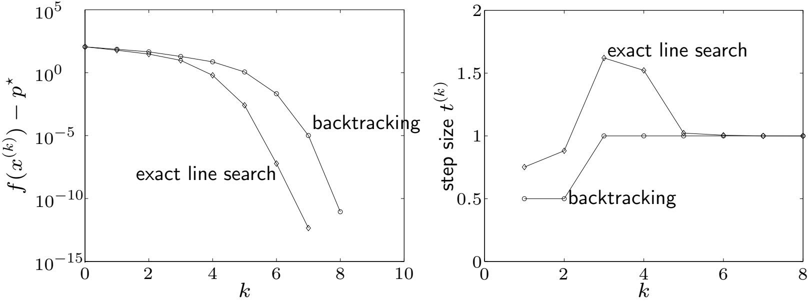
example in $\mathbf{R}^2$ (page 10–9)





- backtracking parameters $\alpha = 0.1$, $\beta = 0.7$
- converges in only 5 steps
- quadratic local convergence

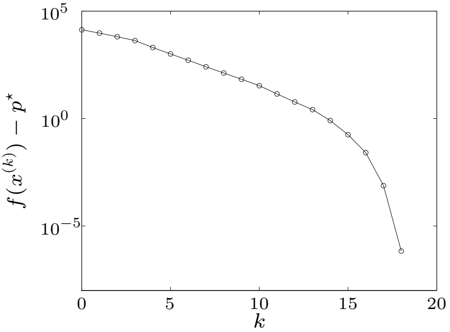
example in $\mathbf{R}^{100}$ (page 10–10)



- backtracking parameters $\alpha = 0.01$, $\beta = 0.5$
- backtracking line search almost as fast as exact l.s. (and much simpler)
- clearly shows two phases in algorithm

example in $\mathbf{R}^{10000}$ (with sparse a_i)

$$f(x) = -\sum_{i=1}^{10000} \log(1 - x_i^2) - \sum_{i=1}^{100000} \log(b_i - a_i^T x)$$



- backtracking parameters $\alpha = 0.01$, $\beta = 0.5$.
- performance similar as for small examples

Outline

Variable metric methods

$$x^+ = x - tH^{-1}\nabla f(x)$$

$H \succ 0$ is approximation of the Hessian at x, chosen to:

- avoid calculation of second derivatives
- simplify computation of search direction

'Variable metric' interpretation (EE236B, lecture 10, page 11)

$$\Delta x = -H^{-1}\nabla f(x)$$

is steepest descent direction at x for quadratic norm

$$\|z\|_H = \left(z^T H z\right)^{1/2}$$

Quasi-Newton methods

given starting point $x^{(0)} \in \operatorname{dom} f$, $H_0 \succ 0$

1. compute quasi-Newton direction $\Delta x = -H_{k-1}^{-1} \nabla f(x^{(k-1)})$
2. determine step size t (*e.g.*, by backtracking line search)
3. compute $x^{(k)} = x^{(k-1)} + t\Delta x$
4. compute H_k

- different methods use different rules for updating H in step 4
- can also propagate H_k^{-1} to simplify calculation of Δx

Broyden-Fletcher-Goldfarb-Shanno (BFGS) update

BFGS update

$$H_k = H_{k-1} + \frac{yy^T}{y^T s} - \frac{H_{k-1} s s^T H_{k-1}}{s^T H_{k-1} s}$$

where

$$s = x^{(k)} - x^{(k-1)}, \qquad y = \nabla f(x^{(k)}) - \nabla f(x^{(k-1)})$$

Inverse update

$$H_k^{-1} = \left(I - \frac{sy^T}{y^T s}\right) H_{k-1}^{-1} \left(I - \frac{ys^T}{y^T s}\right) + \frac{ss^T}{y^T s}$$

- note that $y^T s > 0$ for strictly convex f; see page 1-9
- cost of update or inverse update is $O(n^2)$ operations

Positive definiteness

if $y^T s > 0$, BFGS update preserves positive definitess of H_k

Proof: from inverse update formula,

$$v^T H_k^{-1} v = \left(v - \frac{s^T v}{s^T y} y\right)^T H_{k-1}^{-1} \left(v - \frac{s^T v}{s^T y} y\right) + \frac{(s^T v)^2}{y^T s}$$

- if $H_{k-1} \succ 0$, both terms are nonnegative for all v
- second term is zero only if $s^T v = 0$; then first term is zero only if $v = 0$

this ensures that $\Delta x = -H_k^{-1} \nabla f(x^k)$ is a descent direction

Secant condition

the BFGS update satisfies the *secant condition* $H_k s = y$, *i.e.*,

$$H_k(x^{(k)} - x^{(k-1)}) = \nabla f(x^{(k)}) - \nabla f(x^{(k-1)})$$

Interpretation: define second-order approximation at $x^{(k)}$

$$f_{\text{quad}}(z) = f(x^{(k)}) + \nabla f(x^{(k)})^T (z - x^{(k)}) + \frac{1}{2}(z - x^{(k)})^T H_k (z - x^{(k)})$$

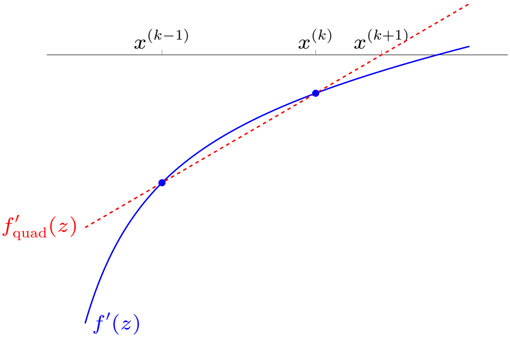
secant condition implies that gradient of f_{quad} agrees with f at $x^{(k-1)}$:

$$\begin{aligned} \nabla f_{\text{quad}}(x^{(k-1)}) &= \nabla f(x^{(k)}) + H_k(x^{(k-1)} - x^{(k)}) \\ &= \nabla f(x^{(k-1)}) \end{aligned}$$

Secant method

for $f : \mathbf{R} \to \mathbf{R}$, BFGS with unit step size gives the secant method

$$x^{(k+1)} = x^{(k)} - \frac{f'(x^{(k)})}{H_k}, \qquad H_k = \frac{f'(x^{(k)}) - f'(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}$$



Convergence

Global result

if f is strongly convex, BFGS with backtracking line search (EE236B, lecture 10-6) converges from any $x^{(0)}$, $H_0 \succ 0$

Local convergence

if f is strongly convex and $\nabla^2 f(x)$ is Lipschitz continuous, local convergence is *superlinear*: for sufficiently large k,

$$\|x^{(k+1)} - x^\star\|_2 \leq c_k \|x^{(k)} - x^\star\|_2 \to 0$$

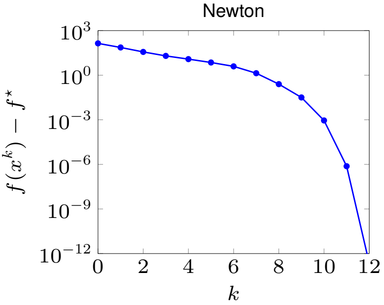
where $c_k \to 0$

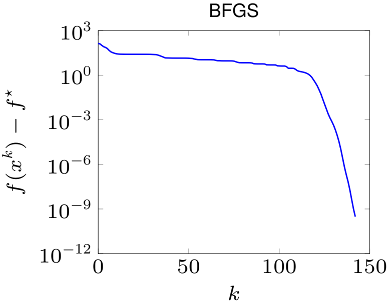
(*cf.*, quadratic local convergence of Newton method)

Example

$$\text{minimize} \quad c^T x - \sum_{i=1}^{m} \log(b_i - a_i^T x)$$

$n = 100$, $m = 500$





- cost per Newton iteration: $O(n^3)$ plus computing $\nabla^2 f(x)$
- cost per BFGS iteration: $O(n^2)$

Limited memory quasi-Newton methods

main disadvantage of quasi-Newton method is need to store H_k or H_k^{-1}

Limited-memory BFGS (L-BFGS): do not store H_k^{-1} explicitly

- instead we store the m (*e.g.*, $m = 30$) most recent values of

$$s_j = x^{(j)} - x^{(j-1)}, \qquad y_j = \nabla f(x^{(j)}) - \nabla f(x^{(j-1)})$$

- we evaluate $\Delta x = H_k^{-1} \nabla f(x^{(k)})$ recursively, using

$$H_j^{-1} = \left(I - \frac{s_j y_j^T}{y_j^T s_j} \right) H_{j-1}^{-1} \left(I - \frac{y_j s_j^T}{y_j^T s_j} \right) + \frac{s_j s_j^T}{y_j^T s_j}$$

  for $j = k, k-1, \ldots, k-m+1$, assuming, for example, $H_{k-m}^{-1} = I$

- cost per iteration is $O(nm)$; storage is $O(nm)$