Chapter 09 Unconstrained minimization

Last update on 2022-05-18 17:47

Table of contents

Unconstrained minimization

unconstrained minimization problem

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

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

optimality condition (review)

$$x^\star \text{ is optimal} \quad \Longleftrightarrow \quad x^\star \in \mathbf{dom}\, f, \quad \nabla f(x^\star) = 0$$

Unconstrained minimization methods

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

$$f(x^{(k)}) \quad \longrightarrow \quad p^*$$

- can be interpreted as iterative methods for solving optimality condition

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

Initial point and sublevel set

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

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

second condition hard to verify, except when all sublevel sets are closed (i.e. f is closed)

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

examples of differentiable functions with closed sublevel sets

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

Strong convexity and implications

f is **strongly convex** on S if there exists an $m > 0$ such that

$$\nabla^2 f(x) \succeq mI \qquad \text{for all} \qquad x \in S$$

implications

- $p^\star > -\infty$
- for $x, y \in S$

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

  hence S is bounded
- for $x \in S$

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

  useful as stopping criterion (if you know m)

Descent methods

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

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

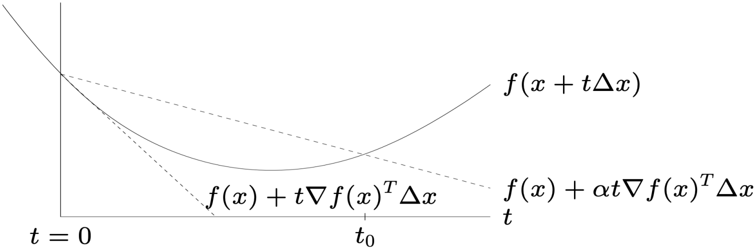
$$t = \operatorname*{\mathbf{argmin}}_{t>0} f(x + t\Delta x)$$

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

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

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

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



Gradient descent method

gradient descent direction $\Delta x = -\nabla f(x)$

---

given a starting point $x \in \mathbf{dom}\, f$

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

- general descent method with $\Delta x = -\nabla f(x)$
- stopping criterion usually of the form

$$\|\nabla f(x)\|_2 \leq \epsilon$$

- convergence result: for strongly convex f

$$f(x^{(k)}) - p^* \leq c^k \left(f(x^{(0)}) - p^* \right)$$

  $c \in (0, 1)$ depends on m, $x^{(0)}$, line search type

- very simple, but often very slow; rarely used in practice

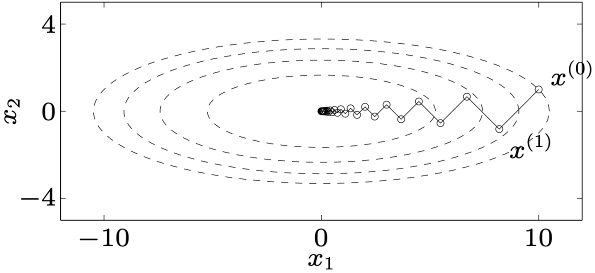
Quadratic example in $\mathbb{R}^2$

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

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

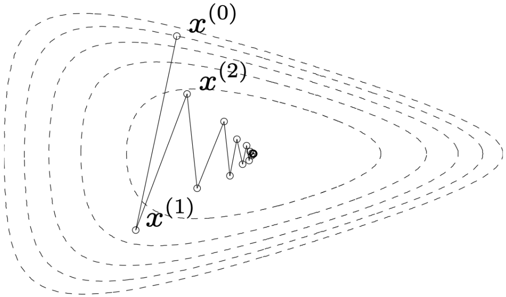
$$x_1^{(k)} = \gamma \left(\frac{\gamma - 1}{\gamma + 1} \right)^k, \qquad x_2^{(k)} = \left(-\frac{\gamma - 1}{\gamma + 1} \right)^k$$

very slow if $\gamma \gg 1$ or $\gamma \ll 1$, following example for $\gamma = 10$

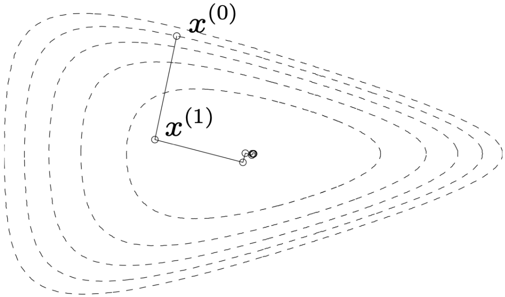


Nonquadratic example in $\mathbb{R}^2$

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



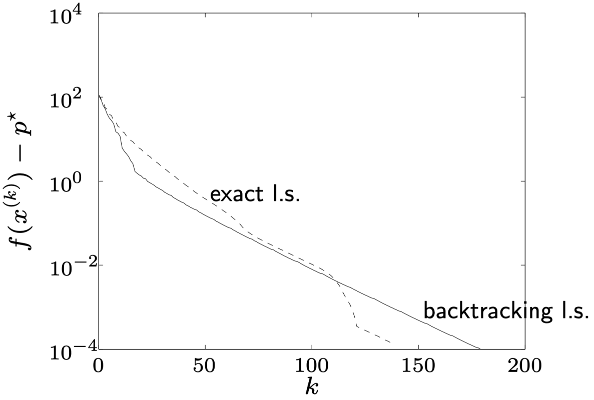
backtracking line search



exact line search

Example in $\mathbb{R}^{100}$

$$f(x) = c^T x - \sum_{i=1}^{500} \log\left(b_i - a_i^T x\right)$$



"linear" convergence (straight line on a semilog plot)

Steepest descent method

normalized steepest descent direction (for norm $\|\cdot\|$)

$$\Delta x_{\mathrm{nsd}} = \mathbf{argmin}\{\nabla f(x)^T v \mid \|v\| = 1\}$$

- for small v we have $f(x+v) \approx f(x) + \nabla f(x)^T v$
- direction Δx_{nsd} is unit-norm step with most negative directional derivative

unnormalized steepest descent direction

$$\Delta x_{\mathrm{sd}} = \|\nabla f(x)\|_* \Delta x_{\mathrm{nsd}}$$

satisfies $\nabla f(x)^T \Delta x_{\mathrm{sd}} = -\|\nabla f(x)\|_*^2$

- general descent method with $\Delta x = \Delta x_{\mathrm{sd}}$
- convergence properties similar to gradient descent

Examples

- Euclidean norm $\|x\|_2$

$$\Delta x_{\mathrm{sd}} = -\nabla f(x)$$

  same as gradient descent

- quadratic norm $\|x\|_P = (x^T P x)^{1/2}$ for $P \in \mathbb{S}^n_{++}$

$$\Delta x_{\mathrm{sd}} = -P^{-1}\nabla f(x)$$

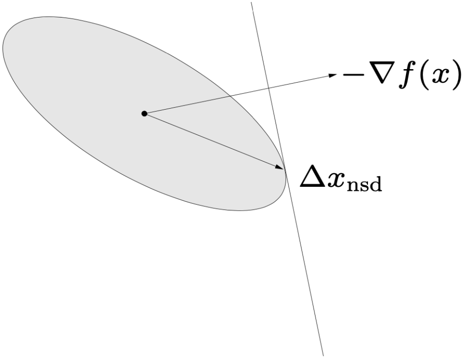
  gradient descent after change of variables $\bar{x} = P^{1/2}x$

- ℓ_1-norm

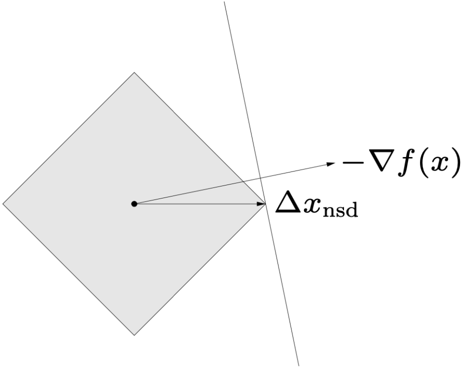
$$\Delta x_{\mathrm{sd}} = -(\partial f(x)/\partial x_i)e_i$$

  where $|\partial f(x)/\partial x_i| = \|\nabla f(x)\|_\infty$

unit balls and normalized steepest descent directions

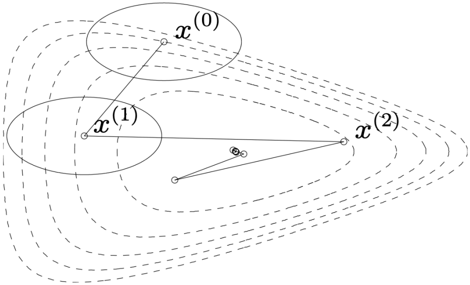


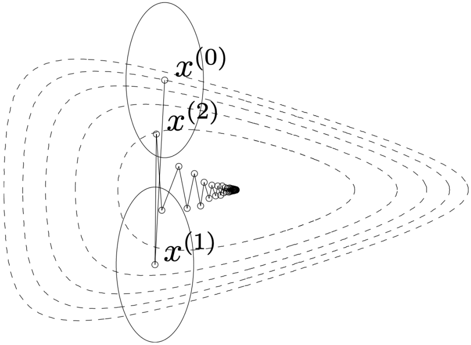
a quadratic norm



the ℓ_1-norm

steepest descent with backtracking line search for two quadratic norms





- dashed lines are contour lines of $f(x)$
- ellipses show $\{x \mid \|x - x^{(k)}\|_P = 1\}$
- choice of P has strong effect on speed of convergence

Newton step

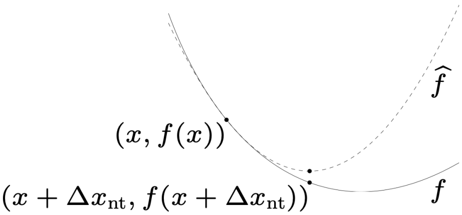
$$\Delta x_{\mathrm{nt}} = -\nabla^2 f(x)^{-1} \nabla f(x)$$

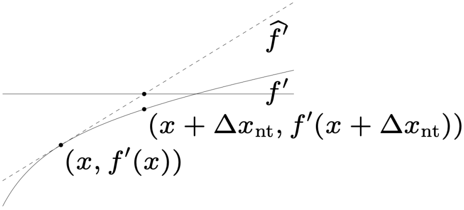
- $x + \Delta x_{\mathrm{nt}}$ minimizes second order approximation

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

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

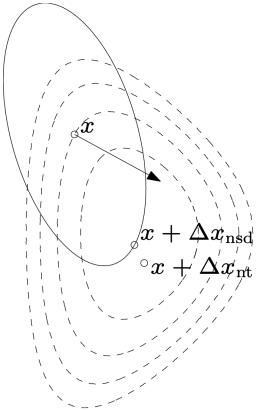
$$\nabla f(x+v) \approx \nabla \widehat{f}(x+v) = \nabla f(x) + \nabla^2 f(x) v = 0$$





- Δx_{nt} is steepest descent direction at x in local Hessian norm

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



ellipse is $\{x + v \mid v^T \nabla^2 f(x) v = 1\}$, arrow shows $-\nabla f(x)$

Newton decrement

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

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

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

- 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 Newton direction

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

properties

- a measure of proximity of x to x^*
- an affine invariant (independent of linear change of coordinates, unlike $\|\nabla f(x)\|_2$)

Newton's method

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

repeat

- *Compute Newton step and decrement.*

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

- *Stopping criterion.* **quit** if $\lambda^2/2 \leq \epsilon$
- *Line search.* Choose step size t by backtracking line search
- *Update.* $x := x + t\Delta x_{\text{nt}}$

affine invariance

Newton iterates for

$$\tilde{f}(y) = f(Ty)$$

with starting point

$$y^{(0)} = T^{-1}x^{(0)}$$

are

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

Classical convergence analysis

assumptions

- f strongly convex on S with constant $m > 0$

$$\nabla^2 f(x) \succeq mI$$

- $\nabla^2 f$ Lipschitz continuous on S with constant $L > 0$

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

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

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

- if $\|\nabla f(x)\|_2 \geq \eta$, then
$$f\left(x^{(k+1)}\right) - f\left(x^k\right) \leq -\gamma$$
- if $\|\nabla f(x)\|_2 < \eta$, then
$$\frac{L}{2m^2}\left\|\nabla f\left(x^{(k+1)}\right)\right\|_2 \leq \left(\frac{L}{2m^2}\left\|\nabla f\left(x^k\right)\right\|_2\right)^2$$

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

- most iterations require backtracking steps
- function value decreases by at least γ
- if $p^* > -\infty$, this phase ends after at most $\left(f(x^{(0)}) - p^*\right)/\gamma$ iterations

quadratically convergent phase $\|\nabla f(x)\|_2 < \eta$

- all iterations use step size $t = 1$
- $\|\nabla f(x)\|_2$ converges to zero quadratically

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

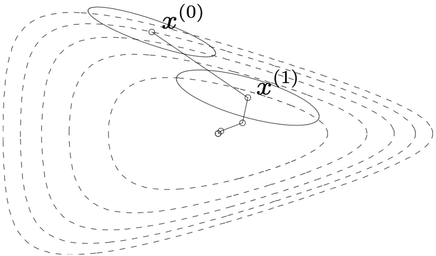
holds for $l \geq k$ if $\|\nabla f(x^{(k)})\|_2 < \eta$

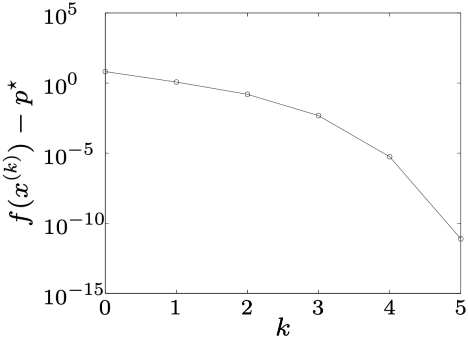
conclusion number of iterations until $f(x) - p^* \leq \epsilon$ is bounded above by

$$\frac{f\left(x^{(0)}\right) - p^*}{\gamma} + \log_2 \log_2 \left(\frac{\epsilon_0}{\epsilon}\right)$$

- γ, ϵ_0 are constants that depend on m, L, $x^{(0)}$
- second term is small and almost constant for practical purposes (say 5 or 6)
- constants m, L are usually unknown in practice
- provides qualitative insight in convergence properties

Example in $\mathbb{R}^2$ $\quad f(x_1, x_2) = e^{x_1+3x_2-0.1} + e^{x_1-3x_2-0.1} + e^{-x_1-0.1}$

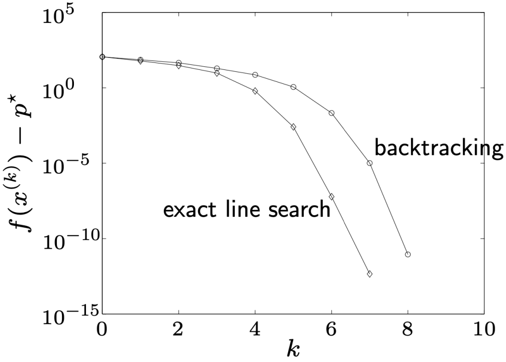


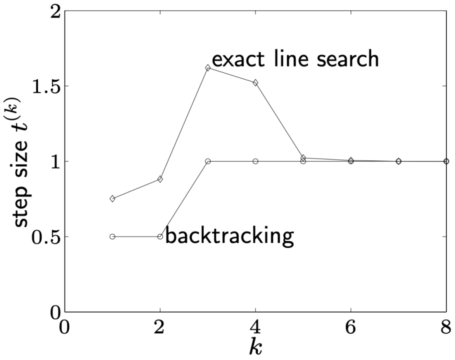


- backtracking parameters $\alpha = 0.1$, $\beta = 0.7$
- converges in only 5 steps
- clearly shows quadratic convergence

Example in $\mathbb{R}^{100}$

$$f(x) = c^T x - \sum_{i=1}^{500} \log\left(b_i - a_i^T x\right)$$

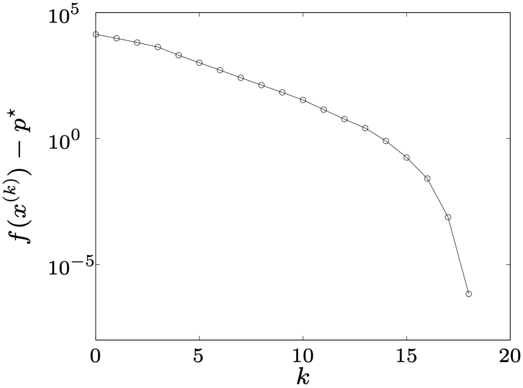




- backtracking parameters $\alpha = 0.01$, $\beta = 0.5$
- backtracking line search almost as fast as exact line search (and much simpler)
- clearly shows two phases in algorithm

Example in $\mathbb{R}^{10000}$

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



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

Self-concordance

shortcomings of classical convergence analysis

- depends on unknown constants (m, L, ...)
- bound is not affine invariant, although Newton's method is

convergence analysis via self-concordance (Nesterov and Nemirovski)

- does not depend on any unknown constants
- gives affine invariant bound
- applies to special class of convex functions ('self-concordant' functions)
- developed to analyze polynomial-time interior-point methods for convex optimization

Self-concordant functions

- convex function $f : \mathbb{R} \to \mathbb{R}$ is **self-concordant** if

$$\left| f'''(x) \right| \leq 2 f''(x)^{3/2}$$

  for all $x \in \mathbf{dom}\, f$

- function $f : \mathbb{R}^n \to \mathbb{R}$ is **self-concordant** if

$$g(t) = f(x + tv)$$

  is self-concordant for all $x \in \mathbf{dom}\, f$ and $v \in \mathbb{R}^n$

examples on $\mathbb{R}$

- linear and quadratic functions
- negative logarithm

$$f(x) = -\log x$$

- negative entropy plus negative logarithm

$$f(x) = x \log x - \log x$$

affine invariance

$$f : \mathbb{R} \to \mathbb{R} \text{ is self-concordant} \quad \Longrightarrow \quad \tilde{f}(y) = f(ay + b) \text{ is self-concordant}$$

$$\tilde{f}'''(y) = a^3 f'''(ay + b), \qquad \tilde{f}''(y) = a^2 f''(ay + b)$$

Self-concordant calculus

properties

- preserved under sum and positive scaling $\alpha \geq 1$
- preserved under composition with affine function
- if g is convex with

$$\mathbf{dom}\, g = \mathbb{R}_{++} \qquad \text{and} \qquad |g'''(x)| \leq 3g''(x)/x$$

then

$$f(x) = \log(-g(x)) - \log x$$

is self-concordant

examples

$$f(x) = -\sum_{i=1}^{m} \log\left(b_i - a_i^T x\right) \quad \text{on} \quad \{x \mid a_i^T x < b_i, i = 1, \cdots, m\}$$

$$f(X) = -\log\det X \quad \text{on} \quad \mathbb{S}_{++}^{n}$$

$$f(x, y) = -\log\left(y^2 - x^T x\right) \quad \text{on} \quad \{(x, y) \mid \|x\|_2 < y\}$$

Convergence analysis for self-concordant functions

summary there exist constants $\eta \in (0, 1/4]$, $\gamma > 0$ such that

- if $\lambda(x) > \eta$, then
$$f\left(x^{(k+1)}\right) - f\left(x^{(k)}\right) \leq -\gamma$$
- if $\lambda(x) \leq \eta$, then
$$2\lambda(x^{(k+1)}) \leq \left(2\lambda(x^{(k)})\right)^2$$

where η and γ only depend on backtracking parameters α and β

complexity bound number of Newton iterations bounded by

$$\frac{f(x^{(0)}) - p^*}{\gamma} + \log_2 \log_2 (1/\epsilon)$$

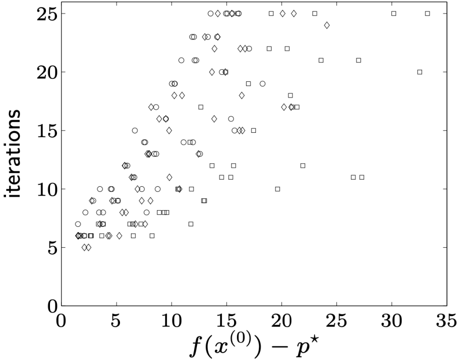
for $\alpha = 0.1$, $\beta = 0.8$, $\epsilon = 10^{-10}$, bound evaluates to

$$375 \left(f(x^{(0)}) - p^* \right) + 6$$

numerical example 150 randomly generated instances of

$$\text{minimize} \qquad f(x) = -\sum_{i=1}^{m} \log\left(b_i - a_i^T x\right)$$

○: $m = 100$, $n = 50$
□: $m = 1000$, $n = 500$
◇: $m = 1000$, $n = 50$



- number of iterations much smaller than $375\left(f(x^{(0)}) - p^*\right) + 6$
- bound of the form $c\left(f(x^{(0)}) - p^*\right) + 6$ with smaller c (empirically) valid

Implementation

main effort in each iteration

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

requires to evaluate derivatives and solve Newton system

$$H \Delta x = -g$$

where $H = \nabla^2 f(x)$, $g = \nabla f(x)$

Cholesky factorization

$$H = LL^T, \qquad \Delta x_{\mathrm{nt}} = -L^{-T}L^{-1}g, \qquad \lambda(x) = \|L^{-1}g\|_2$$

- cost $(1/3)n^3$ flops for unstructured system
- cost $\ll (1/3)n^3$ if H sparse or banded

example of dense Newton system with structure

$$f(x) = \sum_{i=1}^{n} \psi_i(x_i) + \psi_0(Ax + b)$$

assume $A \in \mathbb{R}^{p \times n}$, dense with $p \ll n$, then

$$H = D + A^T H_0 A$$

where

$$D = \mathbf{diag}(\psi_1''(x_1), \ldots, \psi_n''(x_n)), \qquad H_0 = (\nabla^2 \psi_0)(Ax + b)$$

first method solve via dense Cholesky factorization, cost $\approx (1/3)n^3$

second method solve via block elimination

- factor $H_0 = L_0 L_0^T$, write Newton system as

$$D\Delta x + A^T L_0 w = -g, \qquad L_0^T A \Delta x - w = 0$$

- eliminate Δx from first equation, compute w and Δx from

$$(I + L_0^T A D^{-1} A^T L_0) w = -L_0^T A D^{-1} g, \qquad D\Delta x = -g - A^T L_0 w$$

- cost $\approx 2p^2 n$ (dominated by computation of $L_0^T A D^{-1} A^T L_0$)