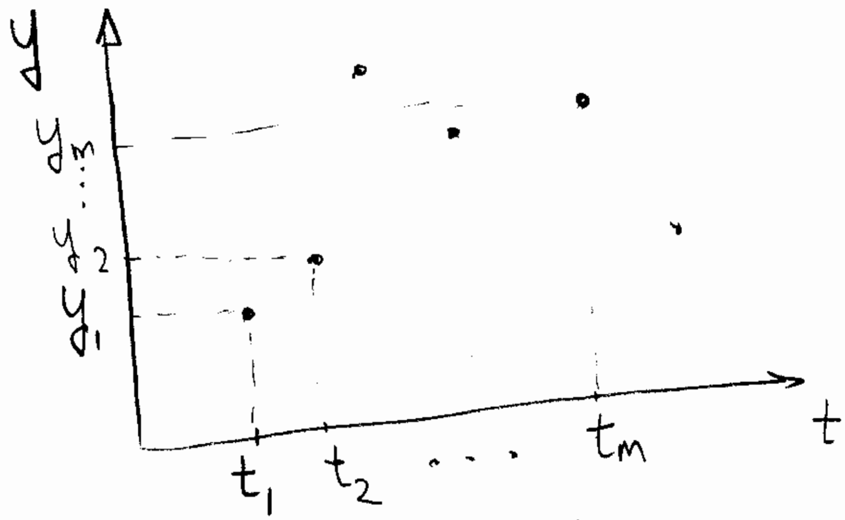
Chapter 2. Fundamentals of Unconstrained Optimization

A non-linear least-squares example



Consider the model:

$$\phi(t;x) = x_1 + x_2 e^{-(x_3-t)^2/x_4} + x_5 \cos(x_6 t)$$

We need to find $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_6 \end{bmatrix}$

such that the error is minimized:

residual: $r_j(x) = y_j - \phi(t_j; x)$

<u>Compute</u> $\min\limits_{x \in \mathbb{R}^6} f(x) = r_1^2(x) + \dots + r_m^2(x)$

How do we know we are there?

2.1 What is a solution?

Global minimizer:

x^* is a global minimizer if:

$$f(x^*) \leq f(x), \text{ for all } x.$$

Local minimizer:

x^* is a local minimizer if:

$\exists N(x^*)$ such that:

$$f(x^*) \leq f(x) \text{ for } x \in N.$$

Here $N(x^*)$ refers to an open set that contains x^*.

Open sets in $\mathbb{R}^1$:

Ex 1: $S = \{x \mid a < x < b\}$

Ex 2: $S = \{x \mid a < x < b \text{ or } c < x < d\}$

Open sets in $\mathbb{R}^2$:

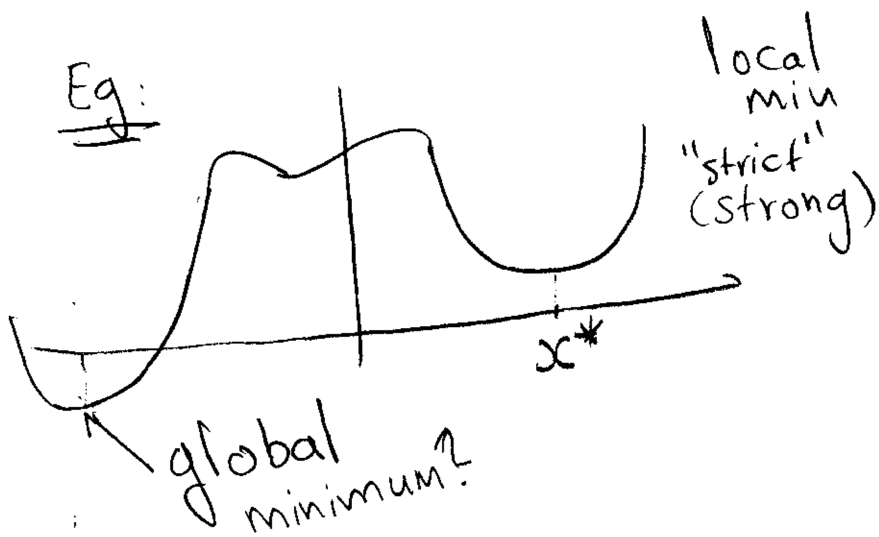
Ex 1: $S = \{(x_1, x_2) \mid a < x_1, x_2 < b\}$

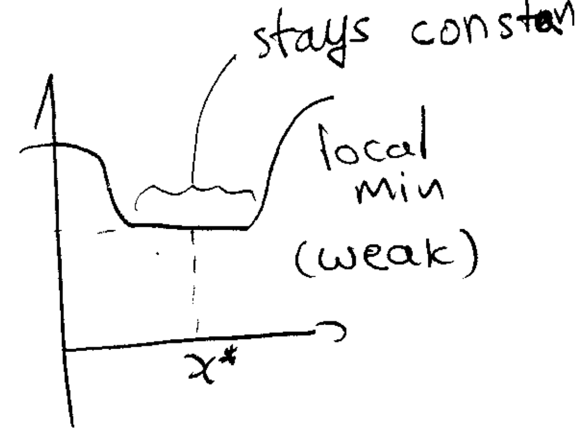
... Note that there are no equalities.

A point x^* is a <u>strict local minimizer</u> (or strong local minimizer) if there is a neighborhood N of x^* such that:

$$f(x^*) < f(x) \quad \text{for all } x \in N,\ x \neq x^*$$

<u>Eg:</u>





A point x^* is an <u>isolated local minimizer</u> if there is a neighborhood N of x^* such that x^* is the only local minimizer in N.

<u>Difficult Case:</u>

$$f(x) = x^4\cos(1/x) + 2x^4, \quad f(0) = 0$$

At $x^* = 0$, cannot isolate one and only one minimum.

Thm 2.1 Taylor's thm

Let $f: \mathbb{R}^n \to \mathbb{R}$ be continuously-diff'ble.
Let $p \in \mathbb{R}^n$. Then:

$$f(x+p) = f(x) + \nabla f(x+tp)^T p \quad (*)$$

for some $t \in (0,1)$.

Moreover, if f is twice continuously-differentiable, we have that:

$$\nabla f(x+p) = \nabla f(x) + \int_0^1 \nabla^2 f(x+tp)\, p \, dt$$

and that:

$$f(x+p) = f(x) + \nabla f(x)^T p + \tfrac{1}{2} p^T \nabla^2 f(x+tp)\, p \quad (**)$$

for some $t \in (0,1)$.

Understanding the notation

Suppose that:

$$f(x_1, x_2) = x_1^2 + 2x_2^2 + 3x_1 + 2x_1x_2$$

$$\nabla f(x_1, x_2) = \begin{bmatrix} \dfrac{\partial f(x_1,x_2)}{\partial x_1} \\ \dfrac{\partial f(x_1,x_2)}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 + 2x_2 + 3 \\ 4x_2 + 2x_1 \end{bmatrix}$$

$$\nabla^2 f(x_1, x_2) = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1 \partial x_2} \\ \dfrac{\partial^2 f}{\partial x_1 \partial x_2} & \dfrac{\partial^2 f}{\partial x_2^2} \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 4 \end{bmatrix}$$

Pick a direction. Say $p = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

The theorem (*) says that we can solve for t, the following:

$$f(x_1, x_2+1) = f(x_1, x_2) + \underbrace{4(x_2+t) + 2x_1}_{\text{replace } \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \text{ by } \begin{bmatrix} x_1+0 \\ x_2+t \end{bmatrix}}$$

for $t \in (0,1)$.

Also, (**) says that:

$$f(x_1, x_2+1) = f(x_1, x_2) + (4x_2 + 2x_1) + 2$$

Thm 2.2 First Order Necessary Conditions

If x^* is a <u>local minimizer</u> and f is <u>continuously diff'ble</u> in an open neighborhood of x^*, then $\nabla f(x^*) = 0$.

<u>Proof</u> Assume $\nabla f(x^*) \neq 0$.

Assume that it is not true, and derive a contradiction

Define $p = -\nabla f(x^*)$

and
$$p^T \nabla f(x^*) = -\nabla f(x^*)^T \nabla f(x^*)$$
$$= -\|\nabla f(x^*)\|^2 < 0.$$

Since ∇f is conts near x^*, $\exists T$
$$p^T \nabla f(x^* + tp) < 0, \quad \text{all } t \in [0, T]$$

Then, for $\bar{t} \in (0, T]$, use thm 2.1:
$$f(x^* + \bar{t}p) = f(x^*) + \bar{t} p^T \nabla f(x + t''p)$$
for some $t'' \in (0, \bar{t})$

Note that $t'' \in [0, T]$ and thus
$$\bar{t} p^T \nabla f(x + t''p) < 0$$
$$\Rightarrow f(x^* + \bar{t}p) < f(x^*)$$

$\boxed{\text{not possible}}$ ∎

x^* is a stationary point if
$\nabla f(x^*) = 0$.

Thm 2.3 (Second-Order Necessary Conds)
If x^* is a local minimizer of f and $\nabla^2 f$ is continuous in an open neighborhood of x^*, then $\nabla f(x^*)=0$ and $\nabla^2 f(x^*)$ is positive semidefinite.

Review:
① A matrix is positive definite if $p^T B p > 0$, $\forall p \neq 0$.

② A matrix is positive semi-definite if $p^T B p \geq 0$, $\forall p \neq 0$.

Proof: omitted, (but see 2.4).

Thm 2.4 (Second-Order Sufficient Conditions)

Suppose:

(I) $\nabla^2 f$ is continuous in an open neighborhood N of x^*,

(II) $\nabla^2 f$ is positive definite in N,

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

Then: x^* is a strict local minimizer of f.

Proof: Choose some $r > 0$, so that $\nabla^2 f(x)$ is positive definite for

$$x \in D \subseteq N \text{ and } D = \{z \mid \|z - x^*\| < r\}$$

Pick $p \neq 0$ with $\|p\| < r$.
At $x^* + p \in D$, we have:

$$f(x^* + p) = f(x^*) + p^T \nabla f(x^*) + \tfrac{1}{2} p^T \nabla^2 f(z) p$$

for some $z = x^* + tp$, $t \in (0,1)$

Since $\nabla f(x^*) = 0$, we have:

$$f(x^* + p) = f(x^*) + \underbrace{\tfrac{1}{2} p^T \nabla^2 f(z)\, p}_{\text{positive by assumption.}}$$

$\Rightarrow \quad f(x^* + p) > f(x^*) \quad (*)$

$\Rightarrow \quad x^*$ is a strict local minimizer of f. This proves that it is sufficient

Note that $\nabla^2 f(z)$ must be zero or positive definite, else we can find p with:

$\tfrac{1}{2} p^T \nabla^2 f(z)\, p$ that is negative, violating $(*)$.

This proves that $\nabla^2 f$ must be positive semi-definite. (necessary).

To show that $\nabla^2 f = 0$ will work:

Let $f(x) = x^4$, $\nabla^2 f(0) = 0$,

but $x = 0$ is a strict-local minimum.

Thm 2.5 Suppose that f is convex.

(meaning: $f(\lambda x + (1-\lambda) y) \leq \lambda f(x) + (1-\lambda) f(y)$)

Then:

(I) x^* is a local minimizer $\Longrightarrow$ x^* is a global minimizer

(II) x^* is a stationary point (f diff'ble there) $\Longrightarrow$ x^* is a global minimizer

Proof: The idea is that to show $A \Rightarrow B$, we show $\neg B \Rightarrow \neg A$ (contra-positive)

We thus assume $\neg B$, and show that we have $\neg A$.

(I) Assume x^* is not a global minimizer ($\neg B$)

By definition, this means that we can find a point $z \in \mathbb{R}^n$ with $f(z) < f(x^*)$.

Consider $x = \lambda z + (1-\lambda) x^*$, $\lambda \in (0,1]$

By convexity:

$$f(\underbrace{\lambda z+(1-\lambda)x^*}_{x}) \leq \lambda f(z) + (1-\lambda) f(x^*)$$

Clearly:

$$\lambda f(z) + (1-\lambda) f(x^*) < \underbrace{\lambda f(x^*) + (1-\lambda) f(x^*)}_{= f(x^*)}$$

$\Rightarrow \quad f(x) < f(x^*)$

$\Rightarrow$ x^* is not a local minimizer for λ small-enough so that $\lambda z + (1-\lambda)x^* \in N(x^*)$ where x^* would have been allowed to be a local minimizer. $(\neg A)$

$\blacksquare$

(II) Assume x^* is <u>not</u> a global minimizer $(\neg B)$. Define z as in (I).

By the definition of the directional derivative:

$$\lim_{\lambda \downarrow 0} \frac{f(x^* + \overbrace{\lambda(z-x^*)}^{\text{in this direction}}) - f(x^*)}{\lambda}$$

$$= \nabla f(x^*)^T (z - x^*)$$ Δ (identity)

$$= \lim_{\lambda \downarrow 0} \frac{f(\lambda z + (1-\lambda) x^*) - f(x^*)}{\lambda}$$

$$\leq \lim_{\lambda \downarrow 0} \frac{\lambda f(z) + (1-\lambda) f(x^*) - f(x^*)}{\lambda}$$

$$= \lim_{\lambda \downarrow 0} \frac{\lambda f(z) + \cancel{f(x^*)} - \lambda f(x^*) - \cancel{f(x^*)}}{\lambda}$$

$$= f(z) - f(x^*) < 0$$

<u>Since:</u> $f(z) < f(x^*)$

Note that:

$$\nabla f(x^*)^T (z - x^*) < 0$$

which implies $\nabla f(x^*) \neq 0$
(a contradiction that x^* is a stationary point) $(\neg A)$ $\blacksquare$

Non-smooth problems: See Fletcher.

2.2 Overview of Algorithms

Starting point x_0:

* must supply to algorithms:
  - ⇒ estimate x_0 to be close to optimal,
  - ⇒ if no estimate is available, try $x_0 = \vec{0}$ or random x_0 that still satisfies constraints (test them)

⇒ Algorithm generates

$$x_1, x_2, \ldots, x_{k-1}$$

at the kth iteration to compute x_k

⇒ Algorithms require:

either $f(x_k) < f(x_{k-1}) < \cdots < f(x_0)$

or:

$$f(x_k) < f(x_{k-m})$$

(reduction after m iterations)

Line Search

Find $\alpha > 0$ so that:

$$\min_{\alpha > 0} f(x_k + \alpha p_k) \sim (*)$$

along some given direction p_k.

Repeat at the new point: $x_{k+1} = x_k + \alpha p_k$ with a new direction p_{k+1}.

Usually $(*)$ is approximately solved for a few trial values for α, until a solution is approximately found.

Trust Regions

Form a region around the current guess x_k, and form a model m_k, applicable for this region. Select direction p_k from the model so that:

$$\min_{\underline{p}} m_k(x_k + \underline{p}), \quad x_k + \underline{p} \text{ inside trust region.}$$

Usually:

$$m_k(x_k + \underline{p}) = f_k + \underline{p}^T \nabla f_k + \frac{1}{2} \underline{p}^T B_k \underline{p}$$

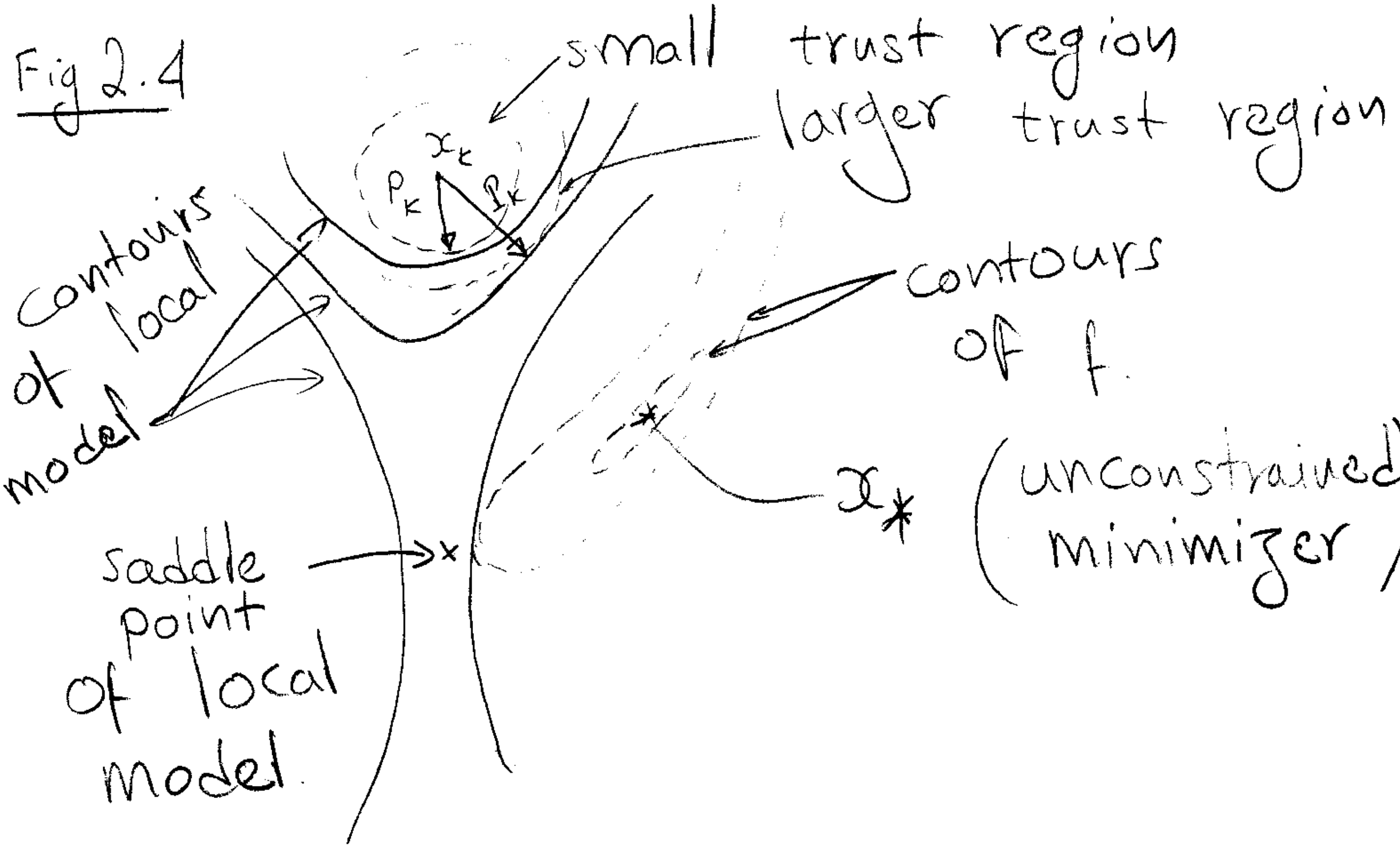
where B_k approximates $\nabla^2 f_k$.

E.g. $f(x) = 10(x_2 - x_1^2)^2 + (1 - x_1)^2$

At $x_k = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, we have:

$$\nabla f_k = \begin{bmatrix} -2 \\ 20 \end{bmatrix}, \quad \nabla^2 f_k = \underbrace{\begin{bmatrix} -38 & 0 \\ 0 & 20 \end{bmatrix}}_{\text{saddle-point behavior } (\lambda_1 < 0,\ \lambda_2 > 0)}$$

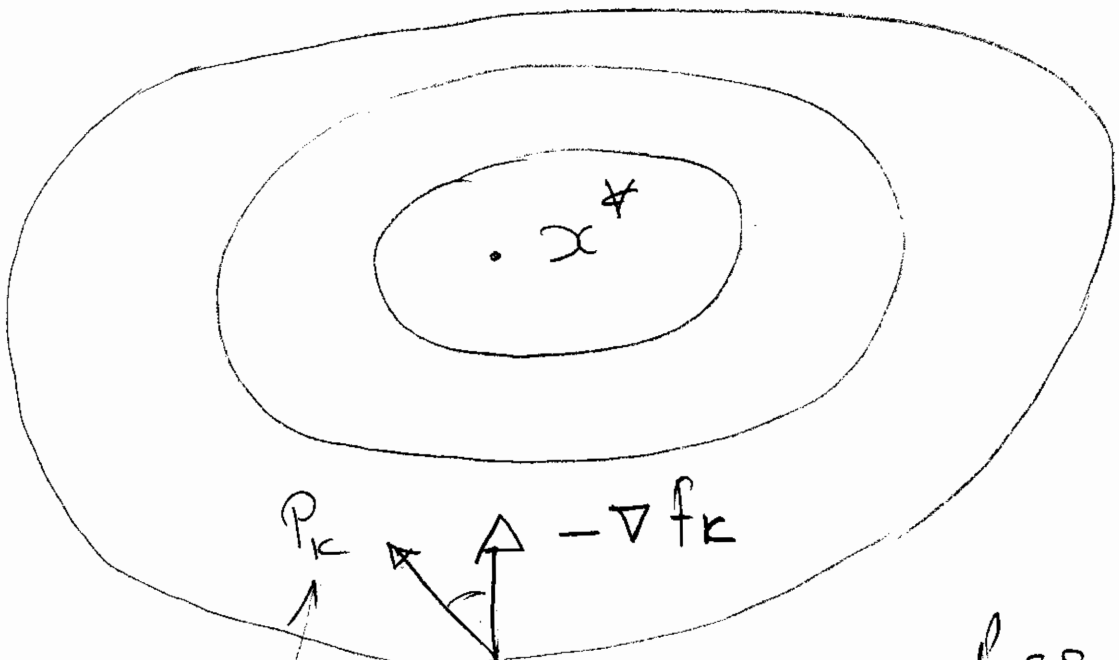
Fig 2.4


small trust region
larger trust region
x_k
p_k
p_k
contours of local model
contours of f.
x_* (unconstrained minimizer)
saddle point of local model

Search Directions for Line Search

Steepest descent: $-\nabla f_k$ for p_k.

set $$x_{k+1} = x_k + \alpha_k p_k$$
$$= x_k - \alpha_k \nabla f_k.$$



Any direction that forms less than 90° with $-\nabla f_k$ will work as a **descent direction**.

Watch-out for SCALING problems

Newton Direction

From the Taylor series expansion up to the 2nd term (one beyond what we did ...):

$$f(x_k + p) \approx \underbrace{f_k + p^T \nabla f_k + \tfrac{1}{2} p^T \nabla^2 f_k\, p}_{m_k(p) \text{ the model}}$$

Assume $\nabla^2 f_k$ is positive definite.

Then, for the optimal p:

$$\min_p \; f_k + p^T \nabla f_k + \tfrac{1}{2} p^T \nabla^2 f_k\, p_k$$

or $\nabla_p m_k(p) = 0$.

for Newton

$$\Rightarrow \quad \nabla f_k + \underbrace{\tfrac{1}{2} \times 2 \times \nabla^2 f_k \times p_k^N}_{} = 0$$

from $\tfrac{1}{2} p^T \nabla^2 f_k\, p = \tfrac{1}{2} \sum_i \sum_j (\nabla^2 f_k)_{ij}\, p_i p_j$

$$\Rightarrow \boxed{p_k^N = -\left(\nabla^2 f_k^{-1}\right) \nabla f_k}$$

Our approximation by Newton's method is exact up to $\|p\|^2$ terms. (MVT for one more term!).

The error is $O(\|p\|^3)$, assuming that $\|p\| \ll 1$ so that:

$$\|p\|^3 >> \|p\|^4, \|p\|^5, \ldots$$

E.g.: For $\|p\| = 0.1$, $\|p\|^3 = 0.001$, while $\|p\|^2 = 0.01$, $\|p\|^4 = 0.0001$, ...

$\therefore$ For $\|p\|$ small, the approximation $f(x_k + p) \approx m_k(p)$ is accurate.

If $\nabla^2 f_k$ is positive definite, then:

$$\nabla f_k^T p_k^N = \underbrace{\nabla f_k^T (\nabla^2 f_k)^{-1}}_{-p_k^N} (\nabla^2 f_k) p_k^N$$

$$= -p_k^N \nabla^2 f_k \, p_k^N \leq 0 \Leftarrow$$ it is a descent direction

Quasi-Newton (corrected proof)

start from:

$$\nabla f(x+p) = \nabla f(x) + \int_0^1 \nabla^2 f(x+tp)\, p \, dt$$

Add:

$$+ \nabla^2 f(x+p)p - \nabla^2 f(x+p)\, p$$

to get:

$$\nabla f(x+p) = \nabla f(x) + \nabla^2 f(x+p)\, p + \int_0^1 \left[\nabla^2 f(x+tp) - \underbrace{\nabla^2 f(x+p)}\right] p \, dt$$

since this is a constant w.r.t. t.

Set: $x = x_k$, $p = x_{k+1} - x_k$.

$$\Rightarrow x_{k+1} = x_k + p$$

$$\Rightarrow \begin{cases} \nabla f(x+p) & \text{is } \nabla f_{k+1} \\ \nabla f(x) & \text{is } \nabla f_k \\ \nabla^2 f(x+p) & \text{is } \nabla^2 f_{k+1} \end{cases}$$

which gives:

$$\nabla f_{k+1} = \nabla f_k + \nabla^2 f_{k+1}(x_{k+1} - x_k) + \underbrace{o(\|x_{k+1} - x_k\|)}_{}$$

$$\Uparrow\Vert$$

W/out proof, for $\int_0^1 [\nabla^2 f(x+tp) - \nabla^2 f(x+p)] \times p \, dt$

(due to continuity).

$\Rightarrow$ Approximate using:

$$\underbrace{\nabla^2 f_{k+1}}_{B_{k+1}} \underbrace{(x_{k+1} - x_k)}_{s_k} \approx \underbrace{\nabla f_{k+1} - \nabla f_k}_{= y_k}$$

Set: $B_{k+1} \times s_k = y_k$

which is used to estimate the approximation to the Hessian:

$$B_{k+1} \approx \nabla^2 f_{k+1},$$

$s_k = x_{k+1} - x_k$, $y_k = \nabla f_{k+1} - \nabla f_k$

An update rule is defined through

Symmetric-rank-one (SR1):

$$B_{k+1} = B_k + \underbrace{\frac{(y_k - B_k s_k)(y_k - B_k s_k)^T}{(y_k - B_k s_k)^T s_k}}$$

single-vector $\Rightarrow$ rank-one matrix update.

or:

BFGS formula:

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

rank-2 update.

Soln is then:

$$p_k = - B_k^{-1} \nabla f_k \quad \text{as before,}$$

but now approximately.

Even better, avoid computing B_k^{-1} by updating $H_k = B_k^{-1}$:

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

$$\rho_k = \frac{1}{y_k^T s_k},$$

and then: $\boxed{p_k = -H_k \nabla f_k}$

For <u>large problems</u>:

<u>partially-separable</u> and <u>limited-memory</u> updating in chapter 9.

Rates of convergence (Q = quotient)

Q linear: $\exists r \in (0,1)$ such that $\dfrac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} \leq r$,

for all k sufficiently large.

Q-superlinear:

Require: $$\lim_{k \to \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0.$$

Q-quadratic Convergence:

$\exists M$, for all k sufficiently large so that:

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^2} \leq M > 0.$$

Q-order p ($p>1$)

$\exists M$, for all k sufficiently large so that:

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^p} \leq M.$$

Q-order $p \Rightarrow$ Q-order $p-1 \Rightarrow \cdots$

Examples:

Ex 1 $x_k = 1 + (0.5)^k \longrightarrow 1$

gives:

$$\frac{\| x_{k+1} - 1 \|}{\| x_k - x^* \|} = \frac{0.5^{k+1}}{0.5^k} = 0.5 \in (0,1)$$

$\Rightarrow$ Q-linear convergence.

Ex 2 $x_k = 1 + k^{-k} \longrightarrow 1$, as $k \to \infty$

gives

$$\frac{(k+1)^{-(k+1)}}{k^{-k}} = \frac{(k+1)^{-k}}{k^{-k}} \cdot \frac{1}{(k+1)} \to 0$$

as $k \to \infty$

$\Rightarrow$ Q-superlinear convergence.

Ex 3 $x_k = 1 + (0.5)^{2^k} \longrightarrow 1.$

gives:

$$\frac{(0.5)^{2^{k+1}}}{\left((0.5)^{2^k} \right)^2} = 1 \leq M, \ k \text{ large}$$

$\Rightarrow$ Q-quadratic convergence.

R-rates of Convergence (R=root)

R-linear

Suppose that there is a sequence of nonnegative scalars $\{\nu_k\}$ such that:

$$\|x_k - x^*\| \le \nu_k, \text{ for all } k,$$

and $\{\nu_k\}$ converges Q-linearly to zero.

R-superlinear

... if $\{\|x_k - x^*\|\}$ is dominated by a Q-superlinear sequence.

R-Quadratically

... if $\{\|x_k - x^*\|\}$ is dominated by a Q-quadratic sequence.

Eg:

$$x_k = \begin{cases} 1+(0.5)^k, & k \text{ even} \\ 1, & k \text{ odd} \end{cases}$$

converges R-linearly to 1, w/out requiring decrease on every step.