Ch. 9 Unconstrained Minimization

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

- Assume $f : \mathbb{R}^n \to \mathbb{R}$ is convex and twice continuously differentiable.

~~- Assume this problem is solvable.~~

- x^* is optimal $\Leftrightarrow$ $\nabla f(x^*) = 0$.
  necessary and sufficient condition

However it is not necessary that an <u>analytical solution</u> exists.

Ex. minimize $f(x) = \log\left(\sum_{i=1}^{m} e^{a_i^T x + b_i} \right)$.

$$0 = \nabla f(x^*) = \frac{1}{\sum_{i=1}^{m} e^{a_i^T x + b_i}} \left(\sum_{i=1}^{m} e^{a_i^T x + b_i} a_i \right).$$

where $a_i \in \mathbb{R}^n$. $x \in \mathbb{R}^n$.

$\Rightarrow$ Need to use an iterative algorithm.

The algorithms described in this chapter produce a minimizing sequence $x^{(k)}$, $k=1,2,\dots$

$$\left(f(x^{(k+1)}) < f(x^{(k)}) \right)$$

where $x^{(k+1)} = x^{(k)} + t^{(k)} \Delta x^{(k)}$

— $\Delta x \in \mathbb{R}^n$: step or search direction

— $t^{(k)} \geq 0$: step size.

— one iteration: $\begin{cases} x^+ = x + t\Delta x. \\ x := x + t\Delta x. \end{cases}$

From convexity of f, $f(y) \geq f(x) + \nabla f(x)^T (y-x)$

Let $y = x + t\Delta x$. $\Rightarrow f(x+t\Delta x) \geq f(x) + \underset{t\geq 0}{\underline{\underline{t}}} \nabla f(x)^T \Delta x$.

for $\underline{f(x+t\Delta x) < f(x)}$, we must have

$$\nabla f(x)^T \Delta x < 0.$$

Decent Method

Given a starting point $x \in \operatorname{dom} f$

Repeat 1. Determine a decent direction Δx.

2. Line search. Choose a step size $t > 0$

3. Update $x := x + t\Delta x$.

Until stopping criterion is achived.

Gradient dencent method

- decent direction $\underline{\Delta x} = -\nabla f(x)$.

Thus $\nabla f(x)^T \Delta x = -\|\nabla f(x)\|_2^2 \le 0$.

- Stopping criterion: $\eta \sim 10^{-10}$.

$$\|\nabla f(x)\|_2^2 \le \eta$$

where η is a small and positive number.

Line search type

1. Exact line search $\quad t = \underset{t>0}{\arg\min} f(x+t\Delta x)$

Given Δx,

minimize $f(x)$ restricted to a ray

$$\{x + t\Delta x : t \geq 0\}$$

— An exact line search is used when the cost of the minimization problem is low compared to the cost of computing the search direction itself.

2. Backtracking line search. (subroutine)

Given a decent direction Δx for f

at $x \in \mathrm{dom} f$,

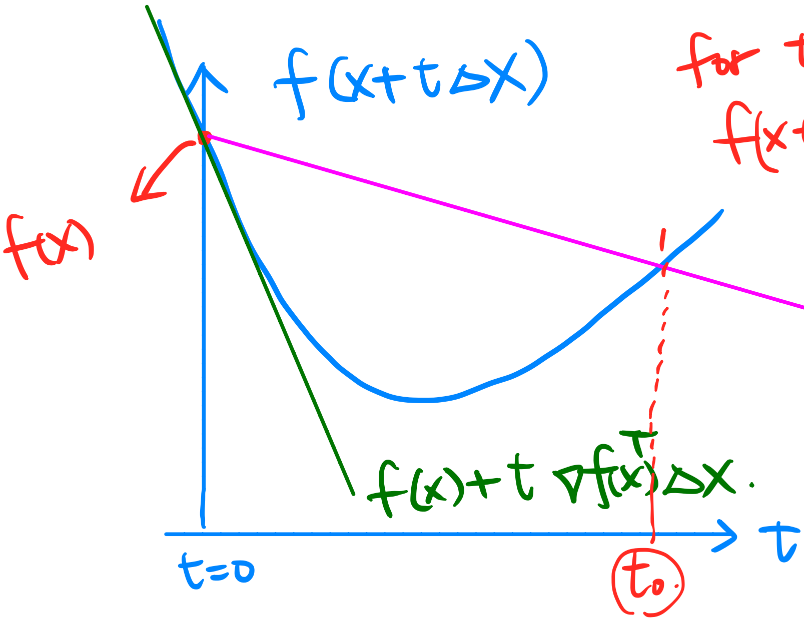
$\alpha \in (0, 0.5)$, $\beta \in (0, 1)$.

(for convergent analysis).

$t := 1$

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

$t := \beta t$ (t is reduced).



for $t \in [0, t_0]$.
$f(x+t\Delta x) \le f(x)+\alpha t \nabla f(x)^T \Delta x$.

For a chosen direction Δx, $\nabla f(x)^T \Delta x < 0$.

For small value of t

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

Thus the backtracking line search will terminate eventually.

Example: A quadratic problem $\mathbb{R}^2$.

minimize $f(x) = \frac{1}{2}\left(x_1^2 + \gamma x_2^2\right)$, where $\gamma > 0$.

— Clearly the optimal point is $x^* = 0$.
and the optimal value is 0.

— $\nabla f(x) = \begin{pmatrix} x_1 \\ \gamma x_2 \end{pmatrix}$

— Initial point $x^{(0)} = \begin{pmatrix} \gamma \\ 1 \end{pmatrix}$

update $x := x + t\Delta x = x - t\nabla f(x)$ (Gradient descent)

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

$$= \begin{pmatrix} \gamma \\ 1 \end{pmatrix} + t\begin{pmatrix} -\gamma \\ -\gamma \end{pmatrix} = \begin{pmatrix} (1-t)\gamma \\ 1-t\gamma \end{pmatrix}$$

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

Exact line search: $0 = \dfrac{\partial f(x^{(1)})}{\partial t} \Rightarrow \underline{t^{(0)} = \dfrac{2}{1+\gamma}}$

$$x^{(1)} = \begin{pmatrix} \frac{\gamma-1}{\gamma+1}\cdot\gamma \\ -\frac{\gamma-1}{\gamma+1} \end{pmatrix}$$

We have

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

$$f(x^{(k)}) = \frac{\gamma(\gamma+1)}{2}\left(\frac{\gamma-1}{\gamma+1}\right)^{2k} = \boxed{\left(\frac{\gamma-1}{\gamma+1}\right)^{2k}} f(x^{(0)}).$$

$\gamma > 0$. $\frac{\gamma-1}{\gamma+1} < 1$

- For this example. the error of the algorithm

$$f(x^{(k)}) - p^* = f(x^{(k)}) = \left(\frac{\gamma-1}{\gamma+1}\right)^{2k} f(x^{(0)}).$$

converges to zero at least as fast as a geometric series.

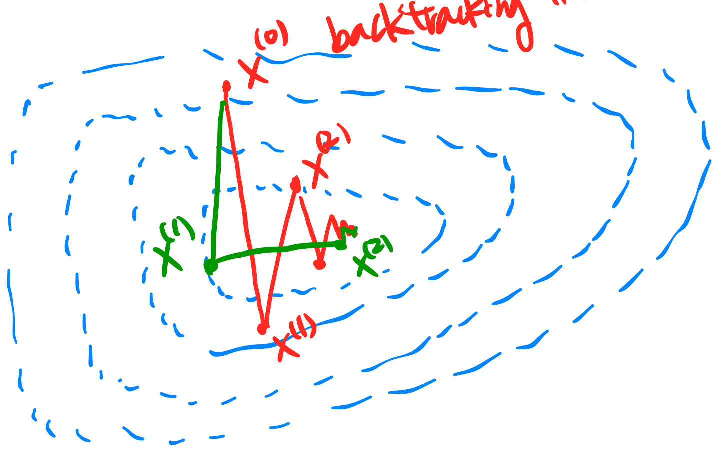
$\Rightarrow$ This is called linear convergence since the error lies below a line on a log-linear plot of error versus iteration number.

Example: A nonquadratic problem in $\mathbb{R}^2$.

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

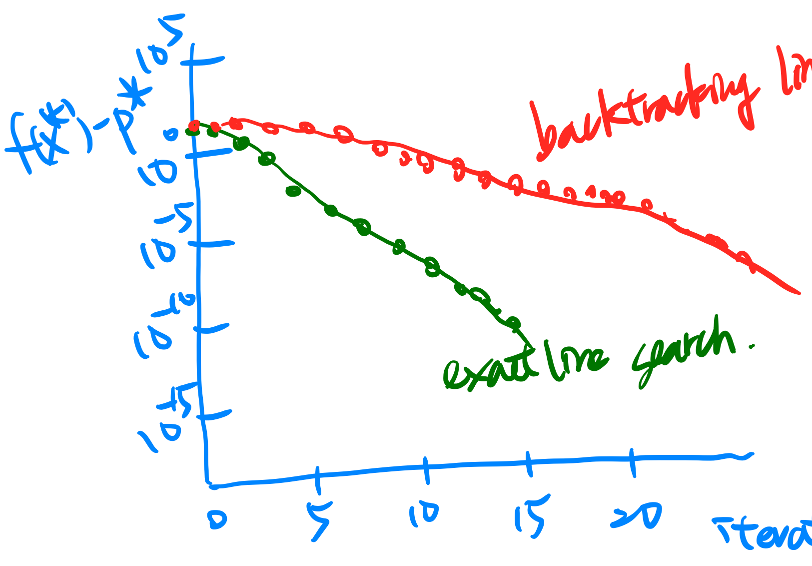
- Apply the gradient decent method
- with a backtracking line search with $\alpha = 0.1, \beta = 0.7$

Fig 9.3



level curves.

- with exact line search (Fig 9.5)



error: Initial 10, 20 iterations 10^{-7}

$10^{-8/20} \simeq 0.4$ / iteration

Initial 10, 15 iterations 10^{-10}.

$10^{-11/15} \simeq 0.2$ / iteration

Newton step: $\Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x)$.

If $\nabla^2 f(x) > 0$,

$$\nabla f(x)^T \cdot \Delta x_{nt} = -\nabla f(x)^T \left(\nabla^2 f(x)^{-1}\right) \nabla f(x) < 0.$$

$x^T A x > 0 \quad \forall x \neq 0$ if $A > 0$.

unless $\nabla f(x) = 0$. or x is optimal.

Thus the Newton step is a decent direction.

— Δx_{nt} minimizes the second-order Taylor approximation of f at x:

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

$$0 = \frac{\partial \hat{f}(x+v)}{\partial v} \Rightarrow v^* = \Delta x_{nt}. \quad \text{(Check this)}.$$

— If f is quadratic, then

$x + \Delta x_{nt}$ is the exact minimizer of f.

— For f twice differentiable, the quadratic model of f will very accurate when x is near x^*.

— $x + \Delta x_{nt}$ is a good estimate of x^*.

The Newton decrement

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

- The Newton decrement is an estimate of $f(x) - p^*$ based on the quadratic approximation $\hat{f}$ at x.

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