Unconstrained Optimization

- Optimality Conditions
    - First order conditions
    - Second order conditions

- Gradient Method
    - Descent direction
    - Steepest descent
    - Line search algorithm: Armijo, Wolfe
    - Convergence: (1) Steepest Descent with exact line search
      (2) Steepest Descent with line search
      (3) General Descent with line search
    - Trust Region method

- Newton Method
    - Newton Direction
    - convergence
    - Quasi-Newton Method
    - Barzilai-Borwein step size

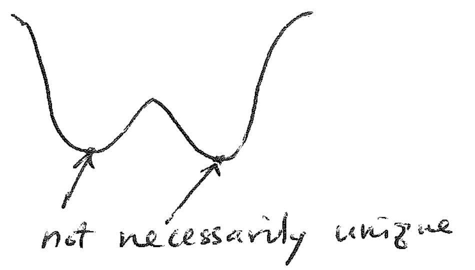
- Gradient Projection Method
    - Gradient Projection
    - Case study: GPSR by Figueiredo, Nowak and Wright
    - Proximal Gradient
    - Case study: ISTA by Beck and Teboulle

Unconstrained Optimization

General Form: $\min f(x)$
$\text{s.t. } x \in \mathbb{R}^n$

Global Minimum: Given a function $f: \mathbb{R}^n \to \mathbb{R}$, the point $x^* \in \mathbb{R}^n$ is a global minimizer if

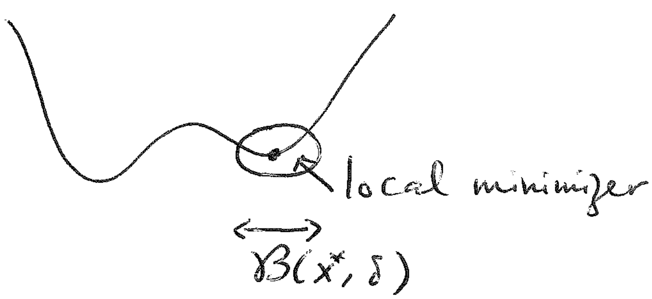
$$f(x^*) \le f(x) \quad \forall x \in \mathbb{R}^n.$$



Local Minimum: Given a function $f: \mathbb{R}^n \to \mathbb{R}$, the point x^* is a local minimizer if

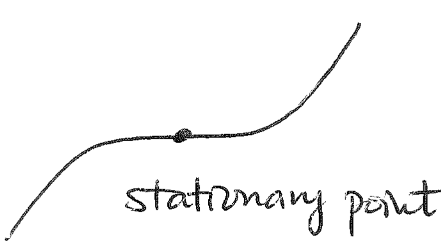
$$f(x^*) \le f(x) \quad \forall x \in \mathcal{B}(x^*, \delta).$$

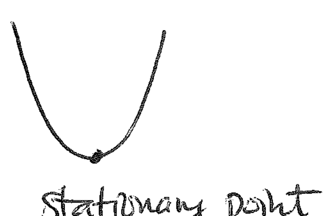
$$\mathcal{B}(x^*, \delta) = \{ x \mid \|x - x^*\| \le \delta \}$$



Stationary Point: Given a function $f: \mathbb{R}^n \to \mathbb{R}$, and assume $f \in C^1$, then a point x^* is a stationary point if

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





Optimality Condition for Local Minimizer

(Both necessary & sufficient)

Let $f: \mathbb{R}^n \to \mathbb{R}$ ~~be~~ [illegible], and assume that $\nabla f(x^*)$ exists and $\nabla^2 f(x^*)$ exists. Then, x^* is a local minimizer if and only if

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

(2) $\nabla^2 f(x^*) \geq 0$. (For sufficiency we need $\nabla^2 f(x^*) > 0$)

otherwise we can have

Some "simple" intuition:

if x^* is a local minimizer, then

$$f(x^* + th) \geq f(x^*)$$, for all t, h so that $x^* + th \in \mathcal{B}(x^*, \delta)$.

$\Rightarrow$ Taylor approximation:

$$\frac{1}{t}\left[f(x^* + th) - f(x^*)\right] = \nabla f(x^*)^T h + \frac{t}{2} h^T \nabla^2 f(\hat{x}) h$$

$x^* < \hat{x} < x^* + h$

$$\Rightarrow \lim_{t \to 0} \underbrace{\frac{1}{t}\left[f(x^* + th) - f(x^*)\right]}_{\geq 0} = \nabla f(x^*)^T h$$

, and so $\boxed{\nabla f(x^*)^T h \geq 0}$ Important!!

Since $\nabla f(x^*)^T h \geq 0$ holds for all h, the only possibility is that $\nabla f(x^*) = 0$ and so $\nabla f(x^*)^T h = 0$.

if f is twice differentiable at x^*, then Taylor approximation again

$$\lim_{t \to 0} \underbrace{\frac{1}{t^2}\left[f(x^* + th) - f(x^*)\right]}_{\geq 0} = \overbrace{\cancel{\frac{\nabla f(x^*)^T h}{t}}}^{=0} + \overbrace{\boxed{h^T \nabla f(x^*) h}}^{\text{has to be } \geq 0} + \overbrace{\frac{t}{6}\cancel{O(h^3)}}^{0 \text{ as } t \to 0} + \ldots$$

Why care about $\nabla f(x)^T h$?

If $x \neq x^*$, then we want

$$\lim_{t \to 0} \underbrace{\frac{1}{t}\left[f(x+th) - f(x)\right]}_{\leq 0} = \nabla f(x)^T h$$

→ Therefore, we want $\nabla f(x)^T h < 0$

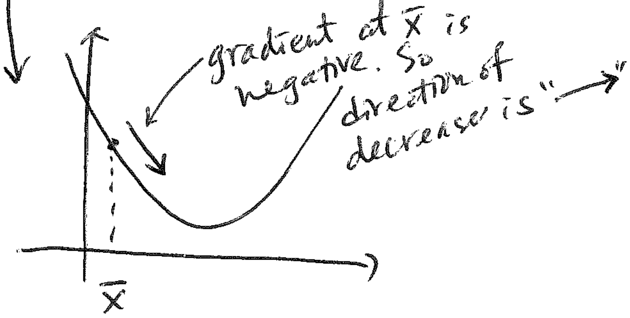
≤ 0 (so that the objective value reduces!)

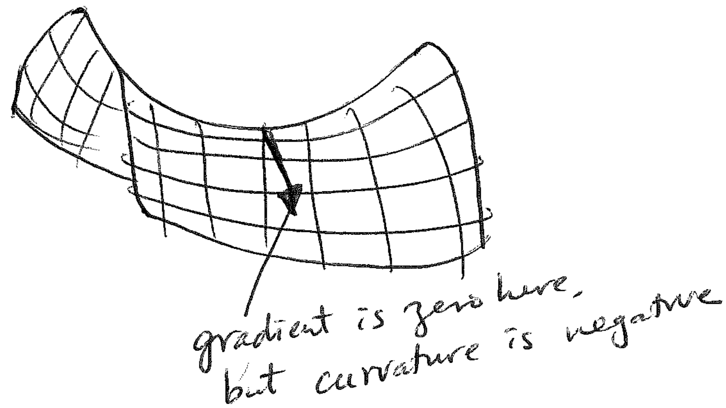
Direction of decrease for f at $\bar{x}$:

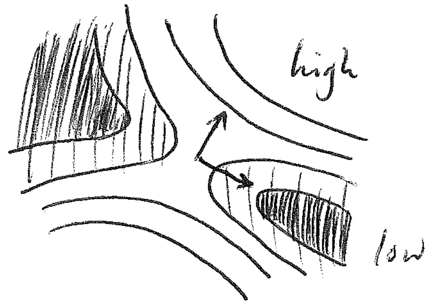
(1) h s.t. $\nabla f(\bar{x})^T h < 0$ ← direction of downhill

(2) h s.t. $\nabla f(\bar{x})^T h = 0$ and $h^T \nabla^2 f(\bar{x}) h < 0$.

direction of negative curvature







Steepest Descent Direction:

$$h = -\nabla f(\bar{x})$$

Ensures that $-\nabla f(\bar{x})^T \nabla f(\bar{x}) = -\|\nabla f(\bar{x})\|$
< 0.

Why "<u>steepest</u>" descent?

Given the current estimate x_k, the next direction should be

$$h_k = \underset{h}{\operatorname{argmin}} \nabla f(x_k)^T h$$

this optimization is unbounded.

So we put a constraint

$$h_k = \underset{\|h\|_2 = \delta}{\operatorname{argmin}} \nabla f(x_k)^T h$$

$$= \underset{h \neq 0}{\operatorname{argmin}} \frac{\delta \nabla f(x_k)^T h}{\|h\|_2}$$

(at optimal $\|h\|_2$ has to satisfy $\|h\|_2 = \delta$, so that $\frac{\delta}{\|h\|_2} = 1$).

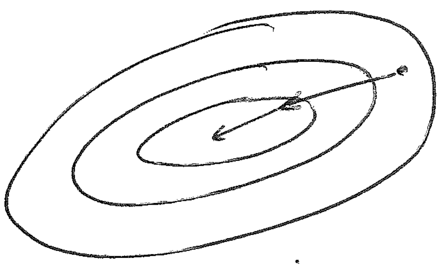
this can be solved as (by Cauchy)

$$\nabla f(x_k)^T h \geq - \|\nabla f(x_k)\|_2 \|h\|_2$$

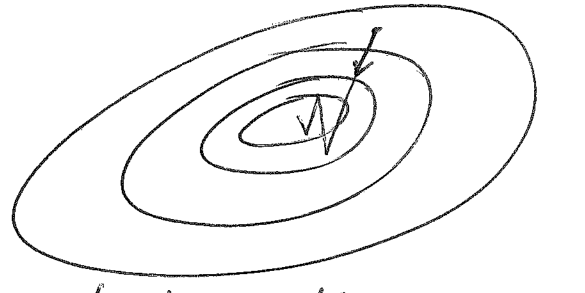
$$\Longrightarrow \quad \frac{\nabla f(x_k)^T h}{\|h\|_2} \geq - \|\nabla f(x_k)\|_2 .$$

The lower bound is attainable at $\boxed{h = -\nabla f(x_k)}$

So $h_k = -\nabla f(x_k)$ will minimizes the direction $\nabla f(x_k)^T h$.



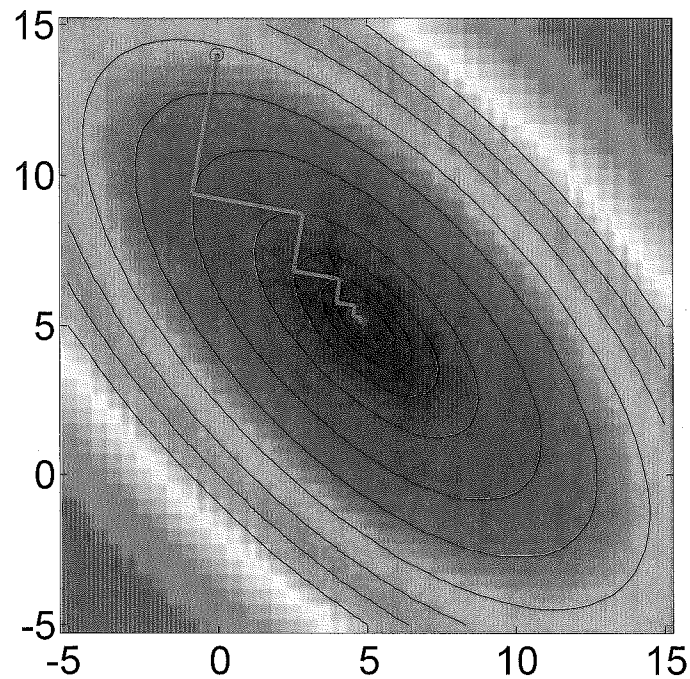
good situation



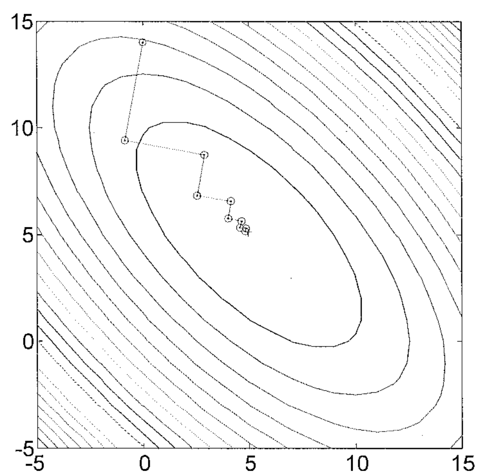
bad situation

Choosing the direction 2: steepest descent

Move in the direction of the gradient $\nabla f(\mathbf{x}_n)$



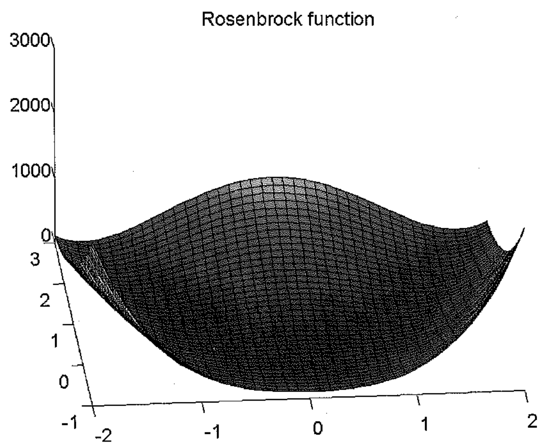
Steepest descent

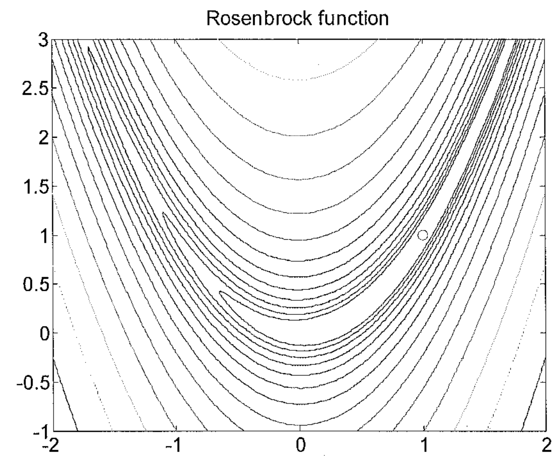


- The gradient is everywhere perpendicular to the contour lines.
- After each line minimization the new gradient is always *orthogonal* to the previous step direction (true of any line minimization.)
- Consequently, the iterates tend to zig-zag down the valley in a very inefficient manner

A harder case: Rosenbrock's function

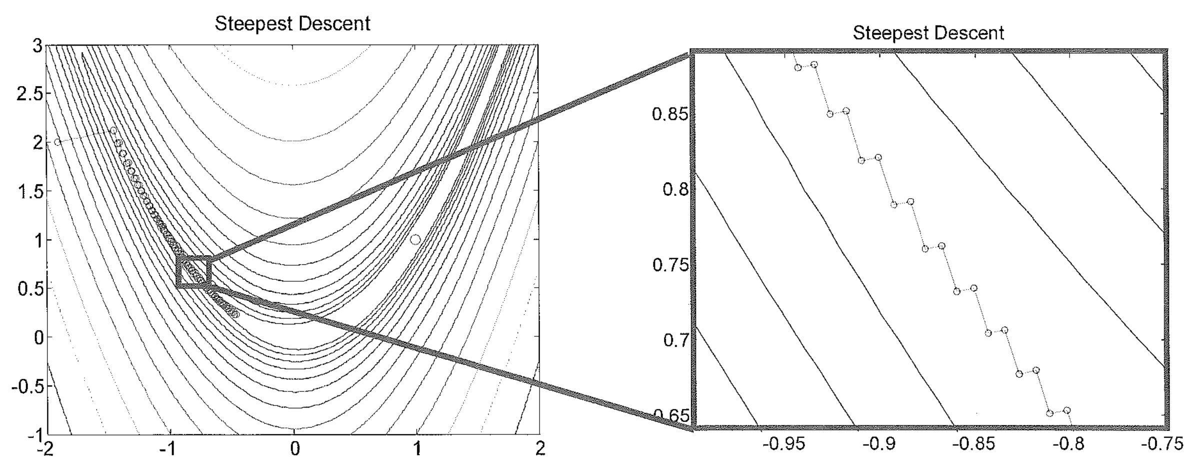
$$f(x, y) = 100(y - x^2)^2 + (1 - x)^2$$





Minimum is at $[1, 1]$

Steepest descent on Rosenbrock function



- The zig-zag behaviour is clear in the zoomed view (100 iterations)
- The algorithm crawls down the valley

Gradient Methods for Unconstrained Optimization

$$x_{k+1} = x_k + \alpha_k h_k,$$

h_k can be an descent direction: (as long as it satisfies)

(i) $\nabla f(x_k)^T h_k < 0$

(ii) $h_k = 0$ if $\nabla f(x_k) = 0$

~~(iii) h_k $\nabla f(x_k)^T h_k < 0$ and $h_k^T \nabla^2 f(x_k) h_k > 0$.~~

α_k is the step size.

How to determine step size?

(a) Minimization Rule: (Exact Line Search)

$$\alpha_k = \arg\min_{\alpha} f(x_k + \alpha h_k).$$

E.g. ~~if $h_k = -\nabla f(x_k)$, then~~

if $f(x) = \frac{1}{2} x^T H x + c^T x$, ← Quadratic programming.

then

$$f(x_k + \alpha h_k) = \frac{1}{2}(x_k + \alpha h_k)^T H (x_k + \alpha h_k) + c^T(x_k + \alpha h_k)$$

$$= \frac{1}{2} x_k^T H x_k + \frac{1}{2}\alpha^2 h_k^T H h_k + \frac{2}{2}\alpha x_k^T H h_k + c^T x_k + \alpha c^T h_k.$$

$$\frac{d}{d\alpha} = 0 \Rightarrow \alpha h_k^T H h_k + x_k^T H h_k + c^T h_k = 0$$

$$\Rightarrow \alpha = -\frac{(x_k^T H h_k + c^T h_k)}{h_k^T H h_k}.$$

But $\nabla f(x_k) = \cancel{\tfrac{1}{2}}\, H x_k + c$. So

$$\alpha = -\frac{\nabla f(x_k)^T h_k}{h_k^T H h_k}.$$

(b) Armijo Line Search

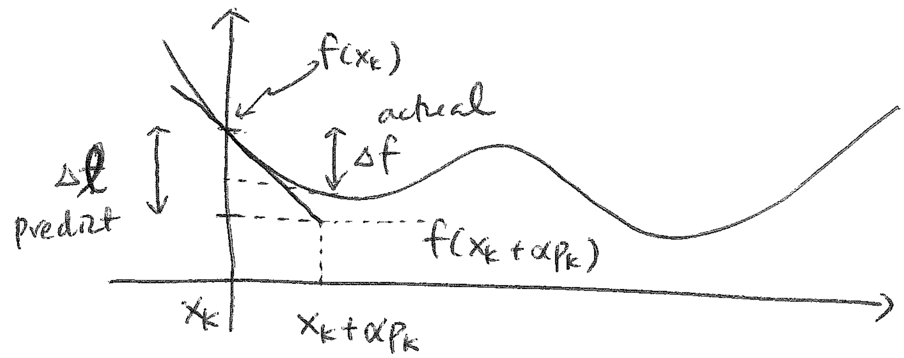
Assume that p_k is a downhill direction:

$$\nabla f(x_k)^T p_k < 0$$

Define two quantities:

$$\Delta l(\alpha p_k) \stackrel{\text{def}}{=} \alpha \nabla f(x_k)^T p_k \quad : \text{ predicted reduction}$$

$$\Delta f(\alpha p_k) \stackrel{\text{def}}{=} f(x_k + \alpha p_k) - f(x_k) : \text{ actual reduction}$$



The ratio $\dfrac{\Delta f(\alpha p_k)}{\Delta l(\alpha p_k)}$ determines the relative drop.

Note: $\dfrac{\Delta f(\alpha p_k)}{\Delta l(\alpha p_k)} \longrightarrow 1$ as $\alpha \longrightarrow 0$

happens when step size too small.

Armijo Condition:

Objective: Want Δf to be large enough.

So let $\dfrac{\Delta f(\alpha p_k)}{\Delta l(\alpha p_k)} \geq \eta_s$, $\quad 0 < \eta_s < 1$.

$$\Longleftrightarrow \boxed{f(x_k + \alpha p_k) - f(x_k) \leq \eta_s \, \alpha \nabla f(x_k)^T p_k}$$

(sign flipped because $\Delta l < 0$)

Wolfe Condition :

Armijo condition can be satisfied for small α.
We don't want α to be too small. One solution:

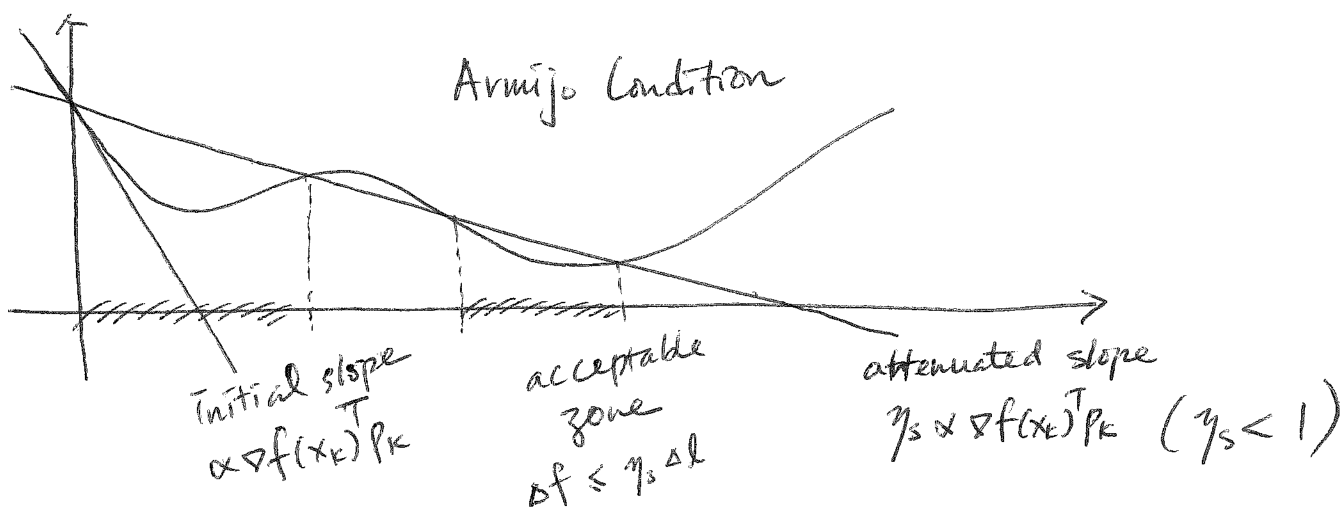
$$\left|\nabla f(x_k + \alpha p_k)^T p_k\right| \leq \left|\nabla f(x_k)^T p_k\right|$$

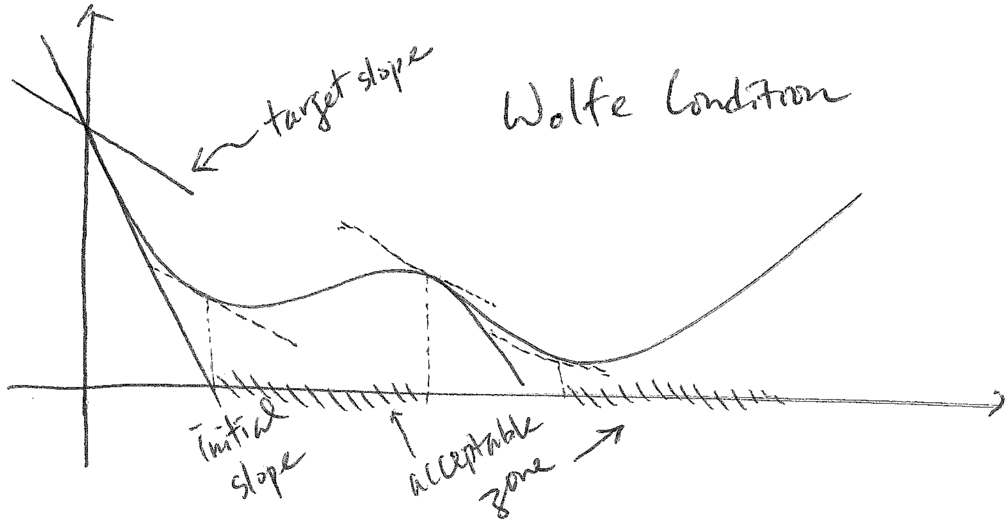
magnitude of the new downhill

magnitude of the old downhill

In practice, we can have $\eta_w < 1$

$$\boxed{\left|\nabla f(x_k + \alpha p_k)^T p_k\right| \leq \eta_w \left|\nabla f(x_k)^T p_k\right|}.$$





Convergence of Gradient Descent

1. Steepest Descent $(p_k = -\nabla f(x_k))$ with exact line search

A. Quadratic Case

Let's assume that

$$f(x) = \frac{1}{2}x^T H x + c^T x$$

Then, steepest descent has

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k),$$

where $$\alpha_k = \frac{\|\nabla f(x_k)\|^2}{\nabla f(x_k)^T H \nabla f(x_k)}.$$

Then, if we define:

$$\|x_k - x^*\|_H^2 \overset{\text{def}}{=} \frac{1}{2}(x_k - x^*)^T H (x_k - x^*)$$

$$\frac{\|x_k - x^*\|_H^2 - \|x_{k+1} - x^*\|_H^2}{\|x_k - x^*\|_H^2} = \frac{\frac{1}{2}(x_k - x^*)^T H (x_k - x^*) - \frac{1}{2}(x_{k+1} - x^*)^T H (x_{k+1} - x^*)}{\frac{1}{2}(x_k - x^*)^T H (x_k - x^*)}$$

$$= \frac{\frac{1}{2}e_k^T H e_k - \frac{1}{2}(x_k - \alpha\nabla f(x_k) - x^*)^T H (x_k - \alpha\nabla f(x_k) - x^*)}{\frac{1}{2}e_k^T H e_k}$$

$$= \frac{\frac{1}{2}e_k^T H e_k - \frac{1}{2}(e_k - \alpha\nabla f(x_k))^T H (e_k - \alpha\nabla f(x_k))}{\frac{1}{2}e_k^T H e_k}$$

$$= \frac{\cancel{\frac{1}{2}e_k^T H e_k} - \cancel{\frac{1}{2}e_k^T H e_k} + \alpha\nabla f(x_k)^T H e_k - \frac{\alpha^2}{2}\nabla f(x_k)^T H \nabla f(x_k)}{\frac{1}{2}e_k^T H e_k}$$

$$= \frac{2\alpha\nabla f(x_k)^T H e_k - \alpha^2\nabla f(x_k)^T H \nabla f(x_k)}{e_k^T H e_k}$$

Since $f(x) = \frac{1}{2}x^T H x + c^T x$, we have

$$\nabla f(x_k) = H x_k + c$$

Note that since f is quadratic, $x^* = -H^{-1}c$.

So $$He_k = H(x_k - x^*)$$

$$= Hx_k - Hx^* = Hx_k + c = \nabla f(x_k).$$

Hence,

$$\frac{\|x_k - x^*\|_H^2 - \|x_{k+1} - x^*\|_H^2}{\|x_k - x^*\|_H^2}$$

$$= \frac{2\left(\frac{\|\nabla f(x_k)\|^2}{\nabla f(x_k)^T H \nabla f(x_k)}\right)\nabla f(x_k)^T \nabla f(x_k) - \left(\frac{\|\nabla f(x_k)\|^2}{\nabla f(x_k)^T H \nabla f(x_k)}\right)^2 \nabla f(x_k)^T H \nabla f(x_k)}{\nabla f(x_k)^T H^{-1} \nabla f(x_k)}$$

$$= \frac{\left(\nabla f(x_k)^T \nabla f(x_k)\right)^2}{\left(\nabla f(x_k)^T H \nabla f(x_k)\right)\left(\nabla f(x_k)^T H^{-1} \nabla f(x_k)\right)}$$

<u>Theorem</u> (Nocedal & Wright Theorem 3.3)

if we apply steepest descent ~~to~~ with exact line search to a quadratic problem with $H > 0$, then

$$\|x_{k+1} - x^*\|_H^2 \leq \left(\frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}}\right)^2 \|x_k - x^*\|_H^2,$$

where $\lambda_{max} = \max \lambda(H)$ and $\lambda_{min} = \min \lambda(H)$.

To prove this theorem, one needs to use something called the Kantorovich Inequality

<u>Kantorovich Inequality</u> (Luenberger & Ye Ch. 8)
if $H \succ 0$, then $\forall x$,

$$\frac{x^T x}{(x^T H x)(x^T H^{-1} x)} \geq \frac{4 \lambda_{min} \lambda_{max}}{(\lambda_{min} + \lambda_{max})^2}.$$

Once we have this inequality, the proof becomes

$$\|x_{k+1} - x^*\|_H^2 \leq \left[1 - \frac{(\nabla f_k^T \nabla f_k)^2}{(\nabla f_k^T H \nabla f_k)(\nabla f_k^T H^{-1} \nabla f_k)}\right] \|x_k - x^*\|_H^2$$

$$\leq \left[1 - \frac{4 \lambda_{min} \lambda_{max}}{(\lambda_{min} + \lambda_{max})^2}\right] \|x_k - x^*\|_H^2$$

$$= \left(\frac{\lambda_{max} - \lambda_{min}}{\lambda_{min} + \lambda_{max}}\right)^2 \|x_k - x^*\|_H^2 .$$

<u>Interpretation</u>: Rate of convergence depends on the condition number $\frac{\lambda_{max}}{\lambda_{min}}$. Also, steepest descent will find the solution in single shot if $\lambda_{min} = \lambda_{max}$. This happens when $H = I$, i.e. circle.

B. Non-Quadratic Case

Theorem (Nocedal & Wright Theorem 3.4)

Assume f is twice differentiable (so $\nabla^2 f$ exists).
~~Assume~~ Use steepest descent with exact line search to find a solution x^*.
Assume $\nabla^2 f(x^*) > 0$. Then

$$f(x_{k+1}) - f(x^*) \le \left(\frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}}\right)^2 \left(f(x_k) - f(x^*)\right)$$

where $\lambda_{max} = \max \lambda\left(\nabla^2 f(x^*)\right)$.

The proof of this theorem is not given in the book.
However, there is a simpler version:

Theorem (Luenberger & Ye Ch. 8)

Same conditions as above. Then

$$f(x_{k+1}) - f(x^*) \le \left(1 - \frac{\lambda_{min}}{\lambda_{max}}\right)^2 \left(f(x_k) - f(x^*)\right).$$

~~Proof~~ (Assumption: We further assume that

$$\lambda_{min} I \le \nabla^2 f(x) \le \lambda_{max} I, \quad \forall x.$$

where λ_{min} and λ_{max} are two constants.
(Note: λ_{min} here is not necessarily ~~=~~ $\min \lambda(\nabla^2 f(x^*))$.))

Proof:

$$f(x_k - \alpha \nabla f_k) = f(x_k) - \alpha \nabla f_k^T \nabla f_k + \frac{\alpha^2}{2} \nabla f_k^T \nabla^2 f(\overbrace{\bar{x}_k}^{\text{some } \bar{x}_k}) \nabla f_k$$

$$\le f(x_k) - \alpha \|\nabla f_k\|^2 + \frac{\alpha^2 \lambda_{max}}{2} \|\nabla f_k\|^2$$

So

$$\min_{\alpha} f(x_k - \alpha \nabla f_k) \le \min_{\alpha} \left[f(x_k) - \alpha \|\nabla f_k\|^2 + \frac{\alpha^2}{2} \lambda_{max} \|\nabla f_k\|^2 \right]$$

$$\begin{array}{ccc} \| & & \| \\ f(x_{k+1}) & \le & f(x_k) - \frac{1}{2\lambda_{max}} \|\nabla f_k\|^2 \end{array}$$

$$\Rightarrow f(x_{k+1}) - f(x^*) \le f(x_k) - f(x^*) - \frac{1}{2\lambda_{max}} \|\nabla f_k\|^2 \longrightarrow (1)$$

Similarly, for any x,

$$f(x) = f(x_k) + \nabla f_k^T (x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(\bar{x})(x - x_k)$$

$$\ge f(x_k) + \nabla f_k^T (x - x_k) + \frac{\lambda_{min}}{2} \|x - x_k\|^2$$

$$\min_{x} f(x) \ge \min_{x} \left\{ f(x_k) + \nabla f_k^T (x - x_k) + \frac{\lambda_{min}}{2} \|x - x_k\|^2 \right\}$$

$$f(x^*) = f(x_k) - \frac{1}{2\lambda_{min}} \|\nabla f_k\|^2 \longrightarrow (2)$$

So (2) $\Rightarrow$ $$\frac{-1}{2} \|\nabla f_k\|^2 \le \lambda_{min} \left(f(x^*) - f(x_k) \right)$$

$$\Rightarrow f(x_{k+1}) - f(x^*) \le \left(1 - \frac{\lambda_{min}}{\lambda_{max}}\right) \left(f(x_k) - f(x^*) \right).$$

Interpretation/Remark:

(1) $$\left(\frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}} \right)^2 \cong \left(\frac{\lambda_{max} - \lambda_{min}}{\lambda_{max}} \right)^2 \text{ for large } \lambda_{max}$$

$$= \left(1 - \frac{\lambda_{min}}{\lambda_{max}} \right)^2.$$

(2) ~~what if~~ When f is not quadratic, we use a local quadratic model to approximate f.

③ Boyd & Vandenberghe (P. 467).

Let $c = 1 - \frac{\lambda_{min}}{\lambda_{max}}$. Then

$$f(x_{k+1}) - f(x^*) \le c\,(f(x_k) - f(x^*))$$
$$\le c^k (f(x_0) - f(x^*))$$

So if we want

$$f(x_{k+1}) - f(x^*) \le \epsilon,$$

then we should require

$$c^k (f(x_0) - f(x^*)) \le \epsilon$$

$$\Rightarrow k \log c \le \log\left(\frac{\epsilon}{f(x_0) - f(x^*)}\right)$$

$$\Rightarrow k \ge \frac{\log\left(\frac{\epsilon}{f(x_0) - f(x^*)}\right)}{\log c}$$

sign flip because $c<1$

$$= \frac{\log\left(\frac{f(x_0) - f(x^*)}{\epsilon}\right)}{\log\left(\frac{1}{c}\right)}$$

$$\simeq O\left(\log\left(\frac{1}{\epsilon}\right)\right).$$

This is called linear convergence in the log-linear plot of error vs iteration #.

(Supplementary)

Steepest Descent with Armijo Rule

Armijo:

$$f(x_k + \alpha p_k) \le f(x_k) + \eta_s \alpha \nabla f_k^T p_k$$

The actual algorithm:

While

$$f(x_k + \alpha p_k) > f(x_k) + \eta_s \alpha \nabla f_k^T p_k,$$

then

$$\alpha \longleftarrow \gamma \alpha, \quad \text{where } \gamma < 1$$

Theorem (Boyd & Vandenberghe p. 468)

Assume $\lambda_{min} I \le \nabla^2 f(x) \le \lambda_{max} I \quad \forall x.$

Then steepest descent with armijo rule satisfies

$$f(x_{k+1}) - f(x^*) \le c\,\big(f(x_k) - f(x^*)\big),$$

where

$$c = 1 - \min\left\{ 2\lambda_{min}\eta_s,\ 2\frac{\lambda_{min}}{\lambda_{max}}\eta_s \gamma \right\}$$

Proof: See reference.

Interpretation:

$0 < \eta_s < 0.5, \qquad 0 < \gamma < 1.$

So the extreme case is $\eta_s = 0.5,\ \gamma = 1.$

Then

$$2\frac{\lambda_{min}}{\lambda_{max}}\eta_s\gamma = \frac{\lambda_{min}}{\lambda_{max}} < \lambda_{min} \quad (\text{if } \lambda_{max} > 1).$$

$\Rightarrow \quad c = 1 - \dfrac{\lambda_{min}}{\lambda_{max}} \Rightarrow$ same ~~behavior~~ rate as exact line search

2. Gradient Descent With Line Search

- Any gradient descent direction:

$$\nabla f(x_k)^T p_k < 0$$

<u>Theorem</u> (Nocedal & Wright)

Consider a gradient descent method

$$x_{k+1} = x_k + \alpha p_k ,$$

where $\nabla f(x_k)^T p_k < 0$, and α is determined using a line search. Assume that f is bounded below, i.e. $f(x) > -\infty$, and assume that ∇f is Lipschitz:

$$\| \nabla f(x) - \nabla f(x') \| \leq L \| x - x' \| , \quad \forall x, x' ,$$

then $\quad \| \nabla f(x_k) \| \longrightarrow 0 \quad$ as $\quad k \longrightarrow \infty$.

<u>Proof</u>: Wolfe condition implies

$$| \nabla f(x_k + \alpha p_k)^T p_k | \leq \eta_w | \nabla f(x_k)^T p_k |$$

$$\Longrightarrow \underbrace{-\eta_w | \nabla f(x_k)^T p_k | \leq \nabla f(x_k + \alpha p_k)^T p_k} \leq \eta_w | \nabla f(x_k)^T p_k |$$

$$\Longrightarrow - \nabla f(x_k + \alpha p_k)^T p_k \leq -\eta_w \nabla f(x_k)^T p_k \leq \nabla f(x_k + \alpha p_k)^T p_k$$

$$\Longrightarrow \underbrace{\nabla f(x_k + \alpha p_k)^T p_k \geq \eta_w \nabla f(x_k)^T p_k} \geq - \nabla f(x_k + \alpha p_k)^T p_k$$

So we have

$$\nabla f(x_k + \alpha p_k)^T p_k \geq \eta_\omega \nabla f(x_k)^T p_k$$

$$\Rightarrow \quad \nabla f(x_{k+1})^T p_k \geq \eta_\omega \nabla f(x_k)^T p_k$$

$$\Rightarrow \quad \nabla f(x_{k+1})^T p_k - \nabla f(x_k)^T p_k \geq (\eta_\omega - 1) \nabla f(x_k)^T p_k$$

$$\Rightarrow \quad \left[\nabla f(x_{k+1}) - \nabla f(x_k)\right]^T p_k \geq (\eta_\omega - 1) \nabla f(x_k)^T p_k$$

By Lipschitz condition of f, we have that

$$\begin{aligned}
&\left[\nabla f(x_{k+1}) - \nabla f(x_k)\right]^T p_k \\
&\leq \| \nabla f(x_{k+1}) - \nabla f(x_k) \|_2 \, \| p_k \|_2 \\
&\leq L \, \| x_{k+1} - x_k \|_2 \, \| p_k \|_2 \\
&= L \alpha \| p_k \|_2^2 .
\end{aligned}$$

Therefore, $L \alpha \| p_k \|^2 \geq (\eta_\omega - 1) \nabla f(x_k)^T p_k$

$$\Rightarrow \quad \alpha \geq \frac{\eta_\omega - 1}{L \| p_k \|^2} \nabla f(x_k)^T p_k$$

The Armijo condition gives

$$\begin{aligned}
f(x_k + \alpha p_k) &\leq f(x_k) + \eta_s \alpha \overset{<0}{\boxed{\nabla f(x_k)^T p_k}} \\
&\leq f(x_k) + \eta_s \left(\frac{\eta_\omega - 1}{L \| p_k \|^2} \right) \nabla f(x_k)^T p_k \\
&= f(x_k) + \frac{\eta_s (\eta_\omega - 1)}{L} \underbrace{\frac{\nabla f(x_k)^T p_k}{\| \nabla f(x_k) \|^2 \| p_k \|^2}}_{= \cos^2 \theta_k} \| \nabla f(x_k) \|^2
\end{aligned}$$

So $\underbrace{f(x_k + \alpha p_k)}_{f(x_{k+1})} \leq f(x_k) + \dfrac{\eta_s(\eta_\omega - 1)}{L} \cos^2\theta_k \, \|\nabla f(x_k)\|^2$

$$\Rightarrow \quad f(x_k) - f(x_{k+1}) \geq \frac{\eta_s(1-\eta_\omega)}{L} \cos^2\theta_k \, \|\nabla f(x_k)\|^2$$

$$\Rightarrow \quad \underbrace{f(x_0) - f(x_{k+1})}_{<\infty} \geq \frac{\eta_s(1-\eta_\omega)}{L} \sum_{j=0}^{k} \cos^2\theta_j \, \|\nabla f(x_j)\|^2$$

because f is bounded below

So $\displaystyle\sum_{j=0}^{k} \cos^2\theta_j \, \|\nabla f(x_j)\|^2 < \infty$.

Since $\nabla f(x_k)^T p_k < 0 \;\forall k$, $\cos^2\theta_k > 0 \;\forall k$.

Therefore, we must have

$$\|\nabla f(x_k)\|^2 \longrightarrow 0 \text{ as } k \longrightarrow \infty.$$

Limitation of this Theorem:

(i) only guarantees $\|\nabla f\| \to 0$

(ii) this could happen for stationary points

So if we want something stronger, we need to use $\nabla^2 f$.

Trust Region Method

Define a trust-region radius δ_k.

Then, solve the minimization

$$\min_{d} \nabla f(x_k)^T d$$
$$\text{s.t. } \|d\|_2 \le \delta_k . \qquad \text{———} \quad (*)$$

Since $\nabla f(x_k)^T d$ is unbounded below, the solution of $(*)$ must lie on the boundary of δ_k. Note that $\{d \mid \|d\|_2 \le \delta_k\}$ is a closed bounded set, the existence of the solution of $(*)$ is guaranteed by the Extreme Value Theorem.

Solution to $(*)$: $$d_k = - \frac{\delta_k}{\|\nabla f(x_k)\|_2} \nabla f(x_k)$$

(Can be found using Cauchy-inequality).

How to update δ_k?

Compute the ratio

actual value

$$\rho_k = \frac{f(x_k + d_k) - f(x_k)}{\nabla f(x_k+d_k)^T d_k - \nabla f(x_k)^T d_k}$$

predicted value

if $\rho_k \ge \eta_s$, then $x_{k+1} = x_k + d_k$

if $\rho_k < \eta_s$, then $x_{k+1} = x_k$, and $\delta_{k+1} = \gamma \delta_k$.