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: \[ \min f(x) \quad \text{s.t. } x \in \mathbb{R}^n \]

Global Minimum: Given a function \( f: \mathbb{R}^n \to \mathbb{R} \), the point \( x^* \in \mathbb{R}^n \) is a global minimizer if \( f(x^*) \leq f(x) \quad \forall x \in \mathbb{R}^n \).

Local Minimum: Given a function \( f: \mathbb{R}^n \to \mathbb{R} \), the point \( x^* \) is a local minimizer if \[ f(x^*) \leq f(x) \quad \forall x \in B(x^*, \delta) \]

where \( B(x^*, \delta) = \left\{ x \mid \| x - x^* \| \leq \delta \right\} \)

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 \( f: \mathbb{R}^n \rightarrow \mathbb{R} \) be \( \mathcal{C}^2 \), 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

\[ f(x^* + th) \geq f(x^*) \quad \text{for all } t, h \text{ so that } x^* + th \in B(x^*, \delta). \]

\[ \Rightarrow \text{Taylor approximation:} \]

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

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

\[ \geq 0 \quad \text{and so} \]

\[ \nabla f(x^*)^T h \geq 0 \]

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

\[ \lim_{t \to 0} \frac{1}{t^2} \left[ f(x^* + th) - f(x^*) \right] = \nabla^2 f(x^*)^T h + \frac{1}{2} h^T \nabla^2 f(x^*) h + \frac{1}{6} O(h^3) \]

\[ \geq 0 \quad \text{has to be} \]

\[ \geq 0 \quad \text{and} \]

\[ \geq 0 \]
Why care about $\nabla f(x)^T h$?

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

$$\lim_{t \to 0} \frac{1}{t} \left[ f(x + th) - f(x) \right] = \nabla f(x)^T h \leq 0 \quad \text{(so that the objective value reduces!)}$$

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

Direction of decrease for $f$ at $x$:

1. $h$ s.t. $\nabla f(x)^T h < 0$ (direction of downhill)
2. $h$ s.t. $\nabla f(x)^T h \leq 0$ and $h^T \nabla^2 f(x) h < 0$ (direction of negative curvature)

Steepest Descent Direction:

$$h = -\nabla f(x)$$

Ensures that $-\nabla f(x)^T \nabla f(x) = -\|\nabla f(x)\|^2 < 0$. 

\[8\]
Why "steepest" descent?

Given the current estimate $x_k$, the next direction should be

$$h_k = \arg\min_h \nabla f(x_k)^T h$$

this optimization is unbounded.

So we put a constraint

$$h_k = \arg\min_{\|h\|_2 = \delta} \nabla f(x_k)^T h$$

$$= \arg\min_{h \neq 0} \frac{\delta \nabla f(x_k)^T h}{\|h\|_2}$$

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

This can be solved as (by Cauchy)

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

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

So the lower bound is attainable at $h = -\nabla f(x_k)$.

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

![good situation](image1)
![bad situation](image2)
Choosing the direction 2: steepest descent

Move in the direction of the gradient $\nabla f(x_n)$

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) \( h_k \neq 0 \)  \( \nabla f(x_k)^T h_k < 0 \)

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

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

then

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

\[ = \frac{1}{2} x_k^T H x_k + \frac{1}{2} \alpha^2 h_k^T H h_k + \frac{\alpha}{2} x_k^T H h_k \]

\[ + c^T x_k + \alpha c^T h_k. \]

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

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

But \( \nabla f(x_k) = \frac{1}{2} H x_k + c \). 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:

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

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

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

Note:

$$\frac{\Delta f(\alpha p_k)}{\Delta l(\alpha p_k)} \rightarrow 1 \text{ as } \alpha \rightarrow 0$$

happens when step size too small.

**Armijo Condition:**

Objective: Want $\Delta f$ to be large enough.

So let

$$\frac{\Delta f(\alpha p_k)}{\Delta l(\alpha p_k)} \geq \eta_s , \quad 0 < \eta_s < 1$$

$$(\Rightarrow) \quad f(x_k + \alpha p_k) - f(x_k) \leq \eta_s \alpha \nabla f(x_k)^T p_k$$

(sign flipped because $\Delta l < 0$)
Wolfe Condition:

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

$$|\nabla f(x_k + \alpha p_k)^T p_k| \leq \frac{|\nabla f(x_k)^T p_k|}{\alpha}$$

and

magnitude of the downhill

magnitude of the downhill

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

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

\[\text{Armijo Condition}\]

Initial slope $\alpha \nabla f(x_k)^T p_k$

acceptable zone $\nabla f \leq \gamma_w \alpha k$

attenuated slope $\gamma_w \alpha \nabla f(x_k)^T p_k$ ($\gamma_w < 1$)

\[\text{Wolfe Condition}\]

initial slope

target slope

acceptable zone
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:
\[
\|x_k - x^*\|_H^2 \overset{\text{def}}{=} \frac{1}{2} (x_k - x^*)^T H (x_k - x^*)
\]
\[
\|x_k - x^*\|_H^2 - \|x_{k+1} - x^*\|_H^2 = \frac{\|x_k - x^*\|_H^2}{2 (x_k - x^*)^T H (x_k - x^*)} \left( \frac{1}{2} (x_k - x^*)^T H (x_k - x^*) - \frac{1}{2} (x_{k+1} - x^*)^T H (x_{k+1} - x^*) \right)
\]
\[
= \frac{1}{2} e_k^T H e_k - \frac{1}{2} \left( \frac{(x_k - \alpha \nabla f(x_k) - x^*)^T H (x_k - \alpha \nabla f(x_k) - x^*)}{2 (x_k - x^*)^T H (x_k - x^*)} \right)
\]
\[
= \frac{1}{2} e_k^T H e_k - \frac{1}{2} \left( e_k + \alpha \nabla f(x_k) \right)^T H \left( e_k + \alpha \nabla f(x_k) \right)
\]
\[
= \frac{1}{2} e_k^T H e_k - \frac{1}{2} e_k^T H e_k + \alpha \nabla f(x_k)^T H \alpha e_k + \frac{\alpha^2}{2} \nabla f(x_k)^T H \nabla f(x_k)
\]
\[
= \frac{1}{2} e_k^T H e_k - \frac{1}{2} e_k^T H e_k + \alpha \nabla f(x_k)^T H \alpha e_k + \frac{\alpha^2}{2} \nabla f(x_k)^T H \nabla f(x_k)
\]
\[
= 2 \alpha \nabla f(x_k)^T H e_k - \alpha^2 \nabla f(x_k)^T H \nabla f(x_k)
\]
\[
= \alpha \nabla f(x_k)^T H e_k
\]
Since \( f(x) = \frac{1}{2}x^T H x + c^T x \), we have
\[
\nabla f(x_k) = Hx_k + c
\]
Note that since \( f \) is quadratic, \( x^* = -H^{-1}c \).
So \( Hx_k = H(x_k - x^*) \)
\[
= Hx_k - (Hx^* = Hx_k + c = \nabla f(x_k)).
\]

Hence,
\[
\frac{||x_k - x^*||_H^2 - ||x_{k+1} - x^*||_H^2}{||x_k - x^*||_H^2} = 2 \left( \frac{||\nabla f(x_k)||_H^2}{\nabla f(x_k)^T H \nabla f(x_k)} \right) \nabla f(x_k)^T H^{-1} \nabla f(x_k) \left( \frac{||\nabla f(x_k)||_H^2}{\nabla f(x_k)^T H \nabla f(x_k)} \right)
\]
\[
= \left( \frac{\nabla f(x_k)^T H \nabla f(x_k)}{\nabla f(x_k)^T H \nabla f(x_k)} \right)^2 \left( \frac{||\nabla f(x_k)||_H^2}{\nabla f(x_k)^T H \nabla f(x_k)} \right)
\]

**Theorem (Nocedal & Wright Theorem 3.3)**
If we apply steepest descent 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.

**Kantorovich Inequality** (Luenberger & Ye Ch. 8)

If $H > 0$, 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\[ \|x^{k+1} - x^*\|^2_H \leq \left[ \frac{\|Df_k^T Df_k\|^2}{(Df_k^T HDf_k)(Df_k^T H^{-1} Df_k)} \right] \|x_k - x^*\|^2_H \]
\[ \leq 1 - \frac{4 \lambda_{\min} \lambda_{\max}}{(\lambda_{\min} + \lambda_{\max})^2} \|x_k - x^*\|^2_H \]
\[ = \left( \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\min} + \lambda_{\max}} \right)^2 \|x_k - x^*\|^2_H. \]

**Interpretation**: 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).
Use steepest descent to find a solution $x^*$, with exact line search.
Assume $\nabla^2 f(x^*) > 0$. Then
\[
f(x_{k+1}) - f(x^*) \leq \left( \frac{\lambda_{\text{max}} - \lambda_{\text{min}}}{\lambda_{\text{max}} + \lambda_{\text{min}}} \right)^2 (f(x_k) - f(x^*))
\]
where $\lambda_{\text{max}} = \max \lambda(\nabla^2 f(x^*))$.

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
\[
f(x_{k+1}) - f(x^*) \leq (1 - \frac{\lambda_{\text{min}}}{\lambda_{\text{max}}})^2 (f(x_k) - f(x^*))
\]

(Thesis: We further assume that $\lambda_{\text{min}} I \leq \nabla^2 f(x) \leq \lambda_{\text{max}} I$, $\forall x$. where $\lambda_{\text{min}}$ and $\lambda_{\text{max}}$ are two constants.

(Note: $\lambda_{\text{min}}$ here is not necessarily $\min \lambda(\nabla^2 f(x^*))$)

Proof:
\[
f(x_k - \alpha \nabla f_k) = f(x_k) - \alpha \nabla f_k^T \nabla f_k + \frac{\alpha^2}{2} \nabla f_k^T \nabla^2 f(x_k) \nabla f_k
\]
\[
\leq f(x_k) - \alpha \| \nabla f_k \|^2 + \frac{\alpha^2 \lambda_{\text{max}}}{2} \| \nabla f_k \|^2
\]
So
\[
\min_{x} f(x_k - \alpha \nabla f_k) \leq \min_{x} \left[ f(x_k) - \alpha \|
abla f_k \|^2 + \frac{\alpha^2}{2} \lambda_{\max} \|
abla f_k \|^2 \right] \leq f(x_{k+1}) \leq f(x_k) - \frac{1}{2\lambda_{\max}} \|
abla f_k \|^2 \Rightarrow f(x_{k+1}) - f(x^*) \leq f(x_k) - f(x^*) - \frac{1}{2\lambda_{\max}} \|
abla f_k \|^2 \rightarrow (1)
\]

Similarly, for any \( x \),
\[
f(x) = f(x_k) + \nabla f_k^T (x - x_k) + \frac{1}{2} (x - x_k)^T \nabla^2 f(x) (x - x_k) \geq f(x_k) + \nabla f_k^T (x - x_k) + \frac{\lambda_{\min}}{2} \|x - x_k\|^2
\]
\[
\min_{x} f(x) \geq \min_{x} \left\{ f(x_k) + \nabla f_k^T (x - x_k) + \frac{\lambda_{\min}}{2} \|x - x_k\|^2 \right\}
\]
\[
f(x^*) = f(x_k) - \frac{1}{2\lambda_{\min}} \|
abla f_k \|^2 \rightarrow (2)
\]

So (2) \( \Rightarrow \) \( \frac{1}{2\lambda_{\max}} \|
abla f_k \|^2 \leq \lambda_{\min} (f(x^*) - f(x_k)) \Rightarrow f(x_{k+1}) - f(x^*) \leq (1 - \frac{\lambda_{\min}}{\lambda_{\max}}) (f(x_k) - f(x^*)) \).

Interpretation/Remark:

1. \( \left( \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \right)^2 \leq \left( \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max}} \right)^2 \) for large \( \lambda_{\max} \)

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

When

2. What if \( f \) 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

$$f(x_{k+1}) - f(x^*) \leq c \left( f(x_k) - f(x^*) \right) \leq c^k (f(x_0) - f(x^*))$$

So if we want

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

then we should require

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

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

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

Sign flip $\log c$ because $c < 1$

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

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

This is called linear convergence in the log-linear plot of error vs iteration $k$. 
Steepest Descent with Armijo Rule

**Armijo:**

\[ f(x_k + \alpha P_k) \leq f(x_k) + \eta_s \alpha \nabla f_k^T P_k \]

The actual algorithm:

While

\[ f(x_k + \alpha P_k) > f(x_k) + \eta_s \alpha \nabla f_k^T P_k, \]

then

\[ \alpha \leftarrow \gamma \alpha, \quad \text{where} \quad \gamma < 1 \]

**Theorem** (Boyd & Vandenberghe p. 468)

Assume \( \lambda_{\max} \leq \nabla^2 f(x) \leq \lambda_{\max} I \) \( \forall x \).

Then steepest descent with armijo rule satisfies

\[ f(x_{k+1}) - f(x^*) \leq C (f(x_k) - f(x^*)) \]

where

\[ C = 1 - \min \left\{ 2 \lambda_{\min} \eta_s, \, 2 \frac{\lambda_{\min} \eta_s}{\lambda_{\max}} \right\} \]

**Proof:** See reference.

**Interpretation:**

\[ 0 < \eta_s < 0.5, \quad 0 < \gamma < 1 \]

So the extreme case is \( \eta_s = 0.5, \quad \gamma = 1 \).

Then

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

\[ \Rightarrow \quad C = 1 - \frac{\lambda_{\min}}{\lambda_{\max}} \quad \Rightarrow \quad \text{same behavior rate as exact line search} \]
2. Gradient Descent with Line Search

- Any gradient descent direction:
  \[ \nabla f(x_k)^T P_k < 0 \]

**Theorem** (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 \( \alpha \) is determined using a line search. Assume that \( f \) is bounded below, i.e. \( f(x) \geq -\infty \), and assume that \( \nabla f \) is Lipschitz:

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

then

\[ \| \nabla f(x_k) \| \to 0 \quad \text{as} \quad k \to \infty. \]

**Proof:** Wolfe condition implies

\[ \left| \nabla f(x_k + \alpha P_k)^T P_k \right| \leq \gamma_{\text{W}} \left| \nabla f(x_k)^T P_k \right| \]

\[ \Rightarrow -\gamma_{\text{W}} \nabla f(x_k)^T P_k \leq \nabla f(x_k + \alpha P_k)^T P_k \leq \gamma_{\text{W}} \nabla f(x_k)^T P_k \]

\[ \Rightarrow \nabla f(x_k^*)^T P_k \leq -\gamma_{\text{W}} \nabla f(x_k)^T P_k \leq \nabla f(x_k + \alpha P_k)^T P_k \]

\[ \Rightarrow \nabla f(x_k + \alpha P_k)^T P_k \geq \gamma_{\text{W}} \nabla f(x_k)^T P_k \geq -\nabla f(x_k + \alpha P_k)^T P_k \]
So we have
\[ \nabla f(x_k + \alpha p_k)^T p_k \geq \eta \nabla f(x_k)^T p_k \]
\[ \Rightarrow \quad \nabla f(x_{k+1})^T p_k \geq \eta \nabla f(x_k)^T p_k \]
\[ \Rightarrow \quad \nabla f(x_{k+1})^T p_k - \nabla f(x_k)^T p_k \geq (\eta - 1) \nabla f(x_k)^T p_k \]
\[ \Rightarrow \quad \left[ \nabla f(x_{k+1}) - \nabla f(x_k) \right]^T p_k \geq (\eta - 1) \nabla f(x_k)^T p_k \]

By Lipschitz condition of \( f \), we have that
\[ \left\| \nabla f(x_{k+1}) - \nabla f(x_k) \right\|_2 \leq L \left\| x_{k+1} - x_k \right\|_2 \cdot \left\| p_k \right\|_2 \]
\[ \leq L \alpha \left\| p_k \right\|_2^2 \]

Therefore,
\[ \alpha \left\| p_k \right\|_2^2 \geq (\eta - 1) \nabla f(x_k)^T p_k \]
\[ \Rightarrow \quad \alpha \geq \frac{\eta - 1}{L \left\| p_k \right\|_2^2} \nabla f(x_k)^T p_k \]

The Armijo condition gives
\[ f(x_k + \alpha p_k) \leq f(x_k) + \eta s \left( \frac{\eta - 1}{L \left\| p_k \right\|_2^2} \nabla f(x_k)^T p_k \right) \]
\[ \leq f(x_k) + \eta s \left( \frac{\eta - 1}{L \left\| p_k \right\|_2^2} \right) \nabla f(x_k)^T p_k \]
\[ = f(x_k) + \eta s \left( \frac{\eta - 1}{L} \right) \frac{\nabla f(x_k)^T p_k}{\left\| \nabla f(x_k) \right\|_2 \left\| p_k \right\|_2} \left\| \nabla f(x_k) \right\|_2^2 \]
\[ = \cos^2 \Theta_k \]
So \( f(x_k + \alpha \beta_k) \leq f(x_k) + \frac{\gamma_s (\gamma - 1)}{L} \cos^2 \theta_k \| Df(x_k) \|^2 \)

\[ \Rightarrow \]

\[ f(x_k) \leq f(x_{k+1}) \]

\[ \Rightarrow f(x_k) - f(x_{k+1}) \geq \frac{\gamma_s (1-\gamma)}{L} \cos^2 \theta_k \| Df(x_k) \|^2 \]

\[ \Rightarrow f(x_k) - f(x_{k+1}) \geq \frac{\gamma_s (1-\gamma)}{L} \sum_{j=0}^{k} \cos^2 \theta_j \| Df(x_j) \|^2 \]

\[ < \infty \]

because \( f \) is bounded below

\[ \sum_{j=0}^{k} \cos^2 \theta_j \| Df(x_j) \|^2 < \infty \]

Since \( Df(x_k)^T \beta_k < 0 \ \forall \ k, \ \cos^2 \theta_k > 0 \ \forall \ k \).

Therefore, we must have

\[ \| Df(x_k) \|^2 \rightarrow 0 \ \text{as} \ k \rightarrow \infty. \]

\[ \text{Limitation of this Theorem:} \]

(i) only guarantees \( \| Df \| \rightarrow 0 \)

(ii) this could happen for stationary points

So if we want something stronger, we need to use \( D^2 f \).
Trust Region Method

Define a trust-region radius $\delta_k$.

Then, solve the minimization

\[
\min_{d} \nabla f(x_k)^T d
\]

s.t. $\|d\|_2 \leq \delta_k$.

Since $\nabla f(x_k)^T d$ is unbounded below, the solution of \((*)\) must lie on the boundary of $\delta_k$. Note that $\{d | \|d\|_2 \leq \delta_k\}$ 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 $\delta_k$?

Compute the ratio

\[
l_k = \frac{f(x_k + d_k) - f(x_k)}{\nabla f(x_k + d_k)^T d_k - \nabla f(x_k)^T d_k}
\]

if $l_k \geq \gamma_s$, then $x_{k+1} = x_k + d_k$

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

17