Unconstrained Optimization

Consider the minimization problem

$$min_{x \in R^n} f(x)$$

- **Definition :** $x^* \in R^n$ is said to be a
  - global minima of f if $f(x^*) \leq f(x)$ for all x $\in R^n$.
  - a strict global minima of f if $f(x^*) < f(x)$ for all x $\in R^n$.
  - a local minima of f if ther exists a neighbourhood of x^* $(N(x^*, \varepsilon), \varepsilon > 0)$ such that $f(x^*) \leq f(x)$ for all x $\in N(x^*, \varepsilon)$.
- If f is a convex function then every local minima of f is a global minima of f.
- If f is a strict convex function then f has unique minima.

- **Descent direction:** Suppose $f: R^n \rightarrow R$ be a continuous function. $d \in R^n$ is said to be a decent direction of f at x if there exists $\overline{\alpha} > 0$ such that

$$f(x + \alpha d) < f(x)$$

for all $\alpha \in (0, \overline{\alpha})$.

Example: Suppose $f(x) = x_1^2 + x_1^2$ and $x = (1,1)^T$. $d = (-4,1)^T$ is a descent direction of f at x since

$$f(x + \alpha d) < f(x)$$

for all $\alpha \in (0, 0.3)$.

- x^* is a local minima of f if and only if there does not exist any descent direction of f at x^*.

- Suppose $f: R^n \to R$ be a differentiable function. If $\nabla f(x)^T d < 0$ for some $d \in R^n$ then d is a descent direction of f at x.

In the previous example $\nabla f(x) = \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix}$. So $\nabla f\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}\right) = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$

Cleary $\left(\nabla f\begin{pmatrix} 1 \\ 1 \end{pmatrix}\right)^T d = \begin{pmatrix} 2 \\ 2 \end{pmatrix}^T \begin{pmatrix} -4 \\ 1 \end{pmatrix} = -8 + 2 = -6 < 0$

Hence $d = \begin{pmatrix} -4 \\ 1 \end{pmatrix}$ is a descent direction of f at $x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

- **First order necessary condition:** Suppose $f: R^n \to R$ be a differentiable function. If $x^* \in R^n$ is a local minima of f then $\nabla f(x^*) = 0$.

Proof: Suppose $x^* \in R^n$ is a local minima of f and $\nabla f(x^*) \neq 0$.

Let $d = -\nabla f(x^*)$. Then

$$\left(\nabla f(x^*)\right)^T d = -\left(\nabla f(x^*)\right)^T \nabla f(x^*) = -\|\nabla f(x^*)\|^2 < 0.$$

This implies d is a descent of f at x^*. Hence there exists $\alpha > 0$ such that $f(x^* + \alpha d) < f(x^*)$. This contradicts that x^* is a local minima of f.

- For $f(x) = x_1^2 + x_1^2$ then $x^* = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ is a local minima.

Clearly $\nabla f(x^*) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$.

- Note that $\nabla f(x^*) = 0$ does not imply x^* is a local minima of f.
- For example let $f(x) = x_1^2 + x_1^3$ then $\nabla f(x) = \begin{pmatrix} 2x_1 \\ 3x_2^2 \end{pmatrix}$. For $x^* = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ we have $\nabla f(x^*) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$. But $x^* = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ is not a local minima of f since $f\left(\begin{pmatrix} 0 \\ -0.01 \end{pmatrix}\right) = -0.000001 < f\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}\right)$.
- Suppose $f: R^n \to R$ be a differentiable function and $\nabla f(x^*) = 0$ for some $x^* \in R^n$. Then x^* is said to be a stationary point of f.
- Clearly every local minima is a stationary point but every stationary point is not a local minima.

- **Second order sufficient condition:** Suppose $f: R^n \to R$ be a twice differentiable function. If for $x^* \in R^n$ $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite, then x^* is a local minima of f.
- For example suppose

$f(x) = 4x_1^2 + x_2^2 - 2x_1x_2$

Then $\nabla f(x) = \begin{pmatrix} 8x_1 - 2x_2 \\ -2x_1 + 2x_2 \end{pmatrix}$ and $\nabla^2 f(x) = \begin{pmatrix} 8 & -2 \\ -2 & 2 \end{pmatrix}$

Clearly for $x^* = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\nabla f(x^*) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\nabla^2 f(x^*) = \begin{pmatrix} 8 & -2 \\ -2 & 2 \end{pmatrix}$

which is positive definite since leading principle minors of $\begin{pmatrix} 8 & -2 \\ -2 & 2 \end{pmatrix}$ are $8\ (>0)$ and $12 (> 0)$.

Hence $x^* = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ is a local minima of f.

Since f is a convex function $x^* = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ is a global minima of f.

Since f is a strictly convex function $x^* = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ is the unique global minima of f.

- Descent methods:

  Input: Unconstrained objective function (f) and initial approximation (x^0)

  Output: An approximate stationary point (x^*)

  Algorithm:

  - **Step 0 (Initialization):** Supply f, x^0, $\varepsilon > 0$ and other related scalars. Set $k := 0$.
  - **Step 1 (Optimality check):** If $\left\|f(x^k)\right\| < \varepsilon$**.** Otherwise go to Step 2.
  - **Step 2 (Descent direction):** Find a suitable descent direction d^k.
  - **Step 3 (Step length):** Select a suitable step length $\alpha_k > 0$ such that
    $$f(x^k + \alpha_k d^k) < f(x^k)$$
  - **Step 4 (Update):** Update $x^{k+1} = x^k + \alpha_k d^k$. Set $k := k + 1$ and go to Step 1.

Selection of step length

- Exact line search: Suppose d^k be a descent direction of f at x^k. Step length α_k is as

$$\alpha_k = argmin_{\alpha>0}\, f(x^k + \alpha d^k)$$

- Suppose $f(x) = x_1^2 + x_1^2$. $d^k = \begin{pmatrix} -4 \\ 1 \end{pmatrix}$ is a descent direction of f at $x^k = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. Using exact line search

$$\alpha_k = argmin_{\alpha>0}(1-4\alpha)^2 + (1+\alpha)^2$$

Suppose $f(\alpha) = (1-4\alpha)^2 + (1+\alpha)^2$. For min of $f(\alpha)$, $f'(\alpha) = 0$.

This implies $-8(1-4\alpha) + 2(1+\alpha) = 0$, i.e $\alpha = \frac{6}{34} = 0.1765$

Clearly $f\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix} + 0.1765\begin{pmatrix} -4 \\ 1 \end{pmatrix}\right) = 1.4706 < f\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}\right)$.

- Inexact line search:
  - In exact line search, we have to solve an optimization problem at every iteration. This is computationally expensive.
  - Inexact line search techniques are developed to overcome these limitations.
  - We have to select step length to ensure:
    - Sufficient decrease in objective function
    - Step length not too small
  - Following **Armijo condition** ensures sufficient decrease in objective function

$$f(x + \alpha d) \leq f(x) + \alpha \beta_1 \nabla f(x)^T d, \; 0 < \beta_1 < 1 \quad (1)$$

  - **Wolfe condition** ensure that step length is not too small

$$\nabla f(x + \alpha d)^T d \geq \beta_2 \nabla f(x)^T d, \qquad \beta_1 < \beta_2 < 1 \quad (2)$$

  - **Backtracking line search:**
    - Set $\alpha := 1$
    - While either (1) or (2) does hold

      update $\alpha := \alpha r$ for some $r \in (0,1)$.

- Suppose $f(x) = 4x_1^2 + x_2^2$. At $x^0 = (1,1)^T$, $d = (-1,2)^T$ is a descent direction of f since $\nabla f(x^0)^T d = (8,2)\begin{pmatrix} -1 \\ 2 \end{pmatrix} = -4 < 0$.
- Set $\beta_1 = 10^{-4}, \beta_2 = 0.9$ and $r = 0.5$.
- For $\alpha = 1$,

$$f(x^0 + \alpha d) - \left\{f(x^0) + \alpha\, \beta_1 \nabla f(x^0)^T d\right\} = 4.0004 > 0$$

Hence Armijo condition does not hold for $\alpha = 1$. So update $\alpha = 1 * 0.5 = 0.5$

- For $\alpha = 0.5$

$$f(x^0 + \alpha d) - \left\{f(x^0) + \alpha\, \beta_1 \nabla f(x^0)^T d\right\} = 0.0002$$

Hence Armijo condition does not hold for $\alpha = 0.5$. So update $\alpha = 0.5 * 0.5 = 0.25$

- For $\alpha = 0.25$

$$f(x^0 + \alpha d) - \left\{f(x^0) + \alpha\, \beta_1 \nabla f(x^0)^T d\right\} = -0.4999$$

Hence Armijo condition holds for $\alpha = 0.25$.

- Next verify Wolfe condition for $\alpha = 0.25$.

$$\nabla f(x^0 + 0.25d)^T d - \beta_2 \nabla f(x^0)^T d = 3.6 > 0$$

- Hence Armijo-Wolfe conditions are satisfied for $\alpha = 0.25$.
- Select $\alpha = 0.25$ and proceed further.