Gradient Descent Methods

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 $f: R^n \to R$ 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 $\mathrm{x} = (1,1)^T$. $d = (-4,1)^T$ 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 $\mathrm{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

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

- 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 $k \coloneqq 0$.
  - **Step 1 (Optimality check):** If ll gradient (f(xk)) ll $\left\|f(x^k)\right\| < \varepsilon$**.** 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 $k \coloneqq k + 1$ 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
$$f(x+\alpha d) \le f(x) + \alpha\beta_1 \nabla f(x)^T d,\ 0 < \beta_1 < 1 \quad (1)$$
  - **Wolfe condition** ensure that step length is not too small
$$\nabla f(x+\alpha d)^T d \ge \beta_2\, \nabla f(x)^T d, \qquad \beta_1 < \beta_2 < 1 \quad (2)$$
  - **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.

- Algorithm:
  - Step 0 (Initialization/Inputs): Choose f, x^0 (initialization), $\varepsilon(> 0)$, β_1, β_2 (0<β_1<$\beta_2 < 1$) and $r \in (0,1)$. (β_1, β_2, and r are required only if we use inexact line search technique). Set $k \coloneqq 0$.
  - Step 1 (Optimality/Stopping condition check): Compute $\nabla f(x^0)$. If $\|\nabla f(x^0)\| < \varepsilon$ then stop. Otherwise go to Step 2.
  - Step 2 (Compute descent direction) : Use different technique to compute suitable descent direction d^k.
  - Step 3 (Step length selection): Use exact/inexact (Armijo-Wolfe backtracking) line search technique to find suitable step length ($\alpha_k > 0$).
  - Step 4 (Update): Update x^{k+1} by $x^{k+1=x^k+\alpha_k}d^k$. Set $k \coloneqq k + 1$ and go to Step 1.
- Output: An approximate stationary point.

- **Steepest Descent Method:**
  - In this method we use $\boldsymbol{d^k} = -\boldsymbol{\nabla f}(\boldsymbol{x^k})$.
  - Clearly d^k is a descent direction since

$$d^{k^T} \nabla f(x^k) = -\left\|\nabla f(x^k)\right\|^2 < 0$$

- **Example:**
  - Suppose $f(x) = 4x_1^2 + x_2^2 - 2x_1x_2$.
  - Then $\nabla f(x) = \begin{pmatrix} 8x_1 - 2x_2 \\ -2x_1 + 2x_2 \end{pmatrix}$. Clearly $\nabla f(x) = 0$ implies $x_1 = 0 = x_2$.
  - Now $\nabla^2 f(x) = \begin{pmatrix} 8 & -2 \\ -2 & 2 \end{pmatrix}$.
  - At $x = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\nabla^2 f(x)$ is positive definite (verify).
  - Hence $x = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ is a local minima of f

Solve $\boldsymbol{min_{x \in R^2} f(x) = 4x_1^2 + x_2^2 - 2x_1x_2}$ ***using steepest descent method.***

- Choose initial approximation $x^0 = \binom{2}{2}$, $\beta_1 = 10^{-4}$, $\beta_2 = 0.9$, $r = 0.5$, and $\varepsilon = 10^{-3}$.
- Now $\nabla f(x^0) = \binom{12}{0}$, clearly $\|\nabla f(x^0)\| = 12 > \varepsilon$. So we calculate d^0 and proceed.
- $d^0 =- \nabla f(x^0) = \binom{-12}{0}$.
- Next select step length using exact line search technique:
  - $\alpha_0 = argmin_{\alpha>0}\, \phi(\alpha) = 4(2 - 12\alpha)^2 + 4 - 2(2 - 12\alpha)2$
  - $\phi'(\alpha) =- 96(2 - 12\alpha) + 48$
  - If $\alpha_0$is a minimizer then $\phi'(\alpha_0) = 0$. This implies$\alpha_0 = 0.125$

- So $x^1 = x^0 + \alpha_0 d^0 = \begin{pmatrix} 2 \\ 2 \end{pmatrix} + 0.125 \begin{pmatrix} -12 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 2 \end{pmatrix}$
- Clearly $f(x^1) = 3 < f(x^0) = 12$
- Now $\nabla f(x^0) = \begin{pmatrix} 0 \\ 3 \end{pmatrix}$, clearly $||\nabla f(x^0)|| = 3 > \varepsilon$. So we calculate d^1 and proceed.
- $d^1 =- \nabla f(x^1) = \begin{pmatrix} 0 \\ -3 \end{pmatrix}$.
- Next select step length using exact line search technique:
    - $\alpha_1 = argmin_{\alpha>0}\, \phi(\alpha) = 1 + (2 - 3\alpha)^2 - 2 * 0.5(2 - 3\alpha)$
    - $\phi'(\alpha) =- 6(2 - 3\alpha) + 3$
    - If α_1 is a minimizer then $\phi'(\alpha_1) = 0$. This implies $\alpha_1 = 0.5$
- Then $x^2 = x^1 + \alpha_1 d^1 = \begin{pmatrix} 0.5 \\ 2 \end{pmatrix} + 0.5 \begin{pmatrix} 0 \\ -3 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$
- Clearly $f(x^2) = 0.75 < f(x^1) =0$

- Now $\nabla f(x^2) = \begin{pmatrix} 3 \\ 0 \end{pmatrix}$, $d^2 = \begin{pmatrix} -3 \\ 0 \end{pmatrix}$.
- Observe that
  - $(d^0)^T d^1 = (-12,0)\begin{pmatrix} 0 \\ -3 \end{pmatrix} = 0$
  - $(d^1)^T d^2 = (0,-3)\begin{pmatrix} -3 \\ 0 \end{pmatrix} = 0$
- **In steepest descent method two consecutive descent directions are perpendicular to each other if step lengths are selected using exact line search technique.**

Solve $\boldsymbol{min_{x\in R^2} f(x) = (x_1 - 2)^2 + (2x_2 - x_1)^2}$ ***using steepest descent method (using Armijo-Wolfe inexact line search technique)***

- Choose initial approximation $x^0 = \begin{pmatrix} 3 \\ 3 \end{pmatrix}, \beta_1 = 10^{-4}, \beta_2 = 0.9, r = 0.5$, and $\varepsilon = 10^{-3}$.
- We, $f(x^0) = 10$.
- Now $\nabla f(x^0) = \begin{pmatrix} -4 \\ 12 \end{pmatrix}$, clearly $\|\nabla f(x^0)\| = 12.65 > \varepsilon$.

  So we calculate d^0 and proceed.
- $d^0 =- \nabla f(x^0) = \begin{pmatrix} 4 \\ -12 \end{pmatrix}$.

- Select step length using inexact line search technique:
  - For $\alpha = 1$, $f(x^0 + \alpha d^0) = 650 > f(x^0) + \alpha\beta_1 \nabla f(x^0)^T d^0$.
  - Update $\alpha = \alpha$r=0.5
  - For $\alpha = 0.5$, $f(x^0 + \alpha d^0) = 130 > f(x^0) + \alpha\beta_1 \nabla f(x^0)^T d^0$.
  - Update $\alpha = \alpha$r=0.25
  - For $\alpha = 0.25$, $f(x^0 + \alpha d^0) = 20 > f(x^0) + \alpha\beta_1 \nabla f(x^0)^T d^0$.
  - Update $\alpha = \alpha$r=0.125
  - For $\alpha = 0.125$, $f(x^0 + \alpha d^0) = 2.5 < f(x^0) + \alpha\beta_1 \nabla f(x^0)^T d^0$.
  - Also $\nabla f(x^0 + \alpha d^0)^T d^0 > \beta_2 \nabla f(x^0)^T d^0$ holds for $\alpha = 0.125$
  - So we choose $\alpha_0 = 0.125$

- Then $x^1 = x^0 + \alpha_0 d^0 = \begin{pmatrix} 3 \\ 3 \end{pmatrix} + 1/8 \begin{pmatrix} 4 \\ -12 \end{pmatrix} = \begin{pmatrix} 3.5 \\ 1.5 \end{pmatrix}$
- Clearly $f(x^1) = 2.5 < f(x^0)$ =10
- Final solution using stopping criteria $\left\|\nabla f(x^k)\right\| < \varepsilon$ is

$$x^{26} = \begin{pmatrix} 2.00022 \\ 1.00017 \end{pmatrix} \cong \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

- Major limitation of steepest descent method is: it converges linearly that is rate of convergence is 1.

Mirror Descent Method

- Consider the problem

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

- Descent direction at x^k is obtained by solving the unconstrained problem

$$\min_{d} \; \Delta f(x^k)^T d + \frac{1}{2} d^T d$$

- The second term is same ae Euclidian distace.
- We observed that steepest descent method converges slowly.
- To increase the convergence rate, we replace Euclidian distance by some non-Euclidian distance called Bregman distance.

- Suppose h: $X \subseteq R^n \to R \cup \{\infty\}$ be a strcitly convex function and X is a closed subset. Bregman distance induced by h is defined by
$$d_h(x,y) = h(x) - h(y) - \nabla h(y)^T(x-y) \qquad (b1)$$
- By mean value theorem $d_h(x,y) = \frac{1}{2}(x-y)^T \nabla^2 h(\zeta)(x-y)$ for some ζ in x and y.
- Since h is strongly convex function
  - $d_h(x,y) \geq 0$ for all x and y.
  - $d_h(x,y) = 0$ iff $x = y$.
- But $d_h(x,y) \neq d_h(y,x)$ *(Verify with counter example)*
- Also triangular inequality does not hold for Bregman distance.
- **Three point identity:** For x, y, z
$$d_h(x,y) = d_h(x,z) + d_h(z,y) - (x-z)^T(\Delta h(y) - \nabla h(z))$$

from RHS, show LHS using (b1).

Examples

- Define $h(x) = x^T x$, then

$$d_h(x, y) = (x - y)^T (x - y) = ||x - y||^2$$

this is same as Euclidian distance.

- Define $h(x) = x^T Q x$, where Q is symmetric positive definite. Then

$$d_h(x, y) = (x - y)^T Q (x - y)$$

This is known as elliptic norm or squared Mahalanobis distance

- Define $h(x) = \sum_{i=1}^{n} x_i log(\, x_i)$, with $0log(0) = 0$. Then

$$d_h(x, y) = \sum_{i=1}^{n} \left(x_i log\left(\frac{x_i}{y_i}\right) \; - x_i + y_i \right)$$

This is called Kullback-Leibler divergence (KL-divergence).

- Descent direction at x^k is obtained by solving the unconstrained problem

$$\min_{d} \Delta f(x^k)^T d + \frac{1}{2} d_h(x^k + d, d)$$

Newton Method

- Consider the unconstrained optimization problem

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

where f is a strictly convex function.

- At any iterating point x^k, consider the quadratic approximation of f as

$$q(x; x^k) = f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{1}{2}(x - x^k)^T \nabla^2 f(x^k)(x - x^k)$$

- Choose next iterating point is the minimizer of $q(x; x^k)$, i.e.

$$x^{k+1} = argmin_{x\in R^n} q(x; x^k)$$

- Then

$$\nabla q(x^{k+1}; x^k) = 0$$

- This implies

$$\nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k) = 0$$

i.e.
$$x^{k+1} = x^k - \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)$$

- Comparing above update formula with $x^{k+1} = x^k + \alpha_k d^k$ we have

$\alpha_k = 1$ and $d^k = -\left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)$

- d^k is a descent direction since

$${d^k}^T \nabla f(x^k) = -\nabla f(x^k)^T \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k) < 0$$

- The last inequality holds since $\nabla^2 f(x^k)$ is positive definite (as f is strictly convex) implies $\left(\nabla^2 f(x^k)\right)^{-1}$ is positive definite.

- Example: Let $f(x) = 3x_1^2 + x_2^2 - 3x_1x_2 + 3x_1 - x_2$
- Choose $x^0 = \begin{pmatrix} 3 \\ 3 \end{pmatrix}$
- $\nabla f(x) = \begin{pmatrix} 6x_1 - 3x_2 + 3 \\ -3x_1 + 2x_2 - 1 \end{pmatrix}$, $\nabla^2 f(x) = \begin{pmatrix} 6 & -3 \\ -3 & 2 \end{pmatrix}$
- $\nabla f(x^0) = \begin{pmatrix} 12 \\ -4 \end{pmatrix}$
- $x^1 = \begin{pmatrix} 3 \\ 3 \end{pmatrix} - \begin{pmatrix} 6 & -3 \\ -3 & 2 \end{pmatrix}^{-1} \begin{pmatrix} 12 \\ -4 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \end{pmatrix} - \begin{pmatrix} \frac{2}{3} & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 12 \\ -4 \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$
- $\nabla f(x^1) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$
- Hence $x^1 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$ is the solution
- If f is a quadratic function then Newton method converges to solution in first iteration.

- Suppose $f(x) = (x_1 - 2)^4 + (x_1 - 2x_2)^2$.
- Then $\nabla f(x) = \begin{pmatrix} 4(x_1 - 2)^3 + 2(x_1 - 2x_2) \\ -4(x_1 - 2x_2) \end{pmatrix}$

and $\nabla^2 f(x) = \begin{pmatrix} 12(x_1 - 2)^2 + 2 & -4 \\ -4 & 8 \end{pmatrix}$

- Choose $x^0 = \begin{pmatrix} 0 \\ 3 \end{pmatrix}$. Then $f(x^0)$ =25

k	x^k	$f(x^k)$	$\nabla f(x^k)$	$\nabla^2 f(x^k)$	$\left(\nabla^2 f(x^k)\right)^{-1}$	d^k	$x^{k+1} = x^k + d^k$
0	$\begin{pmatrix}0\\3\end{pmatrix}$	25	$\begin{pmatrix}-44\\6\end{pmatrix}$	$\begin{pmatrix}56 & -4\\-4 & 2\end{pmatrix}$	$\frac{1}{96}\begin{pmatrix}2 & 4\\4 & 56\end{pmatrix}$	$\begin{pmatrix}2/3\\-5/3\end{pmatrix}$	$\begin{pmatrix}0.6667\\1.3334\end{pmatrix}$
1	$\begin{pmatrix}0.6667\\1.3334\end{pmatrix}$	3.1605	$\begin{pmatrix}-9.4806\\0\end{pmatrix}$	$\begin{pmatrix}29.332 & -4\\-4 & 2\end{pmatrix}$	$\frac{1}{42.664}\begin{pmatrix}2 & 4\\4 & 29.332\end{pmatrix}$	$\begin{pmatrix}0.4444\\0.8888\end{pmatrix}$	$\begin{pmatrix}1.1111\\2.2222\end{pmatrix}$
2	$\begin{pmatrix}1.1111\\2.2222\end{pmatrix}$	0.62433	$\begin{pmatrix}-2.8093\\0\end{pmatrix}$	$\begin{pmatrix}17.4815 & -4\\-4 & 2\end{pmatrix}$	$\frac{1}{18.963}\begin{pmatrix}2 & 4\\4 & 17.482\end{pmatrix}$	$\begin{pmatrix}0.2963\\0.5926\end{pmatrix}$	$\begin{pmatrix}1.4074\\2.8148\end{pmatrix}$
3	$\begin{pmatrix}1.4074\\2.8148\end{pmatrix}$	0.12332	$\begin{pmatrix}-0.8324\\0\end{pmatrix}$	$\begin{pmatrix}12.214 & -4\\-4 & 2\end{pmatrix}$	$\frac{1}{8.428}\begin{pmatrix}2 & 4\\4 & 12.214\end{pmatrix}$	$\begin{pmatrix}0.1975\\0.395\end{pmatrix}$	$\begin{pmatrix}1.6049\\3.2098\end{pmatrix}$
4	$\begin{pmatrix}1.6049\\3.2098\end{pmatrix}$	0.0244	$\begin{pmatrix}-0.2467\\0\end{pmatrix}$	$\begin{pmatrix}9.8732 & -4\\-4 & 2\end{pmatrix}$	$\frac{1}{3.7464}\begin{pmatrix}2 & 4\\4 & 9.8732\end{pmatrix}$	$\begin{pmatrix}0.1317\\0.2634\end{pmatrix}$	$\begin{pmatrix}1.7366\\3.4732\end{pmatrix}$
5	$\begin{pmatrix}1.7366\\3.4732\end{pmatrix}$	0.00481	$\begin{pmatrix}-0.0731\\0\end{pmatrix}$	$\begin{pmatrix}8.8325 & -4\\-4 & 2\end{pmatrix}$	$\frac{1}{1.665}\begin{pmatrix}2 & 4\\4 & 8.8325\end{pmatrix}$	$\begin{pmatrix}0.0877\\0.1755\end{pmatrix}$	$\begin{pmatrix}1.8243\\3.6487\end{pmatrix}$

Observe that $\{x^k\}$ converging to $x^* = (2,4)^T$