Chapter one

Unconstrained optimization

* **Global Minimizer** : A point x^* is a global minimizer if $f(x^*) \leq f(x)$ $\forall x$, where x ranges over all of $\mathbb{R}^n$.

* **Local minimizer or weak local minimizer** : A point x^* is a local minimizer if there exists a neighborhood $\mathcal{N}$ of x^* such that $f(x^*) \leq f(x)$ $\forall x \in \mathcal{N}$.

(*) **Strict Local Minimizer** : A point x^* is a strict local minimizer (also called strong local minimizer) if there is a neighborhood $\mathcal{N}$ of x^* such that $f(x^*) < f(x)$ $\forall x \in \mathcal{N}$ with $x \neq x^*$.

Example : $f(x) = 2$, every point x is a weak minimizer while the function $f(x) = (x-2)^4$ has a strict local minimizer at $x = 2$.

(*) Isolated minimizer : A point x^* is an isolated local minimizer if there is a neighborhood $\mathcal{N}$ of x^* such that x^* is the only local minimizer in $\mathcal{N}$.

Fact : Strict local minimizers are not always isolated but isolated local minimizers are strict.

Fact : When the function is smooth, there are efficient way to identify local minima. In particulare, if f is twice continuously differentiable, we may be able to say that x^* is a local minimizer (and possibly a strict local minimizer) by examining just $\nabla f(x^*)$ and the Hessian $\nabla^2 f(x^*)$.

* The mathematical tool used to study minimizers of smooth function is Taylor's Theorem.

(*) Derivation of Taylor's Theorem

$f: \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable and $p \in \mathbb{R}^n$

then

$$f(x+p) = f(x) + \nabla f(x+tp)^T p, \text{ for some } t \in (0,1)$$

Let $x(t) = x + tp$ with $t \in (0,1)$

$$\phi(t) = f(x+tp)$$

$$\phi'(t) = \nabla f(x+tp) \cdot p \quad \text{(Chain Rule)}$$

$$\phi(1) = f(x+p)$$

Expanding $f(x+p)$ about 0

$$\therefore f(x+p) = f(x) + \underbrace{\nabla f(x+tp)^T \cdot p}_{\phi'(t)}$$

and order 2

$$f(x+p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \cdot \nabla^2 f(x+tp) \cdot p$$

for some $t \in (0,1)$

Theorem 2.1 (Taylor's Theorem).

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and that $p \in \mathbb{R}^n$. Then we have that

$$f(x+p) = f(x) + \nabla f(x+tp)^T p, \tag{2.4}$$

for some $t \in (0,1)$. Moreover, if f is twice continuously differentiable, we have that

$$\nabla f(x+p) = \nabla f(x) + \int_0^1 \nabla^2 f(x+tp)p\,dt, \tag{2.5}$$

and that

$$f(x+p) = f(x) + \nabla f(x)^T p + \tfrac{1}{2} p^T \nabla^2 f(x+tp)p, \tag{2.6}$$

for some $t \in (0,1)$.

Necessary conditions for optimality are derived by assuming that x^* is a local minimizer and then proving facts about $\nabla f(x^*)$ and $\nabla^2 f(x^*)$.

(*) First-order necessary condition :

If x^* is a local minimizer and f is continuously differentiable in an open neighborhood of x^*, then $\nabla f(x^*) = 0$.

Proof : Suppose for contradiction that

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

Define the vector $p = -\nabla f(x^*)$ and note that $p^T \nabla f(x^*) = -\| \nabla f(x^*) \|^2 < 0$.

Because ∇f is continuous near x^*, there exists a scalar $T > 0$ such that

$$p^T \nabla f(x^* + tp) < 0 \quad \forall\, t \in [0, T]$$

For any $\bar{t} \in (0, T]$, we have Taylor's Theorem

$$f(x^* + \bar{t}p) = f(x^*) + \bar{t} p^T \nabla f(x^* + tp)$$

$$\text{for } t \in (0, \bar{t})$$

Therefore, $f(x^* + \bar{t}p) < f(x^*)$ $\forall\, \bar{t} \in (0, T]$. We have found a direction leading away from x^* along which f decreases, so, x^* is not a local minimizer and we have a contradiction.

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

For the next result we recall that a matrix B is positive definite if $p^{\prime}\, Bp > 0$ for all $p \neq 0$, and positive semidefinite if $p^T Bp \geq 0$ for all p (see the Appendix).

Theorem 2.3 (Second-Order Necessary Conditions).

If x^* is a local minimizer of f and $\nabla^2 f$ exists and is continuous in an open neighborhood of x^*, then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive semidefinite.

PROOF. We know from Theorem 2.2 that $\nabla f(x^*) = 0$. For contradiction, assume that $\nabla^2 f(x^*)$ is not positive semidefinite. Then we can choose a vector p such that $p^T \nabla^2 f(x^*) p < 0$, and because $\nabla^2 f$ is continuous near x^*, there is a scalar $T > 0$ such that $p^T \nabla^2 f(x^* + tp) p < 0$ for all $t \in [0, T]$.

By doing a Taylor series expansion around x^*, we have for all $\bar{t} \in (0, T]$ and some $t \in (0, \bar{t})$ that

$$f(x^* + \bar{t}p) = f(x^*) + \bar{t}p^T \nabla f(x^*) + \tfrac{1}{2}\bar{t}^2 p^T \nabla^2 f(x^* + tp) p < f(x^*).$$

As in Theorem 2.2, we have found a direction from x^* along which f is decreasing, and so again, x^* is not a local minimizer. □

We now describe *sufficient conditions*, which are conditions on the derivatives of f at the point z^* that guarantee that x^* is a local minimizer.

Theorem 2.4 (Second-Order Sufficient Conditions).

Suppose that $\nabla^2 f$ is continuous in an open neighborhood of x^* and that $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite. Then x^* is a strict local minimizer of f.

PROOF. Because the Hessian is continuous and positive definite at x^*, we can choose a radius $r > 0$ so that $\nabla^2 f(x)$ remains positive definite for all x in the open ball $\mathcal{D} = \{z \mid \|z - x^*\| < r\}$. Taking any nonzero vector p with $\|p\| < r$, we have $x^* + p \in \mathcal{D}$ and so

$$\begin{aligned} f(x^* + p) &= f(x^*) + p^T \nabla f(x^*) + \tfrac{1}{2} p^T \nabla^2 f(z) p \\ &= f(x^*) + \tfrac{1}{2} p^T \nabla^2 f(z) p, \end{aligned}$$

where $z = x^* + tp$ for some $t \in (0, 1)$. Since $z \in \mathcal{D}$, we have $p^T \nabla^2 f(z) p > 0$, and therefore $f(x^* + p) > f(x^*)$, giving the result. □

(*) Give an example that shows that a second order sufficient condition is not necessary.

Soln:- The second-order sufficient condition states that x^* is a strict local minimizer.

Now consider the function

$$f(x) = x^4$$ for which the point $x^* = 0$ is a strict local minimizer but at x^* the Hessian matrix $\nabla^2 f$ vanishes and therefore it is not positive definite. Hence it fails to satisfy the second-order necessary condition.

(*) convex set

Let S be the set. For any two points $x \in S$ and $y \in S$, the set S is convex if

$$\alpha x + (1-\alpha) y \in S, \quad \forall \alpha \in [0, 1].$$

(*) convex function

The function f is a convex function if its domain S is a convex set and if for any two points $x, y \in S$, the following property holds

$$f(\alpha x + (1-\alpha) y) \leq \alpha f(x) + (1-\alpha) f(y)$$

$$\forall \alpha \in [0, 1].$$

(*) Strictly convex :

A function is strictly convex if the domain of f ie. S is a convex set and if for any two points $x, y \in S$ then $f(\alpha x + (1-\alpha)y) < \alpha f(x) + (1-\alpha) f(y) \quad \forall \alpha \in [0,1]$.

(***) Theorem : If $f: \mathbb{R}^n \to \mathbb{R}^n$ is convex, then any local minimizer x^* is a global minimizer of f.

proof : Suppose that x^* is a local but not a global minimizer. Then we can find a point $z \in \mathbb{R}^n$ with $f(z) < f(x^*)$. consider the line segment that joins x^* and z, that is

$$x = \lambda z + (1-\lambda) x^* \quad \text{for some } \lambda \in [0,1]. \quad \cdots\cdots (1)$$

By the convexity property for f, we have

$$f(x) \leq \lambda f(z) + (1-\lambda) f(x^*) < f(x^*) \quad \cdots\cdots (2)$$

Any neighborhood N of x^* containg a piece of the line segment (1), so there will always be points $x \in N$ at which (2) is satisfied. Hence x^* is not a local minimizer of f.

Therefore, x^* is a global minimizer of f.

(*) Theorem : When f is convex and f is differentiable, then any stationary point x^* is a global minimizer of f.

proof : Suppose that x^* is not a global minimizer and let z be any point in $\mathbb{R}^n$. Then from convexity, we have

$$\nabla f(x^*)^T (z - x^*) = \frac{d}{d\lambda} f(x^* + \lambda(z - x^*))\Big|_{\lambda=0}$$

$$= \lim_{\lambda \to 0} \frac{f(x^* + \lambda(z - x^*)) - f(x^*)}{\lambda}$$

$$\leq \lim_{\lambda \to 0} \frac{f(x^*) + \lambda f(z) - \lambda f(x^*) - f(x^*)}{\lambda}$$

$$= f(z) - f(x^*) < 0$$

Therefore, $\nabla f(x^*) \neq 0$, and so x^* is not a sationary point.

#V.V.I : If f is convex, then the first-order necessary condition in this case is a sufficient condition.

Newton's Method

(*) In Line search direction perhaps the most important one of all search directions is the Newton search direction. This direction is derived from the second-order Taylor series approximation to $f(x_k+p)$, which is

$$f(x_k+p) \approx f(x_k) + p^T \nabla f_k + \frac{1}{2} p^T \nabla^2 f_k p := m_k(p) \quad \text{---------- (1)}$$

Assuming for the moment that $\nabla^2 f_k$ is positive definite, we obtain the Newton direction by finding the vector p that minimizes $m_k(p)$. By simply setting the derivative of $m_k(p)$ to zero, we get the explicit formula:

$$p_k^N = -(\nabla^2 f_k)^{-1} \nabla f_k .$$

Note: If f is twice continuously differentiable, then from Taylor's Theorem,

$$f(x+p) = f(x) + \nabla f^T(x) p + \frac{1}{2} p^T \nabla^2 f(x+tp) p, \text{ for some } t \in (0,1).$$

$$\text{---------- (2)}$$

Newton direction is reliable when the difference between the true function $f(x_k+p)$ and its quadratic model $m_k(p)$ is not too large. By comparing (1) with (2), we see that the only difference between these functions is that matrix $\nabla^2 f(x_k+tp)$ in the third term of the expansion has been replaced by $\nabla^2 f_k$ in (1). If $\nabla^2 f$ is sufficiently smooth, this difference introduces a perturbation of only $O(\|p\|^3)$ into the expansion, so that when $\|p\|$ is small, the approximation $f(x_k+p) \approx m_k(p)$ is quite accurate.

Local rate of convergence properties of Newton's method.

For all x in the vicinity of a solution point x^* such that $\nabla^2 f(x^*)$ is positive definite, the Hessian $\nabla^2 f(x)$ will also be positive definite. Newton's method will be defined in this region and will converge quadratically, provided that the step lengths α_k are eventually always 1.

(***) **Theorem:** Suppose that f is twice differentiable and that the Hessian $\nabla^2 f(x)$ is Lipschitz continuous in a neighborhood of a solution x^* at which the sufficient conditions are satisfied. Consider the iteration $x_{k+1} = x_k + p_k$, where $p_k = -\nabla^2 f_k^{-1} \nabla f_k$, then

(i) if the starting point x_0 is sufficiently close to x^*, the sequence of iterates converges to x^*.

(ii) the rate of convergence of $\{x_k\}$ is quadratic, and

(iii) the sequence of gradient norms $\{\|\nabla f_k\|\}$ converges quadratically to zero.

PROOF. From the definition of the Newton step and the optimality condition $\nabla f_* = 0$ we have that

$$\begin{aligned} x_k + p_k^{\mathrm{N}} - x^* &= x_k - x^* - \nabla^2 f_k^{-1} \nabla f_k \\ &= \nabla^2 f_k^{-1} \left[\nabla^2 f_k (x_k - x^*) - (\nabla f_k - \nabla f_*) \right]. \end{aligned} \tag{3.31}$$

Since Taylor's theorem (Theorem 2.1) tells us that

$$\nabla f_k - \nabla f_* = \int_0^1 \nabla^2 f(x_k + t(x^* - x_k))(x_k - x^*)\, dt,$$

we have

$$\begin{aligned} &\left\| \nabla^2 f(x_k)(x_k - x^*) - (\nabla f_k - \nabla f(x^*)) \right\| \\ &\quad = \left\| \int_0^1 \left[\nabla^2 f(x_k) - \nabla^2 f(x_k + t(x^* - x_k)) \right] (x_k - x^*)\, dt \right\| \\ &\quad \le \int_0^1 \left\| \nabla^2 f(x_k) - \nabla^2 f(x_k + t(x^* - x_k)) \right\| \|x_k - x^*\|\, dt \\ &\quad \le \|x_k - x^*\|^2 \int_0^1 Lt\, dt = \tfrac{1}{2} L \|x_k - x^*\|^2, \end{aligned} \tag{3.32}$$

where L is the Lipschitz constant for $\nabla^2 f(x)$ for x near x^*. Since $\nabla^2 f(x^*)$ is nonsingular, there is a radius $r > 0$ such that $\|\nabla^2 f_k^{-1}\| \le 2\|\nabla^2 f(x^*)^{-1}\|$ for all x_k with $\|x_k - x^*\| \le r$. By substituting in (3.31) and (3.32), we obtain

$$\|x_k + p_k^{\mathrm{N}} - x^*\| \le L \|\nabla^2 f(x^*)^{-1}\| \|x_k - x^*\|^2 = \tilde{L} \|x_k - x^*\|^2, \tag{3.33}$$

where $\tilde{L} = L\|\nabla^2 f(x^*)^{-1}\|$. Choosing x_0 so that $\|x_0 - x^*\| \le \min(r, 1/(2\tilde{L}))$, we can use this inequality inductively to deduce that the sequence converges to x^*, and the rate of convergence is quadratic.

By using the relations $x_{k+1} - x_k = p_k^{\mathrm{N}}$ and $\nabla f_k + \nabla^2 f_k p_k^{\mathrm{N}} = 0$, we obtain that

$$\begin{aligned} \|\nabla f(x_{k+1})\| &= \|\nabla f(x_{k+1}) - \nabla f_k - \nabla^2 f(x_k) p_k^{\mathrm{N}}\| \\ &= \left\| \int_0^1 \nabla^2 f(x_k + t p_k^{\mathrm{N}})(x_{k+1} - x_k)\, dt - \nabla^2 f(x_k) p_k^{\mathrm{N}} \right\| \\ &\le \int_0^1 \left\| \nabla^2 f(x_k + t p_k^{\mathrm{N}}) - \nabla^2 f(x_k) \right\| \|p_k^{\mathrm{N}}\|\, dt \\ &\le \tfrac{1}{2} L \|p_k^{\mathrm{N}}\|^2 \\ &\le \tfrac{1}{2} L \|\nabla^2 f(x_k)^{-1}\|^2 \|\nabla f_k\|^2 \\ &\le 2L \|\nabla^2 f(x^*)^{-1}\|^2 \|\nabla f_k\|^2, \end{aligned}$$

proving that the gradient norms converge to zero quadratically. □

Spring 2018

(a) Explain how Newton's root finding method for a smooth function $g : \mathbb{R}^n \to \mathbb{R}^n$ can be used to optimize a smooth objective function $f : \mathbb{R}^n \to \mathbb{R}$. State an appropriate algorithm.

Soln:

Newton's root finding method is an iterative numerical method used to find the roots of a differential equation. The basic idea is to start with an initial guess and then improve it by iteratively computing the next guess as the intersection of the tangent line at the current guess with the x-axis.

The same idea can be applied to find the minimum of a smooth objective function $f: \mathbb{R}^n \to \mathbb{R}^n$ by using the fact that the minimum occurs at a point where the gradient is zero. In other words, we want to find a root of the gradient function $g(x) = \nabla f(x)$, where $\nabla f(x)$ is the gradient of f at x.

Newton's Algorithm:

We start with an initial guess x_0 and iteratively update the guess as follows:

1. Compute the gradient and Hessian of f at x_k:
$$g_k = \nabla f(x_k), \quad H_k = \nabla^2 f(x_k)$$
2. Solve the linear system $H_k p_k = -g_k$ for the search direction p_k
3. Update the guess $x_{k+1} = x_k + p_k$
4. Repeat steps 1-3 until a stopping criterion is satisfied.

Newton's method is a second-order method, which means it converges faster than the first-order methods such as gradient descent method.

(b) Illustrate Newton's method for the objective function $f(x) = x^2 + 5y^2$ with the starting point $x_o = (1,1)^T$.

Soln : $f(x) = x^2 + 5y^2$, $x_0 = (1,1)^T$

we compute the gradient and Hessian of f.

$$\nabla f(x) = [2x \quad 10y]^T, \qquad H = \nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix}$$

$$\therefore H = \begin{bmatrix} 2 & 0 \\ 0 & 10 \end{bmatrix}$$

Now we apply the Newton's algorithm for optimization to update the initial guess iteratively

(i) compute the gradient and Hessian at $x_0 = (1,1)^T$

$$\therefore g_0 = \nabla f(x_0) = [2, 10]^T$$

$$H_0 = \begin{bmatrix} 2 & 0 \\ 0 & 10 \end{bmatrix}$$

(ii) Solve for p_0 from $H_0 p_0 = -g_0 \Rightarrow p_0 = -H_0^{-1} g_0$

$$p_0 = -\frac{1}{20}\begin{bmatrix} 10 & -0 \\ -0 & 2 \end{bmatrix}\begin{bmatrix} 2 \\ 10 \end{bmatrix} = -\frac{1}{20}\begin{bmatrix} 20 \\ 20 \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$$

$$\therefore p_0 = (-1,-1)^T$$

(iii) update the guess $x_1 = x_0 + p_0 = (1,1)^T + (-1,-1)^T = (0,0)^T$

(iv) Repeat steps (i)-(iii) until a stopping criterion is satisfied.

We have reached a critical point $x_1 = (0,0)^T$, which is the global minimum of the function. At this point, the gradient is zero and Hessian is positive definite, indicating that this a local min.

(c) A secant updating method is a quasi-Newton method defined by $B_{k+1}s_k = y_k$, where $s_k = x_{k+1} - x_k$, $y_k = \text{grad} f(x_{k+1}) - \text{grad} f(x_k)$ and B_{k+1} is an approximation matrix to the Hessian matrix. Show that the matrix defined by

$$B_{k+1} = B_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} \quad \text{——— (1)}$$

satisfies this secant updating condition. Also show that B_{k+1} is symmetric if B_k is.

Soln: To show the matrix B_{K+1} satisfies the secant updating condition, we need to show that $B_{K+1} s_K = y_K$.

we have

$$B_{K+1} = B_K + \frac{y_K y_K^T}{y_K^T s_K} - \frac{B_K s_K s_K^T B_K}{s_K^T B_K s_K}$$

$$\Rightarrow B_{K+1} s_K = B_K s_K + \frac{y_K y_K^T s_K}{y_K^T s_K} - \frac{B_K s_K s_K^T B_K}{s_K^T B_K s_K} s_K$$

$$= B_K s_K + y_K - \frac{B_K s_K (s_K^T B_K s_K)}{(s_K^T B_K s_K)}$$

$$= B_K s_K + y_K - B_K s_K = y_K$$

$\therefore B_{K+1} s_K = y_K$; satisfies the secant updating condition.

To show B_{K+1} is symmetric if B_K is symmetric, we can compute B_{K+1}^T as follows:

$$B_{K+1}^T = \left(B_K + \frac{y_K y_K^T}{y_K^T s_K} - \frac{B_K s_K s_K^T B_K}{s_K^T B_K s_K} \right)^T$$

$$= B_K^T + \frac{(y_K y_K^T)^T}{(y_K^T s_K)^T} - \frac{(B_K s_K s_K^T B_K)^T}{(s_K^T B_K s_K)^T}$$

Scalar

$$= B_K + \frac{y_K y_K^T}{y_K^T s_K} - \frac{B_K s_K s_K^T B_K}{s_K^T B_K s_K} = B_{K+1}$$

$\therefore B_{K+1}$ is symmetric.

(d) State the BFGS algorithm using the Hessian approximation given in the previous part. What is the initial Hessian approximation B_0.

Soln :

The BFGS algorithm using the Hessian approximation given in the previous part can be stated as follows :

1. Choose an initial point x_0 and initial approximation to the Hessian matrix B_0.
2. For $k = 0, 1, 2, \ldots\ldots$ do steps 3 – 8
3. Solve the system of equations $B_k p_k = -\nabla f(x_k)$ to obtain the search direction p_k.
4. Choose the step size α_k by a line search method such as the Armijo – Goldstein condition
5. Set $x_{k+1} = x_k + \alpha_k p_k$
6. compute the new gradient $g_{k+1} = \nabla f(x_{k+1})$
7. compute the vectors $y_k = g_{k+1} - g_k$ and $s_k = x_{k+1} - x_k$
8. update the approximation matrix B_{k+1} using the BFGS formula :

$$B_{k+1} = B_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k}$$

The initial Hessian approximation can be chosen in a variety of ways. one common choice is to set $B_0 = I$, the identity matrix.

(4) Outline Newtons algorithm for minimizing $f : \mathbb{R}^n \to \mathbb{R}$. Be sure to address the need for and implementation details for either a line search or trust region, procedures for dealing with non positive-definite matrices, operation counts, and terminal convergence rate. Be sure to define all terms you use.

Soln:

Newton's Algorithm for Line search / Trust region subproblem

1. Choose initial point x_0
2. At the k-th iteration, solve the Newton system

$H_k \cdot p_k = -\nabla f(x_k)$, where H_k is the Hessian matrix at x_k and p_k is the search direction

3. compute step length α_k using a line search or trust region method.

(a) For the line search, find α_k that satisfies the Armijo - Goldstein condition

$$f(x_k + \alpha_k p_k) \leq f(x_k) + c_1 \alpha_k \nabla f(x_k)^T p_k$$

for $c_1 \in (0, 1)$.

(b) for trust region, solve the subproblem

$$\min m(p_k) = f(x_k) + \nabla f(x_k)^T p_k + \frac{1}{2} p_k^T H_k p :$$

$$\| p_k \| \leq \Delta_k$$

Where Δ is the trust region radius.

(4) Update

$$x_{k+1} = x_k + \alpha_k p_k$$

(5) Repeat (2)-(4) until a stopping criterion is satisfied.

In addition to the above steps, the algorithm must also account for the following implementation details:

If the Hessian is not positive-definite, then this can cause the search direction to be undefined or lead to a step that increases the objective function. To address this, a common approach is to modify the Hessian matrix by adding a multiple of identity matrix to ensure the positive-definiteness. This is known as regularization and a common form of regularization is given by

$$H_k = H_k + \mu_k I \quad \text{where } \mu_k \text{ is a positive constant}$$

Operation counts : Newton's algorithm requires the computation of Hessian matrix and its inverse, which can be computationally expensive. To reduce the computational cost, one can approximate the estimate of Hessian and its inverse. For example, the BFGS algorithm uses a quasi-Newton approximation of the Hessian.

Terminal convergence rate : Newton's algorithm typically converges quadratically to a local minimum of the objective function, meaning that the error between the current and optimal solution decreases quadratically with the number of iterations. However, this convergence rate is only guranteed when certain conditions are met, such as the objective function being twice continuously differentiable and the Hessian matrix being positive-definite. In practice the convergence rate may be slower especially when the objective function has multiple local minimum

(6) Consider the problem of minimizing $f : \mathbb{R}^n \to \mathbb{R}$. Be sure to define all terms you use.

(a) Explain the advantages and disadvantages of Quasi-Newton methods for minizing f.

(b) Provide details on the BFGS Quasi-Newton scheme in both the Hessian and Inverse Hessian forms. Be sure to state the optimization principle motivating the update, how to prevent a non-convex local problem, and a local convergence result. There is no need to provide details on the line search or trust region implementation.

(c) Derive and explain the SR1 update formula. Briefly compare and contrast SR1 to BFGS.

Soln :

(a) Advantages of Quasi-Newton method

(i) No need to compute the exact Hessian. The quasi-Newton methods use approximations to the Hessian matrix based on the gradient of the function, which can be computationally less expensive to compute.

(ii) converges faster than the first-order methods (such as gradient descent method).

(iii) Suitable for large scale problems.

Disadvantages :

(i) The performance of the quasi-Newton's methods can be sensitive to the initial guess, which can lead to convergence to a non-global minimum.

(ii) computationally expensive for high-dimensional problem because the Hessian approximation is updated at each iteration.

(b)

BFGS is a quasi-Newton method that uses an approximation of the Hessian matrix to iteratively update the search direction. The goal is to find a minimizer of the function $f(x)$ using the following optimization principle:

At iteration k, we want to find a search direction p_k that minimizes the quadratic approximation of $f(x)$ around the current point x_k.

$$f(x_k + p) = f(x_k) + \nabla f(x_k)^T p_k + \frac{1}{2} p_k^T B_k p_k$$

Where B_k is the approximation of the Hessian matrix.

The BFGS updates the Hessian approximation at each iteration using the following formula:

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T y_k} \quad \longrightarrow (1)$$

Where

$$s_k = x_{k+1} - x_k \quad \text{and} \quad y_k = \nabla f(x_{k+1}) - \nabla f(x_k) \quad \longrightarrow (2)$$

B_k has to be symmetric positive definite matrix and s_k and y_k needs to satisfy the curvature condition

$$s_k^T y_k > 0. \quad \longrightarrow (3)$$

For the non-convex case, we need to enforce the condition (3) explicitly by imposing restrictions on the line search procedure that chooses step length α. The condition (3) is guranteed to hold if we impose the Wolfe or strong Wolfe conditions on the line search.

Using the wolf condition and equation (2), we note that $\nabla f_{k+1}^T s_k \geq c_2 \nabla f_k^T s_k$ and therefore

$$y_k^T s_k \geq (c_2 - 1) \alpha_k \nabla f_k^T p_k. \quad \text{——— (4)}$$

Since $c_2 < 1$ and p_k is the descent direction, the term on the righ side of (4) is positive and the curvature condition (2) is satisfied.

BFGS for inverse Hessian matrix

The BFGS updates for inverse Hessian matrix is defined by the formula

$$H_{k+1} = (I - \rho_k s_k y_k^T) H_k (I - \rho_k y_k s_k^T) + \rho_k s_k s_k^T$$

Where H_k is the inverse Hessian matrix and

$$\rho_k = \frac{1}{y_k^T s_k}$$

The non-convexity case is checked by ensuring the the curvature condition

$$\rho_k = \frac{1}{y_k^T s_k} > 0,$$

so that the positive definiteness of H_k is preserved.

BFGS method has a local convergence result, which states that if the function $f(x)$ is twice continuously differentiable and the Hessian is positive definite at the solution x^*, then the iterates generated by the BFGS method converge to x^* at least linearly.

1. (a) Find the closest distance from the point $P = (5,0,5,0) \in \mathbb{R}^4$ to the column space of the matrix

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}$$

(b) Find the minimal-norm solution of the system of linear equations $Ax = b$, where $A = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}$ and $b = (2,-1)^T$.

(c) Using a one-dimensional, smooth objective function, $f : \mathbb{R} \to \mathbb{R}$, show that finding an optimizer x^* can be ill-conditioned. Formulate this result in terms of a higher dimensional objective function $f : \mathbb{R}^n \to \mathbb{R}$.

Solution

(a) We want to find the closest distance from the point $P = (5,0,5,0)$ to the column space of the matrix

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}$$

The projection of P onto the column space A is given by

$$P_A = A(A^TA)^{-1}A^TP$$

We have

$$A^TA = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$$

$$\therefore (A^TA)^{-1} = \frac{1}{5}\begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix}$$

$$A^TP = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}\begin{pmatrix} 5 \\ 0 \\ 5 \\ 0 \end{pmatrix} = \begin{pmatrix} 5 \\ 5 \end{pmatrix}$$

$$\therefore P_A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}\frac{1}{5}\begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix}\begin{pmatrix} 5 \\ 5 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 2 \\ 1 \end{pmatrix}$$

Therefore, the closest distance from P to the column space is $\|P - P_A\| = \sqrt{(5-1)^2 + (0-3)^2 + (5-2)^2 + (0-1)^2} = \sqrt{16+9+9+1} = \sqrt{35}$ Ans

(b) we want to find the minimal-norm solution of the system of linear equations $Ax = b$ where

$$A = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}, \quad b = (2, -1)^T$$

The normal equations are given by $A^T A\hat{x} = A^T b$, where $\hat{x}$ is the solution of the least squares problem.

Now

$$A^T A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 1 & 2 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix}$$

$$A^T b = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 2 \\ -1 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ -1 \\ 2 \end{pmatrix}$$

Solving

(c) Consider the one-dimensional smooth function

$$f(x) = x^2$$

The minimum of this function occurs at $x = 0$.

However, if we introduce a small perturbation to this function by adding a small quadratic term with a small coefficient ε, we get

$$f(x) = x^2 + \varepsilon x^2$$

The minimum still occurs at $x = 0$, but as ε becomes smaller, the 2nd-derivative of the function at the minimum becomes larger in magnitude, making the problem ill-conditioned. In other words, a small change in the input ε leads to a large change in the output of the function.

We can generalize this higher dimensional function

$$f(x) = \sum_{i=1}^{n} x_i^2$$

The minimum is at $x = 0$. If we do a small perturbation to this function say

$$f(x) = \sum_i x_i^2 + \varepsilon \sum_i x_i^2$$

The minimum is still at $x = 0$ but as ε-smaller, the condition number of the Hessian matrix at the minimum becomes larger making the problem ill-conditioned.