Newton and Quasi-Newton Methods

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

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

- This implies

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

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

- Comparing above update formula with $x^{k+1} = x^k + \alpha_k d^k$ we have $\alpha_k = 1$ and $d^k = -\left(\nabla^2 f\left(x^k\right)\right)^{-1} \nabla f\left(x^k\right)$
- d^k is a descent direction since

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

- The last inequality holds since $\nabla^2 f\left(x^k\right)$ is positive definite (as f is strictly convex) implies $\left(\nabla^2 f\left(x^k\right)\right)^{-1}$ 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 $x^* = (2,4)^T$

- Limitations on Newton method:
  - d is not a descent direction if f is not strictly convex at this point.
  - Requires Hessian value at every iterating point which increases computational cost.
  - Converges locally: *converges to the solution only when initial approximation is close to solution.*
- Possible steps to avoid these limitations:
  - We can use positive definite approximation of Hessian two avoid Hessian computation as well as to find descent direction at iterating points where the function is not strictly convex.
  - Line search techniques (exact/inexact) can be used to develop globally convergent methods.

Note: An optimization technique is said to be a globally convergent if from any initial approximation we can find a stationary point using this technique.

- BFGS quasi-Newton method:
  - Formula for descent direction at x^{k+1} is

$$d^k := -\left(B^k\right)^{-1} \nabla f\left(x^k\right)$$

where B^k is a positive definite approximation of $\nabla^2 f\left(x^k\right)$.
  - Different techniques are used to update B^{k+1}.
  - We use BFGS update formula as

$$B^{k+1} = B^k + \frac{s^k {s^k}^T}{{s^k}^T \delta^k} - \frac{B^k \delta^k {\delta^k}^T B^k}{{\delta^k}^T B^k \delta^k} \qquad (1)$$

where $\delta^k = x^{k+1} - x^k,\ s^k = \nabla f\left(x^{k+1}\right) - \nabla f\left(x^k\right)$
  - B^{k+1} is positive definite if ${s^k}^T \delta^k > 0$.
  - Note that $\delta^k = x^{k+1} - x^k = \alpha_k d^k$. This implies

$$\begin{aligned} {s^k}^T \delta^k &= \left(\nabla f\left(x^{k+1}\right) - \nabla f\left(x^k\right)\right)^T \alpha_k d^k \\ &\geq \alpha_k (c_2 - 1) \nabla f\left(x^k\right)^T d^k > 0 \end{aligned}$$

Where the second inequality follows from Wolfe condition and the last inequality holds since $c_2 < 1$ and $\nabla f\left(x^k\right)^T d^k < 0$.

- Algorithm:
  - Step 0: Select f, x^0, B^0 β_1, β_2 $(0 < \beta_1 < \beta_2 < 1)$, $r \in (0,1)$, and $\varepsilon > 0$.

  set $k := 0$
  - Step 1: If $\left\|\nabla f\left(x^k\right)\right\| < \varepsilon$ then stop. Otherwise go to Step 2.
  - Step 2: Calculate $d^k = -\left(B^k\right)^{-1} \nabla f\left(x^k\right)$
  - Step 3: Choose step length α_k as first element in the sequence $\{1, r, r^2, r^3, \dots\}$ satisfying Armijo-Wolfe conditions.
  - Step 4: Update $x^{k+1} := x^k + \alpha_k d^k$
  - Step 5: Update B^{k+1} using (1). Set $k := k + 1$ and go to Step 1.

- Example: Consider $f(x) = (x_1 - 1)^2 + (x_2 - x_1^2)^2$
- Then $\nabla f(x) = \begin{pmatrix} 2(x_1 - 1) - 4x_1(x_2 - x_1^2) \\ 2(x_2 - x_1^2) \end{pmatrix}$ and

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

- Choose $x^0 = (0,3)^T$. Then $\nabla^2 f(x^0) = \begin{pmatrix} -10 & 0 \\ 0 & 2 \end{pmatrix}$
- Observe that $\nabla^2 f(x^0)$ is not positive definite. So we can not use Newton method to solve this problem

k	x^k	$f(x^k)$	$\nabla f(x^k)$	B^k	d^k	α_k	x^{k+1}	B^{k+1}
0	$\begin{pmatrix}0\\3\end{pmatrix}$	10.0	$\begin{pmatrix}-2\\6\end{pmatrix}$	$\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}$	$\begin{pmatrix}2\\-6\end{pmatrix}$	0.5	$\begin{pmatrix}1\\0\end{pmatrix}$	$\begin{pmatrix}2.1 & -1.3\\-1.3 & 2.2333\end{pmatrix}$
1	$\begin{pmatrix}1\\0\end{pmatrix}$	1.0	$\begin{pmatrix}4\\-2\end{pmatrix}$	$\begin{pmatrix}2.1 & -1.3\\-1.3 & 2.2333\end{pmatrix}$	$\begin{pmatrix}-2.1112\\-0.3334\end{pmatrix}$	0.25	$\begin{pmatrix}0.4722\\-0.08335\end{pmatrix}$	$\begin{pmatrix}8.9618 & -3.0353\\-3.0353 & 2.5755\end{pmatrix}$
2	$\begin{pmatrix}0.4722\\-0.08335\end{pmatrix}$	0.3724	$\begin{pmatrix}-0.477\\-0.6126\end{pmatrix}$	$\begin{pmatrix}8.9618 & -3.0353\\-3.0353 & 2.5755\end{pmatrix}$	$\begin{pmatrix}0.2227\\0.5003\end{pmatrix}$	1.0	$\begin{pmatrix}0.6949\\0.41695\end{pmatrix}$	$\begin{pmatrix}8.42 & -3.6479\\-3.6479 & 2.5848\end{pmatrix}$
3	$\begin{pmatrix}0.6949\\0.41695\end{pmatrix}$	0.0974	$\begin{pmatrix}-0.4269\\-0.1319\end{pmatrix}$	$\begin{pmatrix}8.42 & -3.6479\\-3.6479 & 2.5848\end{pmatrix}$	$\begin{pmatrix}0.1874\\0.3154\end{pmatrix}$	1.0	$\begin{pmatrix}0.8823\\0.73235\end{pmatrix}$	$\begin{pmatrix}8.5087 & -3.9324\\-3.9324 & 2.4623\end{pmatrix}$
4	$\begin{pmatrix}0.8823\\0.73235\end{pmatrix}$	0.01598	$\begin{pmatrix}-0.0727\\-0.0922\end{pmatrix}$	$\begin{pmatrix}8.5087 & -3.9324\\-3.9324 & 2.4623\end{pmatrix}$	$\begin{pmatrix}0.0987\\0.195\end{pmatrix}$	1.0	$\begin{pmatrix}0.981\\0.9274\end{pmatrix}$	$\begin{pmatrix}9.6864 & -4.0203\\-4.0203 & 2.1487\end{pmatrix}$
5	$\begin{pmatrix}0.981\\0.9274\end{pmatrix}$	0.0016	$\begin{pmatrix}0.1004\\-0.0705\end{pmatrix}$	$\begin{pmatrix}9.6864 & -4.0203\\-4.0203 & 2.1487\end{pmatrix}$	$\begin{pmatrix}0.0146\\0.0601\end{pmatrix}$	1.0	$\begin{pmatrix}0.9956\\0.9872\end{pmatrix}$	$\begin{pmatrix}9.6859 & -3.903\\-3.903 & 1.9878\end{pmatrix}$

Observe that $\{x^k\}$ converging to $x^* = (1,1)^T$. Using stopping criteria $\|\nabla f(x^k)\| < 10^{-3}$ the final solution is obtained as $x^7 = (0.99982, 0.99955)^T \cong (1,1)^T$.

- DFP- Method:
  - Calculating ${B^k}^{-1}$ increases number of computations in BFGS method
  - To avoid calculating matrix inverse, in DFP method we generate a sequence of positive definitive matrices $\{H^k\}$, where $H^k$is an approximation of $(\nabla^2 f(x))^{-1}$.
  - In this method d^k is calculated as $d^k = -H^k \nabla f\left(x^k\right)$
  - H^k is updates using

$$H^{k+1} = H^k + \frac{\delta^k {\delta^k}^T}{{s^k}^T \delta^k} - \frac{H^k s^k {s^k}^T H^k}{{s^k}^T H^k s^k} \qquad (2)$$

where $\delta^k = x^{k+1} - x^k,\ s^k = \nabla\, f\left(x^{k+1}\right) - \nabla\, f\left(x^k\right)$.

  - Consider $f(x) = (x_1 - 1)^2 + (x_2 - x_1^2)^2$ and $x^0 = (0,3)^T$
  - Similar to BFGS method, detail computations are provided in the next table.

k	x^k	$f(x^k)$	$\nabla f(x^k)$	B^k	d^k	α_k	x^{k+1}	B^{k+1}
0	$\begin{pmatrix}0\\3\end{pmatrix}$	10.0	$\begin{pmatrix}-2\\6\end{pmatrix}$	$\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}$	$\begin{pmatrix}2\\-6\end{pmatrix}$	0.5	$\begin{pmatrix}1\\0\end{pmatrix}$	$\begin{pmatrix}0.6733 & 0.38\\0.38 & 0.66\end{pmatrix}$
1	$\begin{pmatrix}1\\0\end{pmatrix}$	1.0	$\begin{pmatrix}4\\-2\end{pmatrix}$	$\begin{pmatrix}0.6733 & 0.38\\0.38 & 0.66\end{pmatrix}$	$\begin{pmatrix}-1.933\\-0.20\end{pmatrix}$	0.5	$\begin{pmatrix}0.0335\\-0.1\end{pmatrix}$	$\begin{pmatrix}0.2235 & 0.1984\\0.1984 & 0.5977\end{pmatrix}$
2	$\begin{pmatrix}0.0335\\-0.1\end{pmatrix}$	0.9443	$\begin{pmatrix}-1.9194\\-0.2022\end{pmatrix}$	$\begin{pmatrix}0.2235 & 0.1984\\0.1984 & 0.5977\end{pmatrix}$	$\begin{pmatrix}0.4691\\0.5017\end{pmatrix}$	1	$\begin{pmatrix}0.5026\\0.4017\end{pmatrix}$	$\begin{pmatrix}0.4697 & 0.3508\\0.3508 & 0.5644\end{pmatrix}$
3	$\begin{pmatrix}0.5026\\0.4017\end{pmatrix}$	0.2696	$\begin{pmatrix}-1.2945\\0.2982\end{pmatrix}$	$\begin{pmatrix}0.4697 & 0.3508\\0.3508 & 0.5644\end{pmatrix}$	$\begin{pmatrix}0.5034\\0.2858\end{pmatrix}$	1	$\begin{pmatrix}1.006\\0.6875\end{pmatrix}$	$\begin{pmatrix}0.3071 & 0.3156\\0.3156 & 0.5688\end{pmatrix}$
4	$\begin{pmatrix}1.006\\0.6875\end{pmatrix}$	0.1054	$\begin{pmatrix}1.318\\-0.6491\end{pmatrix}$	$\begin{pmatrix}0.3071 & 0.3156\\0.3156 & 0.5688\end{pmatrix}$	$\begin{pmatrix}-0.1999\\-0.0468\end{pmatrix}$	1	$\begin{pmatrix}0.8061\\0.6407\end{pmatrix}$	$\begin{pmatrix}0.2016 & 0.2188\\0.2188 & 0.5073\end{pmatrix}$
5	$\begin{pmatrix}0.8061\\0.6407\end{pmatrix}$	0.0376	$\begin{pmatrix}-0.3584\\-0.0182\end{pmatrix}$	$\begin{pmatrix}0.2016 & 0.2188\\0.2188 & 0.5073\end{pmatrix}$	$\begin{pmatrix}0.0762\\0.0876\end{pmatrix}$	1	$\begin{pmatrix}0.8823\\0.7283\end{pmatrix}$	$\begin{pmatrix}0.4045 & 0.5501\\0.5501 & 0.9435\end{pmatrix}$

Observe that $\{x^k\}$ converging to $x^* = (1,1)^T$. Using stopping criteria $\|\nabla f(x^k)\| < 10^{-3}$ the final solution is obtained as $x^{11} = (1.00025, 1.0006)^T \cong (1,1)^T$.

- Advantages of quasi-Newton methods:
  - Does not require Hessian computations
  - Converges globally
  - Order of convergence is superlinear.

Definition: A sequence $\{x^k\}$ is said to converge with order $q \geq 1$ to x^* if $lim_{k\to\infty} \frac{\|x^{k+1}-x^*\|}{\|x^k-x^*\|^q} < M$ for some $M > 0$.

- $q = 1$ is called *linear convergence* $(M < 1)$.
- $q = 2$ is called *quadratic convergence.*
- $\{x^k\}$ is said to converge superlinearly if $lim_{k\to\infty} \frac{\|x^{k+1}-x^*\|}{\|x^k-x^*\|} = 0.$

For example

- $\{x^k\} = \{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \dots, \frac{1}{2^k}, \dots\}$ converges linearly to 0.
- $\{x^k\} = \{\frac{1}{2}, \frac{1}{4}, \frac{1}{16}, \frac{1}{256}, \dots, \frac{1}{2^{2^k}}, \dots\}$ converges quadratically to 0.
- $\{x^k\} = \{1, \frac{1}{4}, \frac{1}{27}, \frac{1}{64}, \dots, \left(\frac{1}{k}\right)^k, \dots\}$ converges superlinearly to 0.

- In line search techniques:
  - Steepest descent method converges linearly.
  - Newton method converges quadratically.
  - Quasi-Newton method converges superlinearly.