Practical Newton's Method

Lecture-20

Newton's Method

$$p_k^n = -(\nabla^2 f(x_k))^{-1} \nabla f(x_k)$$

$$\nabla^2 f(x_k) p_k^n = -\nabla f(x_k)$$

- Pure Newton's method converges rapidly once it is close to x^*.
- It may not converge from the remote starting point

- The search direction to be a descent direction
  - True if the Hessian is Positive Definite
  - Otherwise it may be ascent, or may be excessively long

- Two Strategies:
  - Newton GC: Solve Linear System using GC, terminate if neg curvature encountered
  - Modified Newton: Modify Hessian before or during the solution

Inexact Newton Steps

Iterative method to solve linear system, terminate at some approximate solution.

$$\nabla^2 f(x_k) p_k^n = -\nabla f(x_k)$$

Residual $r_k = \nabla^2 f(x_k) p_k + \nabla f(x_k)$

Scale dependent

Relative Residual $\dfrac{\|r_k\|}{\|\nabla f(x_k)\|} \le \eta_k$

Terminate iterations if:

$\|r_k\| \le \eta_k \|\nabla f(x_k)\| \quad 0 \prec \eta_k \prec 1, \forall k$ $\quad \eta_k$ is the forcing sequence

How about? $\eta_k = 1,$

Theorem 6.1

Suppose that $\nabla(x_k)$ is continuously differentiable in a neighborhood of a minimizer x^*, and assume that $\nabla^2 f(x^*)$ is positive definite. Consider the iteration $x_{k+1} = x_k + p_k$, where p_k satisfies $\|r_k\| \le \eta_k \|\nabla f(x_k\| \quad 0 \prec \eta_k \prec 1, \forall k$, then , if the starting point x_0 is sufficiently near x^*, the sequence $\{x_k\}$ converges to x^* linearly. That is, for all *K* sufficiently large, we have:

$$\|x_{k-1} - x^*\| \le c\|x_k - x^*\|, \quad 0 \prec c \prec 1$$

Theorem 3.7 (Newton) (Lecture-6)

Suppose that f is twice differentiable and that Hessian is Lipschitz continuous. Consider the iteration where p_k is given by

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

Then:

1. If the starting point x_0 is sufficiently close to x^*, the sequence converges to x^*.

2. The rate of convergence is quadratic
3. The sequence of gradient norms $\| \nabla f(x_k) \|$ converges quadratically to zero.

$$r_k = \nabla^2 f(x_k) p_k + \nabla f(x_k)$$

$$p_k = (\nabla^2 f(x_k))^{-1} (r_k - \nabla f(x_k))$$

If Hessian is PD $\quad \| \nabla^2 f(x_k)^{-1} \| \le L$

$$\|p_k\| \le L\left(\|\nabla f(x_k)\| + \|r_k\|\right) \le 2L\|\nabla f(x_k)\| \qquad \|r_k\| \le \eta_k \|\nabla f(x_k)\|$$

Taylor Series

$$\nabla f(x_{k+1}) = \nabla f(x_k) + \nabla^2 f(x_k) p_k + O\left(\|p_k\|^2\right)$$

$$= \nabla f(x_k) - (\nabla f(x_k) - r_k) + O\left(L^2 \|\nabla f(x_k)\|^2\right)$$

$$x_{k+1} = x_k + p$$

$$\nabla f(x_{k+1}) = r_k + O\left(\|\nabla f(x_k)\|^2\right)$$

$$\|\nabla f(x_{k+1})\| \le \eta_k \|\nabla f(x_k)\| + O\left(\|\nabla f(x_k)\|^2\right) \qquad \|r_k\| \le \eta_k \|\nabla f(x_k)\|$$

$$\|\nabla f(x_{k+1})\| \le \eta_k \|\nabla f(x_k)\| + O\left(\|\nabla f(x_k)\|^2\right)$$

$$\frac{\|\nabla f(x_{k+1})\|}{\|\nabla f(x_k)\|} \le \eta_k + \frac{O\left(\|\nabla f(x_k)\|^2\right)}{\|\nabla f(x_k)\|}$$

$$\frac{\|\nabla f(x_{k+1})\|}{\|\nabla f(x_k)\|} \le \eta_k + O\left(\|\nabla f(x_k)\|\right)$$

If x_k is chosen close to x^*, we can expect $\|\nabla f(x)\|$ to decrease by a factor of approximately $\eta_k < 1$ at every iteration.

$$\limsup_{k\to\infty} \frac{\|\nabla \mathrm{f}(\mathrm{x}_{\mathrm{k+1}})\|}{\|\nabla \mathrm{f}(\mathrm{x}_{\mathrm{k}})\|} \le \eta < 1$$

If $r_k = o\left(\|\nabla f(x_k)\|\right)$ $\qquad \limsup_{k\to\infty} \frac{\|\nabla f(x_{k+1})\|}{\|\nabla f(x_k)\|} = 0$

$$\|r_k\| \le \eta_k \|\nabla f(x_k)\|$$

If $r_k = O\left(\|\nabla f(x_k)\|^2\right)$ $\qquad \limsup_{k\to\infty} \frac{\|\nabla f(x_{k+1})\|}{\|\nabla f(x_k)\|^2} = c$

$$\nabla f(x_{k+1}) = r_k + O\left(\|\nabla f(x_k)\|^2\right)$$

$$\nabla f(x_{k+1}) = O\left(\|\nabla f(x_k)\|^2\right) + O\left(\|\nabla f(x_k)\|^2\right)$$

$$\nabla f(x_{k+1}) = O\left(\|\nabla f(x_k)\|^2\right)$$

Theorem 6.2

Suppose that the conditions of Theorem 6.1 hold and assume that the iterates $\{x_k\}$ generated by the inexact Newton method converges to x^*. Then the rate of convergence is super-linear if $\eta_k \to 0$ and quadratic if $\eta_k = O(\|\nabla f(x_k)\|)$.

Quadratic

$$\eta_k = \min(.5, \|\nabla f(x_k)\|)$$

$$\|r_k\| \le \eta_k \|\nabla f(x_k)\|$$

$$\|r_k\| \le \|\nabla f(x_k)\| \, \|\nabla f(x_k)\|$$

$$\|r_k\| \le \|\nabla f(x_k)\|^2$$

$$r_k = O\|\nabla f(x_k)\|^2$$

$$\nabla f(x_{k+1}) = r_k + O\left(\|\nabla f(x_k)\|^2\right)$$

$$\nabla f(x_{k+1}) = O\left(\|\nabla f(x_k)\|^2\right) + O\left(\|\nabla f(x_k)\|^2\right)$$

$$\nabla f(x_{k+1}) = O\left(\|\nabla f(x_k)\|^2\right)$$

$$\lim_{k\to\infty} \frac{\nabla f(x_{k+1})}{\|\nabla f(x_k)\|^2} = c$$

Line-Search Newton-CG Method

1. The starting point for GC iteration is

2. Negative curvature test. If the search direction satisfies

$$\left(p^{(i)}\right)^T A p^{(i)} \le 0$$

If i=0, complete the first GC, compute the new iterate $x^{(1)}$, stop

If i>0, stop the first GC, return most recent solution

3. The Newton step p_k is defined as the final CG iterate $x^{(f)}$

Algorithm 6.1

Algorithm 6.1 (Line Search Newton - CG)
given initial point x_0
for $k = 1,2,\ldots,n$
Compute a search direction p_k by applying the CG method to
$\nabla^2 f(x_k)p = -\nabla f_k$ starting from $x^{(0)} = 0$. Terminate when
$\| r_k \| \le \min(0.5, \sqrt{\| \nabla f_k \|}) \| \nabla f(x_k) \|$, or if the negative curvature is
is encountered
Set $x_{k+1} = x_k + \alpha_k p_k$, where α_k satisfies Wolfe backtracking conditions
end

Problems

- If Hessian is nearly singular, Newton-CG direction can be long, requiring many function evaluations.
- The reduction in function may be very small.
- Normalize the Newton's direction
- Introduce threshold $(p^{(i)})^T A p^{(i)} \leq 0$

Algorithm 6.2

Algorithm 6.2 (Line Search Newton with Modification)
given initial point x_0
for $k = 1,2,\ldots,n$
Factorize the matrix $B_k = \nabla^2 f(x_k) + E_k$, where $E_k = 0$ if $\nabla^2 f(x_k)$ is sufficiently PD; otherwise, E_k is chosen to ensure that B_k is sufficiently PD
Solve $B_k p_k = -\nabla^2 f(x_k)$;
Set $x_{k+1} = x_k + \alpha_k p_k$, where α_k satisfies Wolfe backtracking conditions
end

Bounded Modified Factorization Property

The matrices in the sequence $\{B_k\}$ have bounded condition number whenever the sequence of Hessian $\{\nabla^2 f(x_k)\}$ is bounded, that is:

$$cond(B_k) = \|B_k\|\|B_k^{-1}\| \le C, \text{ for some } C > 0, \forall k$$

Hessian Modification

Choose modification E_k such that matrix $B_k = \nabla^2 f(x_k) + E_k$ is sufficiently PD.

-modification to be well-conditioned
-small, so that second order information is preserved
-modification be computable at moderate cost

Eigenvalue Modification

$$\nabla f(x_k) = (1,-3,2)$$

$$\nabla f^2(x_k) = diag(10,3,-1) = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

Spectral decomposition $Q = I$ and $\Lambda = \text{diag}(\lambda_1, \lambda_2, \lambda_3)$

$$\nabla^2 f(x_k) = Q\Lambda Q^T = \sum_{i=1}^{n} \lambda_i q_i q_i^T$$

$$p_k^N = -(\nabla^2 f(x_k)^{-1} \nabla f(x_k) = -\begin{bmatrix} .1 & & \\ & .33 & \\ & & 1 \end{bmatrix}\begin{bmatrix} 1 \\ -3 \\ 2 \end{bmatrix}$$

$p_k^N = (-.1,1,2)$

$$\nabla f(x_k)^T p_k^N = \begin{bmatrix} 1 \\ -3 \\ 2 \end{bmatrix}^T \begin{bmatrix} -.1 \\ 1 \\ 2 \end{bmatrix} = -.1 - 3 + 4 = .9 > 0$$

It is not a descent direction

Replace all negative eigenvalues by small positive numbers.

$$B_k = \sum_{i=1}^{2} \lambda_i q_i q_i^T + \delta q_3 q_3^T = diag(10,3,10^{-8})$$

$$\delta = 10^{-8}$$

$$p_k = -B_k^{-1} \nabla f_k = -\sum_{i=1}^{2} \frac{1}{\lambda_i} q_i (q_i^T \nabla f_k) - \frac{1}{\delta} q_3 (q_3^T \nabla f(x_k)) \approx -(2x10^8) q_3$$

For small δ this step is nearly parallel to q_3 and very long. Although f decreases along the direction p_k, its extreme length violates the sprit of Newton's method, which relies on the quadratic approximation of the objective function.

Flip the signs of negative eigenvalues, in our case Set

$$\delta = 1$$

Set the last term zero, so that the search direction has no component along the negative curvature directions, adapt the choice of δ to ensure the length of the step is not excessive.

$$p_k = -B_k^{-1}\nabla f_k = -\sum_{i=1}^{2}\frac{1}{\lambda_i}q_i(q_i^T\nabla f_k) - \frac{1}{\delta}q_3(q_3^T\nabla f(x_k)) \approx -(2x10^8)q_3$$