

Optimization Theory

Lecture 09

Fudan University

luoluo@fudan.edu.cn

1 Newton's Method

2 Damped Newton Method

1 Newton's Method

2 Damped Newton Method

Newton's Method

Recall that optimizing smooth function $f(\mathbf{x})$ by gradient descent

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{L} \nabla f(\mathbf{x}_t)$$

is based on minimizing RHS of

$$f(\mathbf{y}) \leq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{y} - \mathbf{x}_t \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}_t\|_2^2.$$

In a local region, we can minimize the RHS of

$$f(\mathbf{y}) \approx f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{y} - \mathbf{x}_t \rangle + \frac{1}{2} \langle \mathbf{y} - \mathbf{x}_t, \nabla^2 f(\mathbf{x}_t)(\mathbf{y} - \mathbf{x}_t) \rangle.$$

Suppose $\nabla^2 f(\mathbf{x}_t)$ is non-singular, then we achieve Newton's method

$$\mathbf{x}_{t+1} = \mathbf{x}_t - (\nabla^2 f(\mathbf{x}_t))^{-1} \nabla f(\mathbf{x}_t).$$

Quadratic Convergence

Theorem

Suppose the twice differentiable function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ has L_2 -Lipschitz continuous Hessian and local minimizer \mathbf{x}^* with $\nabla^2 f(\mathbf{x}^*) \succeq \mu \mathbf{I}$, then the Newton's method

$$\mathbf{x}_{t+1} = \mathbf{x}_t - (\nabla^2 f(\mathbf{x}_t))^{-1} \nabla f(\mathbf{x}_t)$$

with $\|\mathbf{x}_0 - \mathbf{x}^*\|_2 \leq \mu/(2L_2)$ holds that

$$\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2 \leq \frac{L_2}{\mu} \|\mathbf{x}_t - \mathbf{x}^*\|_2^2.$$

Newton's method has local quadratic convergence, which requires

$$T = \mathcal{O}(\ln \ln(1/\epsilon))$$

iterations to achieve $\|\mathbf{x}_T - \mathbf{x}^*\|_2 \leq \epsilon$.

Standard Newton's Method

Strengths:

- ① The quadratic convergence is very fast (even for ill-conditioned case).
- ② The algorithm is affine invariant.

Weakness:

- ① The convergence guarantee is local.
- ② The iteration is expensive for large d .

Outline

1 Newton's Method

2 Damped Newton Method

Newton's Method with Line Search

Algorithm 1 Newton's Method with Line Search

```
1: Input:  $\mathbf{x}_0 \in \mathbb{R}^d, \tau \in (0, 1), c_1 \in (0, 1)$ 
2: for  $t = 0, 1 \dots$ 
3:    $\mathbf{p}_t \leftarrow -(\nabla^2 f(\mathbf{x}_t))^{-1} \nabla f(\mathbf{x}_t)$ 
4:    $\alpha \leftarrow 1$ 
5:   while  $f(\mathbf{x}_t + \alpha \mathbf{p}_t) > f(\mathbf{x}_t) + c_1 \alpha \langle \nabla f(\mathbf{x}_t), \mathbf{p}_t \rangle$  do
6:      $\alpha \leftarrow \tau \alpha$ 
7:   end while
8:    $\mathbf{x}_{t+1} = \mathbf{x}_t + \alpha \mathbf{p}_t$ 
9: end for
```

- ❶ For strongly-convex $f(\cdot)$, the direction \mathbf{p}_t is a descent direction.
- ❷ What is the global convergence rate?

Damped Newton Method

The damped Newton method is based on

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{1 + M_f \lambda_f(\mathbf{x}_t)} (\nabla^2 f(\mathbf{x}_t))^{-1} \nabla f(\mathbf{x}_t),$$

where $M_f > 0$ and

$$\lambda_f(\mathbf{x}_t) = \sqrt{\left\langle \nabla f(\mathbf{x}_t), (\nabla^2 f(\mathbf{x}_t))^{-1} \nabla f(\mathbf{x}_t) \right\rangle}.$$

This method has global convergence guarantee under mild assumptions.