Newton's Method

Yuqian Zhang

Rutgers University

yqz.zhang@rutgers.edu

February 25, 2021

Overview

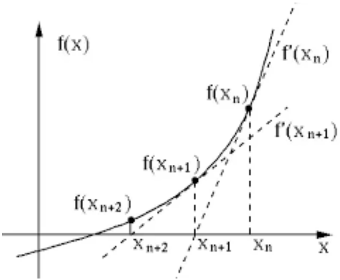
Classical Newton's Method

classical technique for finding the root of a general differentiable function $f : \mathbb{R} \to \mathbb{R}$ such that

$$f(x) = 0$$

- start from some guess x_0
- apply iteration

$$x_{(n+1)} = x_{(n)} - \frac{f(x_{(n)})}{f'(x_{(n)})}$$



Classical Newton's Method

- there can be many roots, and which one we converge to will depend on what we choose for x_0
- classical convergence theory that once we are close enough to a particular root x_0, we will have

$$\left|x_0 - x_{(n+1)}\right| \leq C \left(x_0 - x_{(n)}\right)^2, \quad C = \sup_{x \in \mathcal{I}} \frac{|f''(x)|}{2\,|f'(x)|}$$

- Newton's method exhibits quadratic convergence: the error at the next iteration is proportional to the square of the error at the last iteration.

Newton's Method

- $f(x)$ is convex, twice differentiable, and has a minimizer,

$$x^{(k+1)} = x^{(k)} - \frac{f'(x^{(k)})}{f''(x^{(k)})} \tag{1}$$

- if f is three-times continuously differentiable

$$\left| x_0 - x^{(k+1)} \right| \leq C \left(x_0 - x^{(k)} \right)^2, \quad C = \sup_{x \in \mathcal{I}} \frac{|f'''(x)|}{2\,|f''(x)|}$$

Interpreting Newton's Method

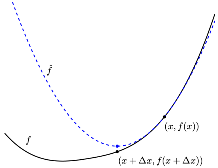
1. At $x^{(k)}$, approximate $f(x)$ using the Taylor expansion

$$f(x) \approx f(x^{(k)}) + f'(x^{(k)})(x - x^{(k)}) + \frac{1}{2}f''(x^{(k)})(x - x^{(k)})^2$$

2. Find the exact minimizer of above quadratic approximation

$$(\hat{x} - x^{(k)})f''(x^{(k)}) = -f'(x^{(k)})$$

3. Take $x^{(k+1)} = \hat{x}$



Interpreting Newton's Method

- Approximate $f(\boldsymbol{x})$ using the Taylor expansion

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}^{(k)}) + \left\langle \nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{x} - \boldsymbol{x}^{(k)} \right\rangle + \frac{1}{2} \left(\boldsymbol{x} - \boldsymbol{x}^{(k)} \right)^T \nabla^2 f(\boldsymbol{x}^{(k)}) \left(\boldsymbol{x} - \boldsymbol{x}^{(k)} \right)$$

- Find the exact minimizer of his quadratic approximation

$$\text{minimize} \quad \boldsymbol{g}^T \left(\boldsymbol{x} - \boldsymbol{x}^{(k)} \right) + \frac{1}{2} \left(\boldsymbol{x} - \boldsymbol{x}^{(k)} \right)^T \boldsymbol{H} \left(\boldsymbol{x} - \boldsymbol{x}^{(k)} \right)$$

with Hessian $\boldsymbol{H} = \nabla^2 f(\boldsymbol{x}^{(k)})$ and gradient $\boldsymbol{g} = \nabla f(\boldsymbol{x}^{(k)})$

Pure Newton step

- The minimizer $\hat{\boldsymbol{x}}$ satisfies

$$\boldsymbol{H}(\boldsymbol{x} - \boldsymbol{x}^{(k)}) = -\boldsymbol{g}$$

- If $\boldsymbol{H}$ is invertible, take

$$\boldsymbol{x}^{(k+1)} = \hat{\boldsymbol{x}} = \boldsymbol{x}^{(k)} - \boldsymbol{H}^{-1}\boldsymbol{g}$$

(Practical) Newton's Method

- $\boldsymbol{d}^{(k)}$: step direction

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

- t_k: backtracking line search

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + t_k \boldsymbol{d}^{(k)}$$

Assumptions

Suppose $f(\boldsymbol{x})$ is strongly convex

$$m\boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq M\boldsymbol{I}, \quad \forall x \in \mathbb{R}^n$$

and that its Hessian is Lipschitz

$$\left\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\right\| \leq L \left\|\boldsymbol{x} - \boldsymbol{y}\right\|_2$$

Main result

We will show that the Newton's algorithm coupled with an *exact line search* converges to precision ϵ

$$f(\boldsymbol{x}^{(k)}) - p^* \leq \epsilon$$

for a number of iterations

$$k \geq C_1 \left(f(\boldsymbol{x}^{(0)}) - p^* \right) + \log_2 \log_2(\epsilon_0/\epsilon)$$

where

$$C_1 = M^2 L^2 / m^5, \quad \epsilon_0 = 2m^3/L^2$$

Convergence of Newton's Method

- *Damped Newton stage*: far from the solution, large ∇f

$$f(\boldsymbol{x}^{(k+1)}) - f(\boldsymbol{x}^{(k)}) \leq 1/C_1$$

- *Quadratic convergence stage*: ∇f is small enough

$$\left\|\nabla f(\boldsymbol{x}^{(k)})\right\|_2 \leq C_2 \cdot 2^{-2^{k-\ell}}, \quad \forall k > \ell$$

where $C_2 = L/(2m^2)$

Damped phase

$$\left\|\nabla f(\boldsymbol{x}^{(k)})\right\|_2 \geq m^2/L$$

- take $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + t_{exact}\boldsymbol{d}^{(k+1)}$
- denote the *Newton decrement* denoted as

$$\lambda_k^2 = -\nabla f(\boldsymbol{x}^{(k)})^T \boldsymbol{d}^{(k+1)}$$

Damped phase

Since the largest eigenvalue of $\left(\nabla^2 f(x^{(k)})\right)^{-1}$ is at most $1/m$

$$\begin{aligned} f(\boldsymbol{x}^{(k)} + t\boldsymbol{d}^{(k+1)}) &\leq f(\boldsymbol{x}^{(k)}) - t\lambda_k^2 + \frac{M}{2}\left\|t\boldsymbol{d}^{(k+1)}\right\|_2^2 \\ &\leq f(\boldsymbol{x}^{(k)}) - t\lambda_k^2 + \frac{M}{2m}t^2\lambda_k^2 \end{aligned}$$

Plug in $t = m/M$, then

$$\begin{aligned} f(\boldsymbol{x}^{(k)} + t_{exact}\boldsymbol{d}^{(k+1)}) - f(\boldsymbol{x}^{(k)}) &\leq -\frac{m}{M}\lambda_k^2 \\ &\leq -\frac{m}{M^2}\left\|\nabla f(\boldsymbol{x}^{(k)})\right\|_2^2 \\ &\leq -\frac{m^5}{L^2 M^2} \end{aligned}$$

Quadratic convergence

$$\left\|\nabla f(\boldsymbol{x}^{(k)})\right\|_2 \leq m^2/L$$

- step size $t = 1$, $\alpha < 1/3$
- by construction $\nabla^2 f(\boldsymbol{x}^{(k)})\boldsymbol{d}^{(k+1)} = -\nabla f(\boldsymbol{x}^{(k)})$, then

$$\begin{aligned}
\nabla f(\boldsymbol{x}^{(k+1)}) &= \nabla f(\boldsymbol{x}^{(k)} + \boldsymbol{d}^{(k+1)}) - \nabla f(\boldsymbol{x}^{(k)}) - \nabla^2 f(\boldsymbol{x}^{(k)})\boldsymbol{d}^{(k+1)} \\
&= \int_0^1 \nabla^2 f(\boldsymbol{x}^{(k)} + t\boldsymbol{d}^{(k+1)})\boldsymbol{d}^{(k+1)} dt - \nabla^2 f(\boldsymbol{x}^{(k)})\boldsymbol{d}^{(k+1)} \\
&= \int_0^1 \left[\nabla^2 f(\boldsymbol{x}^{(k)} + t\boldsymbol{d}^{(k+1)}) - \nabla^2 f(\boldsymbol{x}^{(k)})\right] \boldsymbol{d}^{(k+1)} dt
\end{aligned}$$

Quadratic convergence

$$
\begin{aligned}
\left\|\nabla f(\boldsymbol{x}^{(k+1)})\right\|_2 &\le \int_0^1 \left\|\nabla^2 f(\boldsymbol{x}^{(k)} + t\boldsymbol{d}^{(k+1)}) - \nabla^2 f(\boldsymbol{x}^{(k)})\right\|_2 \left\|\boldsymbol{d}^{(k+1)}\right\|_2 dt \\
&\le \int_0^1 t^2 L \left\|\boldsymbol{d}^{(k+1)}\right\|_2^2 dt \\
&= \frac{L}{2}\left\|\left(\nabla^2 f(\boldsymbol{x}^{(k)})\right)^{-1} \nabla f(\boldsymbol{x}^{(k)})\right\|_2^2 \\
&\le \frac{L}{2m^2}\left\|\nabla f(\boldsymbol{x}^{(k)})\right\|_2^2
\end{aligned}
$$

Since $\left\|\nabla f(\boldsymbol{x}^{(k)})\right\|_2 \le m^2/L$, we have

$$
\frac{L}{2m^2}\left\|\nabla f(\boldsymbol{x}^{(k+1)})\right\|_2 \le \left(\frac{L}{2m^2}\left\|\nabla f(\boldsymbol{x}^{(k)})\right\|_2\right)^2 \le \left(\frac{1}{2}\right)^2
$$

Quadratic convergence

If we entered this stage at iteration ℓ, this means

$$\frac{L}{2m^2}\left\|\nabla f(\boldsymbol{x}^{(k)})\right\|_2 \leq \left(\frac{L}{2m^2}\left\|\nabla f(\boldsymbol{x}^{(l)})\right\|_2\right)^{2^{k-\ell}} \leq \left(\frac{1}{2}\right)^{2^{k-\ell}}$$

By strong convexity of f

$$f(\boldsymbol{x}^{(k)}) - p^* \leq \frac{1}{2m}\left\|\nabla f(\boldsymbol{x}^{(k)})\right\|_2^2 \leq \frac{2m^3}{L^2}\left(\frac{1}{2}\right)^{2^{k-\ell+1}}$$

Hence

$$k - \ell + 1 \geq \log_2 \log_2(\epsilon_0/\epsilon), \quad \epsilon_0 = 2m^3/L^2$$

Convergence criteria: the Newton decrement

- Newton's method is *affine invariant*
- Suppose $\tilde{f}(\boldsymbol{x}) = f(\boldsymbol{T}\boldsymbol{x})$ for some invertible $\boldsymbol{T} \in \mathbb{R}^{n\times n}$
- Newton's algorithm gives iterates $\tilde{\boldsymbol{x}}^{(k)} = \boldsymbol{T}^{-1}\boldsymbol{x}^{(k)}$
- Euclidean norm of the gradient is not affinely invariant:

$$\left\|\nabla \tilde{f}(\boldsymbol{x})\right\|_2 \neq \|\nabla f(\boldsymbol{T}\boldsymbol{x})\|_2$$

- Question: which norm should we use as the stopping criteria?

$$\left\|\nabla f(\boldsymbol{x}^{(k)})\right\|_? \leq \epsilon$$

Convergence criteria: the Newton decrement

A criteria that is affinely invariant is the Newton decrement:

$$\lambda(\boldsymbol{x}) = \|\nabla f(x)\|_{\boldsymbol{H}^{-1}} \doteq \sqrt{\boldsymbol{g}^T \boldsymbol{H}^{-1} \boldsymbol{g}}$$

with $\boldsymbol{g} = \nabla f(\boldsymbol{x})$ and $\boldsymbol{H} = \nabla^2 f(\boldsymbol{x})$.
If $\boldsymbol{d} = -\left(\nabla^2 f(\boldsymbol{x})\right)^{-1} \nabla f(\boldsymbol{x})$, then

$$\langle \boldsymbol{d}, \nabla f(\boldsymbol{x}) \rangle = -\lambda(\boldsymbol{x})^2$$

The convergence criteria for Newton's method is usually whether $\lambda(\boldsymbol{x}^{(k)})$ is below some threshold.

Self-concordant functions

Definition

- A convex function $f : \mathbb{R} \to \mathbb{R}$ is self-concordant if
$$\left| f'''(x) \right| \leq 2 f''(x)^{3/2}, \quad \forall\, x \in \operatorname{dom} f$$
- A function $f : \mathbb{R}^n \to \mathbb{R}$ is self-concordant if $g(t) = f(\boldsymbol{x} + t\boldsymbol{v})$ is self-concordant for all $\boldsymbol{x} \in \operatorname{dom} f$ and $v \in \mathbb{R}^n$.

If f is self-concordant, then Newton iterations coupled with standard backtracking line search will have $f(\boldsymbol{x}^{(k)}) - p^* \leq \epsilon$ after

$$k \geq C \left(f(\boldsymbol{x}^{(0)}) - p^* \right) + \log_2 \log_2 (1/\epsilon)$$

where C depends on backtracking parameters.

Convergence of descent algorithms

Gradient Descent

- Strongly convex

$$k \geq \frac{\log((f(x^{(0)}) - p^*)/\epsilon)}{\log(1 - m/M)}$$

- Lipschitz gradient

$$k \geq \frac{1}{2t\epsilon} \left\| x^{(0)} - x^* \right\|_2^2$$

Convergence of descent algorithms

Newton's Method

- strongly convex and Lipschitz Hessian

$$k \geq C_1 \left(f(\boldsymbol{x}^{(0)}) - p^* \right) + \log_2 \log_2(\epsilon_0/\epsilon)$$

here $C_1 = M^2 L^2 / m^5$ and $\epsilon_0 = 2m^3/L^2$

- Self-concordant functions

$$k \geq C \left(f(\boldsymbol{x}^{(0)}) - p^* \right) + \log_2 \log_2(1/\epsilon)$$