Class #17

Friday, March 28, 2025 10:20 AM

Newton's Method: Convergence Guarantees

Assumptions

① Function $f \in C^2(\mathbb{R}^n)$ and m-strongly convex

$\Rightarrow$ on set $S = \{x : f(x) \le f(x^{(0)})\}$

$$mI \preccurlyeq \nabla^2 f(x) \preccurlyeq MI \quad \forall \; x \in S$$

② We have L-Lipschitz Hessians

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \le L \|x - y\|_2$$

If 3^{rd} derivative exists $\Leftrightarrow$ Equivalent condition on 3^{rd} derivative that it is bounded.

$\Downarrow$ Assump. 2 can be replaced by working with self-concordant functions (optional reading in § 9.6).

e.g., $f(x) = x^T P x$; $P \succcurlyeq mI$

$$\nabla^2 f(x) = P$$

$\Rightarrow$ All quadratics satisfy Assump. 2.

Convergence behavior of Newton's Method

It has two phases of convergence:

① Phase I $\Rightarrow$ Damped phase $\Rightarrow$ It has linear convergence.

In this phase, backtracking provides a step size that satisfies:

$$t \geq \min\left\{\beta \frac{m}{M}, 1\right\}$$

② Phase II $\Rightarrow$ Quadratic Convergent phase

$\Rightarrow$ Full Newton step phase $\Rightarrow t = 1 \;\; \forall\, l \geq k$ (some fixed k)

In this phase, we have convergence behavior is

$$f(x^{(l)}) - p^* = O\left(c^{2^l}\right) \quad ; \quad c = 0.5$$

Summary of discussion

① Rapid (super linear / Quadratic) convergence eventually. Once in Quadratic Convergent phase, we need only six to eight more iterations to reach optimal value.

② Newton's method is also not affected by change of coordinates
↳ It is much less sensitive to the condition number of a problem.

③ Performance scales well with the number of dimensions.

④ Backtracking parameters also do not affect the performance that much.

Drawback: Memory and Computation

Reading: BV: §9.5.4

Damped Phase (Linearly Convergent Phase)

Newton's method gives us linear convergence from $k=0$ upto some finite as long as

$$\|\nabla f(x^{(k)})\|_2 \geq \eta \quad \text{for some}$$

$$0 < \eta < \frac{m^2}{L}$$

and $t \geq \min\{\beta \frac{m}{M}, 1\}$ for these iterations.

Basically, in each iteration, we will reduce the objective function by a constant $\gamma > 0$

$$\gamma = \alpha \beta \eta^2 \frac{m}{M^2}.$$

Reminder: $\lambda(x) = \left(\nabla f(x)^T (\nabla^2 f(x))^{-1} \nabla f(x)\right)^{1/2}$

$$\Updownarrow$$

$$\lambda(x)^2 = \Delta x_{nt}^T \nabla^2 f(x) \Delta x_{nt} \iff \nabla f(x)^T \Delta x_{nt} = -\lambda(x)^2$$

b/c $\Delta x_{nt} = -(\nabla^2 f(x))^{-1} \nabla f(x)$

Quadratic upper bound

$$f(x + t\Delta x_{nt}) \leq f(x) + t \underbrace{\nabla f(x)^T \Delta x_{nt}}_{-\lambda(x)^2} + \frac{Mt^2}{2} \|\Delta x_{nt}\|_2^2$$

$$\leq f(x) - t\lambda(x)^2 + \frac{Mt^2}{2} \|\Delta x_{nt}\|_2^2$$

b/c $\lambda^2(x) = \Delta x_{nt}^T \nabla^2 f(x) \Delta x_{nt} \geq m \|\Delta x_{nt}\|_2^2$

Remember: $\lambda_{min}(A) \le \frac{v^T A v}{\|v\|_2^2} \le \lambda_{max}(A)$; A is diagonalizable

Note: In class, we upperbounded $\lambda(x)^2$; we should have upper bounded $\|\Delta x_{nt}\|_2^2 \Rightarrow \|\Delta x_{nt}\|_2^2 \le \frac{\lambda^2(x)}{m}$. Below are the original steps in class, that led to an extra m factor. I am correcting them below, so you can tell what things were corrected in relation to class notes. Corrections will be in red.

$$\Rightarrow f(x + t\Delta x_{nt}) \le f(x) - \underbrace{\cancel{tm\|\Delta x_{nt}\|_2^2}}_{t\lambda(x)^2} + \underbrace{\cancel{\frac{Mt^2}{2}\|\Delta x_{nt}\|_2^2}}_{\frac{M}{2m}t^2\lambda(x)^2}$$

Pick $\boxed{\hat{t} = \frac{m}{M}}$

$$= f(x) - \frac{m^{\cancel{2}}}{M}\underbrace{\cancel{\|\Delta x_{nt}\|_2^2}}_{\lambda(x)^2} + \frac{M}{2m} \times \frac{m^2}{M^2} \cdot \underbrace{\cancel{\|\Delta x_{nt}\|_2^2}}_{\lambda(x)^2}$$

$$= f(x) - \underbrace{\left(\frac{m^{\cancel{2}}}{M} - \frac{m^{\cancel{2}}}{2M}\right)}_{\frac{m^{\cancel{2}}}{2M}} \underbrace{\cancel{\|\Delta x_{nt}\|_2^2}}_{\lambda(x)^2}$$

$$f(x + t\Delta x_{nt}) \le f(x) - \frac{1}{2}\cdot\frac{m^{\cancel{2}}}{M}\cancel{\|\Delta x_{nt}\|_2^2}\ \lambda(x)^2 \quad \to m$$

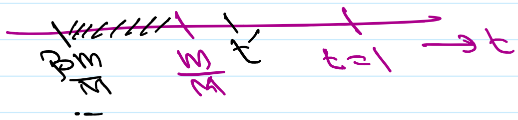
$$f(x + t\Delta x_{nt}) \le f(x) - \frac{1}{2}\frac{m}{M}\cancel{\|\Delta x_{nt}\|_2^2}\ \lambda(x)^2$$

If $\alpha \in (0, \frac{1}{2}]$

$$\Rightarrow \le f(x) - \alpha t \cancel{\|\Delta x_{nt}\|_2^2}\ \lambda(x)^2$$

$\Rightarrow \hat{t} = \frac{m}{M}$ satisfies the back tracking condition

(remember, back tracking in Newton's method uses $f(x) - \alpha t \lambda(x)^2$



(remember, back tracking in Newton's method uses $f(x) - \alpha t \lambda(x)^2$ condition)

we are bound to accept $t \geq \frac{\beta m}{M}$

$$f(x + t\Delta x_{nt}) \leq f(x) - \underbrace{\alpha t \|\cancel{\Delta x_{nt}}\|_2^2}_{\lambda(x)^2}$$

$$\Rightarrow \alpha t \lambda(x)^2 \geq \alpha \beta \frac{m}{M} \times \lambda(x)^2$$

$$\geq \alpha \beta \eta^2 \boxed{\frac{m}{M^2}}$$

Also:

$$\lambda(x)^2 \geq \frac{1}{\lambda_{max}(\nabla^2 f(x))} \times \|\nabla f(x)\|_2^2 \geq \frac{\eta^2}{M}$$

Quadratic Convergence Phase

Once $\|\nabla f(x^{(k)})\|_2$ goes below η, then we always have $t^{(k)} = 1$ and

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

How many iterations for Newton's method?

Linear phase $\Rightarrow$ $f(x)$ decreases by at least γ in each iteration

$$\Rightarrow \# \text{ of iterations} = \frac{f(x^{(0)}) - p^*}{\gamma}$$

Say we want final accuracy to be ϵ

$$f(x^{(l)}) - p^* \leq \epsilon \quad \forall\, l \geq K$$

$$= O\left(0.5^{2^l}\right) \leq \epsilon$$

$$\epsilon_0 = 2m^3/L^2$$

$\log_2 \log_2 \left(\frac{\epsilon_0}{\ } \right) = \#$ of iterations.

$\epsilon_0 = 2m^3/L^2$

$$\log_2 \log_2 \left(\frac{\epsilon_0}{\epsilon}\right) = \# \text{ of iterations.}$$

$$\# \text{ of iterations} \propto \log_2 \log_2 (\epsilon^{-1})$$

vs: $\quad GD \propto \log(\epsilon^{-1})$

$\nabla^2 f(x) =$