Chapter 10: Optimization for Data Science
Newton Method

TANN Chantara

Institute of Technology of Cambodia

October 8, 2022

Table of Contents

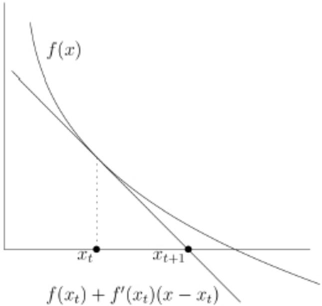
1-dimensional case

- Goal: Find x such that $f(x) = 0$
- Newton-Raphson method

$$x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)}, \quad t \geq 0$$

- x_{t+1} is the solution to

$$f(x_t) = f'(x_t)(x - x_t) = 0$$



Example: $f(x) = x^2 - R, \qquad x_0 = R$

$$x_{t+1} = x_t - \frac{x_t^2 - R}{2x_t} = \frac{1}{2}\left(x_t + \frac{R}{x_t}\right)$$

- How fast can this method converge?

1-dimensional case (cont'd)

- Note that
$$x_{t+1} = \frac{1}{2}\left(x_t + \frac{R}{x_t}\right) \geq \frac{x_t}{2},$$
hence, in order to achieve $x_t \leq 2\sqrt{R}$, we need at least $T \geq \log R/2$ steps.
- If we start closer to $\sqrt{R}$, we can converge faster, i.e., suppose we start at $x_0 - \sqrt{R} < 1/2$. Then we can show via induction that
$$x_{t+1} - \sqrt{R} = \frac{1}{2x_t}\left(x_t - \sqrt{R}\right)^2 \leq \left(x_t - \sqrt{R}\right)^2,$$
which implies
$$x_T - \sqrt{R} \leq \left(x_0 - \sqrt{R}\right)^{2^T} \leq \left(\frac{1}{2}\right)^{2^T}$$
$\Longrightarrow$ To get $x_T - \sqrt{R} < \varepsilon$ we only need $T = \log\log(1/\varepsilon)$ steps

Newton's method for optimization

- $f : \mathbb{R}^n \to \mathbb{R}, \ \ x_0 \in \mathbb{R}^n$ arbitrary
- Recall that solving $\min_{x \in \mathbb{R}^n} f(x)$ for a differentiable function is equivalent to solving $\nabla f(x) = 0$

Newton method:

$$x_{t+1} = x_t - \underbrace{\nabla^2 f(x_t)^{-1}}_{H(x_t)} \nabla f(x_t), \quad t \geq 0$$

- Gradient descent is of this form with $H(x_t) = \gamma \mathrm{I}$
  $\implies$ Netwon method is an *adaptive gradient descent*
- Computing $H(x_t)$ is costly (inversion of a matrix)

Convergence for quadratic functions

Lemma: Consider a nondegenerate quadratic function of the form

$$f(x) = \frac{1}{2}x^\top Mx - q^\top x + c,$$

where $M \in \mathbb{R}^{n\times n}$ is invertible, symmetric, $q \in \mathbb{R}^n$, $c \in \mathbb{R}$. Let $x^\star = M^{-1}q$ be the unique solution of $\nabla f(x) = 0$ (the unique global minimum if f is convex). With any starting point $x_0 \in \mathbb{R}^n$, Newton's method yields $x_1 = x^\star$.

Proof: We have $\nabla f(x) = Mx - q$, which implies $x^\star = M^{-1}q$ and $\nabla^2 f(x) = M$. Hence,

$$x_0 - \nabla^2 f(x_0)\nabla f(x_0) = x_0 - M^{-1}(Mx_0 - q) = M^{-1}q = x^\star$$

Local convergence

Key property of the Newton method: If we start close to the minimum, we reach distance at most ε within $\log\log(1/\varepsilon)$ steps

- fastest convergence we have seen so far
- requires to start close to the minimum already

Theorem: Let $f : \mathrm{dom}(f) \to \mathbb{R}$ be twice differentiable with a critical point $x^\star$ (i.e., $\nabla f(x^\star) = 0$). Suppose there is a ball $X \subset \mathrm{dom}(f)$ with center $x^\star$ such that:

(i) Bounded inverse Hessians: $\exists \mu > 0$ such that

$$\|\nabla^2 f(x)^{-1}\| \leq 1/\mu, \quad \forall x \in X$$

(ii) Lipschitz continuous Hessians: $\exists B > 0$ such that

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq B\|x - y\|, \quad \forall x, y \in X$$

Then, $x_t \in X$ and x_{t+1} resulting from the Newton step satisfy

$$\|x_{t+1} - x^\star\| \leq \frac{B}{2\mu}\|x_t - x^\star\|^2$$

Local convergence (cont'd)

Example: Consider the nondegenerate quadratic function

$$f(x) = \frac{1}{2}x^\top Mx - q^\top x + c,$$

then property (i) holds with $\mu = 1/\|M^{-1}\|$ over $X = \mathbb{R}^n$. Property (ii) is satisfied for $B = 0$. Hence, the Theorem states that

$$\|x_1 - x^\star\| = 0,$$

that is, the Newton method reaches $x^\star$ after one iteration step.

Local convergence rate

Corollary: With the assumptions of the Theorem and if $x_0 \in X$ satisfies $\|x_0 - x^\star\| \le \mu/B$, the Newton method yields

$$\|x_T - x^\star\| \le \frac{\mu}{B}\left(\frac{1}{2}\right)^{2^T-1}, \quad T \ge 0$$

Hence, to achieve $\|x_T - x^\star\| \le \varepsilon$ we need $T = \mathcal{O}(\log\log(1/\varepsilon))$

Proof: We proceed via induction over T. For the induction base we start with $T = 1$ and recall that by the theorem

$$\|x_1 - x^\star\| \le \frac{B}{2\mu}\|x_0 - x^\star\|^2 \le \frac{B}{2\mu}\frac{\mu^2}{B^2} = \frac{\mu}{B}\frac{1}{2}.$$

We then show the induction step $T \to T+1$, again we use the theorem to state that

$$\|x_{t+1} - x^\star \le \frac{B}{2\mu}\|x_t - x^\star\|^2 \le \frac{B}{2\mu}\frac{\mu^2}{B^2}\left(\frac{1}{2}\right)^{2^{T+1}-2} = \frac{\mu}{B}\left(\frac{1}{2}\right)^{2^{T+1}-1},$$

where the second inequality is due to the induction hypothesis.

Quasi-Newton methods

Motivation:

- Computational bottleneck of the Newton method is the computation of the inverse of the Hessian
  $\Longrightarrow$ cost of $\mathcal{O}(n^3)$
- Can this costly step of the Hessian inversion be circumvented?

Secant method

- Focus on 1-dimensional setting to fix ideas, recall the Newton step

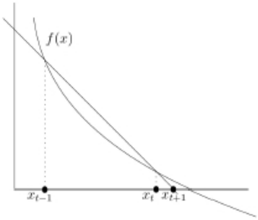
$$x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)}$$

- Using the finite difference approximation of a gradient

$$f'(x_t) \approx \frac{f(x_t) - f(x_{t-1})}{x_t - x_{t-1}}$$

- Secant step is gradient free

$$x_{t+1} = x_t - f(x_t)\frac{x_t - x_{t-1}}{f(x_t) - f(x_{t-1})},$$



The secant condition

- Applying the finite difference approximation to the second derivative of f

$$H_t = \frac{f'(x_t) - f'(x_{t+1})}{x_t - x_{t-1}} \approx f''(x_t)$$

- This implies the secant condition

$$f'(x_t) - f'(x_{t+1}) = H_t(x_t - x_{t-1}) \tag{1}$$

- The secant method works as

$$x_{t+1} = x_t - H_t^{-1} f'(x_t) \tag{2}$$

Whenever we use an update equation (2) with a symmetric matrix H_t satisfying (1), we say we have a Quasi-Newton method.

Quasi-Newton method in n dimensions

- Update scheme

$$x_{t+1} = x_t - Q_t^{-1} \nabla f(x_t)$$

- Symmetric matrix Q_t satisfy the secant condition

$$\nabla f(x_t) - \nabla f(x_{t+1}) = Q_t(x_t - x_{t-1})$$

Locally Hessians are close to constant

Lemma: With the assumptions and terminology of Theorem and if $x_0 \in X$ satisfies

$$\|x_0 - x^\star\| \leq \frac{\mu}{B},$$

then the Hessians in Newton's method satisfy the relative error bound

$$\frac{\|\nabla^2 f(x_t) - \nabla^2 f(x^\star)\|}{\|\nabla^2 f(x^\star)\|} \leq \left(\frac{1}{2}\right)^{2^t - 1}, \quad t \geq 0$$

The Lemma implies that for all $t \geq 0$

$$\|\nabla^2 f(x_t) - \nabla^2 f(x^\star)\| \leq \underbrace{\|\nabla^2 f(x^\star)\| \left(\frac{1}{2}\right)^{2^t - 1}}_{\approx 0 \text{ for } t \text{ large}}$$

$\implies \quad \nabla^2 f(x_t) \approx \nabla^2 f(x^\star)$ for t large

Proof of Lemma

For any two matrices $A, B \in \mathbb{R}^{n\times n}$, the inequality $\|AB\| \leq \|A\|\|B\|$ holds, since

$$\|AB\| = \max_{v\neq 0} \frac{\|ABv\|}{\|v\|} \leq \max_{v\neq 0} \frac{\|A\|\|Bv\|}{\|v\|} = \|A\|\|B\|.$$

Hence,

$$1 = \|\nabla^2 f(x^\star)\nabla^2 f(x^\star)^{-1}\| \leq \|\nabla^2 f(x^\star)\|\|\nabla^2 f(x^\star)^{-1}\| \leq \|\nabla^2 f(x^\star)\|\frac{1}{\mu},$$

which implies that $\|\nabla^2 f(x^\star)\| \geq \mu$. By the Lipschitz assumption and the Corollary from above

$$\|\nabla^2 f(x_T) - \nabla^2 f(x^\star)\| \leq B\|x_T - x^\star\| \leq \mu \left(\frac{1}{2}\right)^{2^T - 1}.$$

Together with $\|\nabla^2 f(x^\star)\| \geq \mu$, the statement follows.

Greenstadt's approach

- Since Hessians are close to constant (see previous Lemma) use $H_t \approx H_{t-1}$ and hence $H_t^{-1} \approx H_{t-1}^{-1}$
- Greenstadt Ansatz: $H_t^{-1} = H_{t-1}^{-1} + E_t$
- Error matrix E_t should be "small"
  $\implies$ $\|AE_tA^\top\|_F$ should be small for $A \in \mathbb{R}^{n\times n}$ invertible

- Fix t and denote
  - $H = H_{t-1}^{-1}$
  - $H' = H_t^{-1}$
  - $E = E_t$
  - $\sigma = x_t - x_{t-1}$
  - $y = \nabla f(x_t) - \nabla f(x_{t-1})$
  - $r = \sigma - Hy$

- Greenstadt Ansatz is

$$H' = H + E$$

- Secant condition $\nabla f(x_t) - \nabla f(x_{t-1}) = H_t(x_t - x_{t-1})$ becomes

$$H'y = \sigma \quad \iff \quad Ey = r$$

Greenstadt's approach (con'd) In

In summary we end up with the following convex optimization problem

$$(\star)\begin{cases} \min_{E\in\mathbb{R}^{n\times n}} & \|AEA^\top\|_F^2 \\ \text{s. t.} & Ey = r \\ & E^\top - E = 0 \end{cases}$$

- Lagrangian dual of $(\star)$ can be solved at low computational cost

Main key points

- **Definitions:** Newton method, Secant method, Quasi-Newton methods
- **Iteration complexity analysis for Newton method:**
    - For arbitrary initial condition: linear convergence rate $\mathcal{O}(\log(1/\varepsilon))$
    - For initial condition close to optimum: quadratic convergence rate $\mathcal{O}(\log\log(1/\varepsilon))$