Lecture 19: Basic Newton's Method

Yudong Chen

1 Second-Order Optimization

From now on, we will assume $\mathcal{X} = \mathbb{R}^d$ (unconstrained optimization) and $f : \mathbb{R}^d \to \mathbb{R}$ is *twice* continuously differentiable.

Second-order oracle model:

$$x \in \mathbb{R}^d \longrightarrow \boxed{\begin{array}{c}\text{2nd order}\\ \text{oracle}\end{array}} \longrightarrow f(x), \nabla f(x), \nabla^2 f(x).$$

Recall our general descent method:

$$x_{k+1} = x_k + \alpha_k p_k,$$

where α_k is the stepsize and p_k is a search direction. If p_k satisfies $\langle p_k, \nabla f(x_k)\rangle < 0$, then it is called a descent direction at x_k.

In this and subsequent lectures, we focus on search directions of the form

$$p_k = -B_k^{-1}\nabla f(x_k),$$

where $B_k \succ 0$. Examples:

- $B_k = I$: standard gradient descent, considered before;
- $B_k = \nabla^2 f(x_k)$: Newton's method;
- $B_k =$ some approximation of $\nabla^2 f(x_k)$: quasi-Newton's methods.

2 Basic Newton's Method

The basic Newton's (BN) method uses $B_k = \nabla^2 f(x_k)$ with a unit stepsize $\alpha_k = 1$, that is,

$$x_{k+1} = x_k - \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k). \tag{BN}$$

One can verify that

$$x_{k+1} = \operatorname*{argmin}_y \left\{ f(x_k) + \langle \nabla f(x_k), y - x_k\rangle + \frac{1}{2}\left\langle \nabla^2 f(x_k)(y - x_k), y - x_k\right\rangle \right\},$$

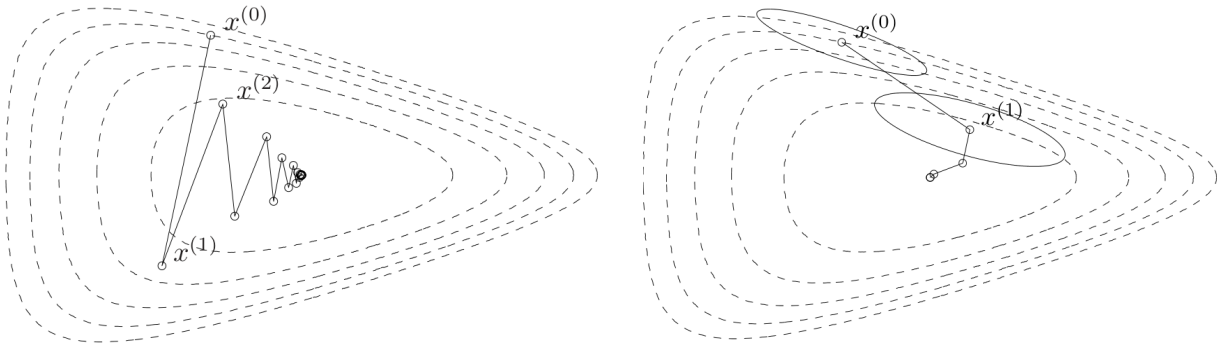
that is, x_{k+1} minimizes the second-order Taylor expansion of f at x_k. (Compare with GD.)

Here we assume that

- $\nabla^2 f(x_k)$ is invertible, so the iteration (BN) is well-defined;
- $\nabla^2 f(x_k) \succ 0$ is positive definite (p.d.), so $p_k = \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$ is a descent direction.

Later we will discuss how to handle situations where these assumptions are not satisfied.

Illustration of the steps taken by gradient descent (left) and Netwon's method (right):[1]



2.1 Terminology for rates of convergence

To discuss the convergence rate of (BN) and other descent methods, we need to introduce some terminology.

Let $\{x_k\}$ be a sequence in $\mathbb{R}^d$ that converges to some $x^* \in \mathbb{R}^d$. We say that the convergence is

1. *Q-linear* (or simply *linear*), if there exists $r \in (0,1)$ such that

$$\|x_{k+1} - x^*\|_2 \leq r \|x_k - x^*\|_2, \qquad \forall k \text{ sufficient large.}$$

   For example, the sequence $x_k = 0.5^k$ converges to 0 linearly. We have shown that when f is m-strongly convex and L-smooth, GD converges linearly with $r = \sqrt{1 - \frac{m}{L}} \approx 1 - \frac{m}{2L}$. Roughly speaking, linear convergence means that an ϵ accuracy can be achieves in $\log \frac{1}{\epsilon}$ iterations.

2. *Q-quadratic*, if there exists a constant $M > 0$ such that

$$\|x_{k+1} - x^*\|_2 \leq M \|x_k - x^*\|_2^2, \qquad \forall k \text{ sufficient large.}$$

   Note the square on the RHS. For example, the sequence $x_k = 0.5^{(2^k)}$ converges to 0 quadratically. Roughly speaking, quadratic convergence means that an ϵ accuracy can be achieved in $\log \log \frac{1}{\epsilon}$ iterations; put differently, the number of correct digits doubles at each iteration. Quadratic convergence is much faster than linear convergence. (A picture)

3. *Q-superlinear*, if for any constant $r > 0$, there exists $K_r < \infty$ such that

$$\|x_{k+1} - x^*\|_2 \leq r \|x_k - x^*\|_2, \qquad \forall k \geq K_r.$$

   This means $\{x_k\}$ converges faster than linear convergence (but not necessarily as fast as quadratic convergence).

[1] The figures are taken from Convex Optimization by Boyd and Vandenberghe.

2.2 Local quadratic convergence of Newton's method

Recall that the condition

$$\nabla f(x^*) = 0, \quad \nabla^2 f(x^*) \succ 0 \tag{1}$$

is a (2nd-order) sufficient condition for x^* being a local minimizer of f. Also recall that $\nabla^2 f(x)$ is said to be Lipschitz-continuous in some set $\mathcal{N}$ if there exists a constant $L_H < \infty$ such that

$$\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\|_2 \leq L_H \left\|x - y\right\|_2, \forall x, y \in \mathcal{N}.$$

The basic Newton's method converges quadratically in a neighborhood of such an x^*.

Theorem 1 (Theorem 3.5 in Nocedal-Wright)**.** *Suppose that f is twice continuously differentiable and that its Hessian is Lipschitz-continuous in a neighborhood of x^*, where x^* satisfies the 2nd-order sufficient condition (1). Let $\{x_k\}$ be given by (BN). Then*

(i) if the initial point x_0 is sufficiently close to x^*, then $\{x_k\}$ converges to x^*;

(ii) the rate of convergence of $\{x_k\}$ is quadratic;

(iii) the sequence of gradient norms $\{\|\nabla f(x_k)\|_2\}$ converges to zero, with a quadratic convergence rate.

Proof. It is clear that if $\|x_{k+1} - x^*\|_2 \leq M \|x_k - x^*\|_2^2$ holds for all k and x_0 is sufficiently close to x^* (e.g., $M \|x_0 - x^*\| < 1$), then we must have $\|x_k - x^*\|_2 \overset{k\to\infty}{\longrightarrow} 0$ and thus (i) holds.

It remains to show that when $\|x_0 - x^*\|_2$ is sufficiently small, then

1) (quadratic convergence of iterates) there exists a constant $M > 0$ such that $\forall k : \|x_{k+1} - x^*\|_2 \leq M \|x_k - x^*\|_2^2$, and

2) (quadratic convergence of gradients) there exists a constant $M' > 0$ such that $\forall k : \|\nabla f(x_{k+1})\|_2 \leq M' \|\nabla f(x_k)\|_2^2$.

Proof of 1): Suppose that x_k is in a neighborhood of x^* where $\nabla^2 f$ is L_H-Lipschitz continuous. Recall $x_{k+1} = x_k - \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$. Then

$$\begin{aligned}
\|x_{k+1} - x^*\|_2 &= \left\|x_k - x^* - \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)\right\|_2 \\
&= \left\|\left(\nabla^2 f(x_k)\right)^{-1} \left[\nabla^2 f(x_k)\left(x_k - x^*\right) - \nabla f(x_k)\right]\right\|_2 \\
&\leq \left\|\left(\nabla^2 f(x_k)\right)^{-1}\right\|_2 \left\|\nabla^2 f(x_k)\left(x_k - x^*\right) - \left(\nabla f(x_k) - \nabla f(x^*)\right)\right\|_2. \quad \text{b/c } \nabla f(x^*) = 0
\end{aligned}$$

We know from Taylor's Theorem that

$$\nabla f(x^*) - \nabla f(x_k) = \int_0^1 \nabla^2 f\left(x_k + t(x^* - x_k)\right)\left(x^* - x_k\right) \mathrm{d}t.$$

It follows that

$$\begin{aligned}
&\|x_{k+1} - x^*\|_2 \\
&\leq \left\|\left(\nabla^2 f(x_k)\right)^{-1}\right\|_2 \left\|\int_0^1 \left[\nabla^2 f(x_k) - \nabla^2 f\left(x_k + t(x^* - x_k)\right)\right](x_k - x^*)\,\mathrm{d}t\right\|_2 \\
&\leq \left\|\left(\nabla^2 f(x_k)\right)^{-1}\right\|_2 \int_0^1 \left\|\left[\nabla^2 f(x_k) - \nabla^2 f\left(x_k + t(x^* - x_k)\right)\right](x_k - x^*)\right\|_2 \mathrm{d}t \quad \text{Jensen} \\
&\leq \left\|\left(\nabla^2 f(x_k)\right)^{-1}\right\|_2 \int_0^1 \underbrace{\left\|\nabla^2 f(x_k) - \nabla^2 f\left(x_k + t(x^* - x_k)\right)\right\|_2}_{\leq L_H t\|x_k - x^*\|_2} \|x_k - x^*\|_2 \,\mathrm{d}t \quad \text{Cauchy-Schwarz} \\
&\leq \frac{L_H}{2}\left\|\left(\nabla^2 f(x_k)\right)^{-1}\right\|_2 \|x_k - x^*\|_2^2 .
\end{aligned}$$

Since $\nabla^2 f(x^*)$ is invertible and $\nabla^2 f$ is Lipschitz-continuous in a neighborhood of x^*, there exists some $r > 0$ such that

$$\left\|\left(\nabla^2 f(x_k)\right)^{-1}\right\|_2 \leq 2\left\|\left(\nabla^2 f(x^*)\right)^{-1}\right\|_2 \qquad \forall x_k : \|x_k - x^*\| \leq r. \tag{2}$$

Hence

$$\|x_{k+1} - x^*\|_2 \leq M\|x_k - x^*\|_2^2$$

for $M = L_H\left\|\left(\nabla^2 f(x^*)\right)^{-1}\right\|_2$.

Proof of 2): From $x_{k+1} = x_k - \left(\nabla^2 f(x_k)\right)^{-1}\nabla f(x_k)$, we can write

$$\nabla f(x_k) = -\nabla^2 f(x_k)\,(x_{k+1} - x_k)\,. \tag{3}$$

Thence

$$\begin{aligned}
\|\nabla f(x_{k+1})\|_2 &= \|\nabla f(x_{k+1}) - \nabla f(x_k) + \nabla f(x_k)\|_2 \\
&= \left\|\int_0^1 \left[\nabla^2 f\left(x_k + t(x_{k+1} - x_k)\right) - \nabla^2 f(x_k)\right](x_{k+1} - x_k)\,\mathrm{d}t\right\|_2 \quad \text{Taylor and (3)} \\
&\leq \int_0^1 \underbrace{\left\|\nabla^2 f\left(x_k + t(x_{k+1} - x_k)\right) - \nabla^2 f(x_k)\right\|_2}_{\leq L_H t\|x_{k+1} - x_k\|_2} \|x_{k+1} - x_k\|_2 \,\mathrm{d}t \quad \text{Jensen's, Cauchy-Schwarz} \\
&\leq \frac{L_H}{2}\|x_{k+1} - x_k\|_2^2 \\
&= \frac{L_H}{2}\left\|\left(\nabla^2 f(x_k)\right)^{-1}\nabla f(x_k)\right\|_2^2 \\
&\leq \frac{L_H}{2}\cdot \underbrace{\left\|\left(\nabla^2 f(x_k)\right)^{-1}\right\|_2^2}_{\leq 4\left\|\left(\nabla^2 f(x^*)\right)^{-1}\right\|_2^2 \text{ by (2)}} \cdot \|\nabla f(x_k)\|_2^2 \\
&\leq M'\|\nabla f(x_k)\|_2^2,
\end{aligned}$$

where $M' = 2L_H\left\|\left(\nabla^2 f(x_k)\right)^{-1}\right\|_2^2$. □

Remark 1. If $f(x) = \frac{1}{2}x^\top Ax - b^\top x$ is a convex quadratic function, the Hessian $\nabla^2 f(x) = A$ is independent of x and hence $\nabla^2 f$ is L_H-Lipschitz continuous on $\mathbb{R}^d$ with $L_H = 0$. In this case, Theorem 1 implies that (BN) converges to a global minimizer x^* in one iteration. Of course, one can prove this result directly by noting that $x_1 = x_0 - A^{-1}(Ax_0 - b) = A^{-1}b = x^*$.

3 Additional remarks (optional)

3.1 Affine invariance

A nice feature of Newton's method is that it is invariant to linear or affine transformations (i.e., changes of coordinates), in the follow sense. Let $\{x_k\}$ be the iterates of (BN) applied to the function $f : \mathbb{R}^d \to \mathbb{R}$. Suppose $T \in \mathbb{R}^{d\times d}$ is a nonsingular matrix. Define a new function $g : \mathbb{R}^d \to \mathbb{R}$ by $g(y) = f(Ty)$. If we apply (BN) to minimize g starting from $y_0 = T^{-1}x_0$, then

$$y_k = T^{-1}x_k, \qquad \forall k.$$

(Proof uses the chain rules $\nabla g(y) = T^\top \nabla f(Ty)$ and $\nabla^2 g(y) = T^\top \nabla^2 f(Ty)T$; left as exercise.) That is, the iterates are related by the same linear transformation. In contrast, gradient descent lacks this property and is very sensitive to changes of coordinates (which strongly affect, e.g., the condition number).

However, the convergence analysis of (BN) in Theorem 1 is *not* affine invariant: it depends very much on the choice of coordinates. If we change the coordinate system, the values of L_H, M and M' all change. There is an elegant way of obtaining affine invariant convergence results, which is based on the notion of self-concordant functions.

3.2 Performance

Newton's method converges very fast near x^*. If x_0 is sufficiently close to x^* such that the quadratic convergence holds, usually at most six or iterations suffice for achieving a very high accuracy.

The main drawback of Netwon's method is the high cost of computing storing the $d \times d$ Hessian matrix $\nabla^2 f(x)$, especially when d is large. There are several ways for reducing the computational cost, including various inexact Newton's methods and quasi-Newton's methods—we will discuss some of them later. However, these methods are still more computationally intensive in general than first-order methods.

3.3 Global convergence?

Theorem 1 is a *local* convergence result: it holds when x_0 is sufficiently close to x^*. If x_0 is far from x^*, the basic Newton's method (BN) may not converge to a stationary point. Additional adjustmen to (BN) is needed to ensure global convergence. We will discuss some of them in the next lecture.