4.4.2 The Newton Method

Example 4.19 Let us apply the Newton method to find the root of the following function

$$g(x) = \frac{x}{\sqrt{1+x^2}}.$$

Clearly $x^* = 0$.

The Newton method will give:

$$x_{k+1} = x_k - \frac{g(x_k)}{g'(x_k)} = x_k - x_k(1+x_k^2) = -x_k^3.$$

Therefore, the method converges if $|x_0| < 1$, it oscillates if $|x_0| = 1$, and finally, diverges if $|x_0| > 1$.

Assumption 4.20

1. $f \in \mathcal{C}_M^{2,2}(\mathbb{R}^n)$;
2. There is a local minimum $\boldsymbol{x}^*$ of the function $f(\boldsymbol{x})$;
3. The Hessian is positive definite at $\boldsymbol{x}^*$:

$$\boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}^*) \succeq \ell \boldsymbol{I}, \quad \ell > 0;$$

4. Our starting point $\boldsymbol{x}_0$ is close enough to $\boldsymbol{x}^*$.

Theorem 4.21 Let the function $f(\boldsymbol{x})$ satisfy the above assumptions. Suppose that the initial starting point $\boldsymbol{x}_0$ is close enough to $\boldsymbol{x}^*$:

$$\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2 < \bar{r} := \frac{2\ell}{3M}.$$

Then $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2 < \bar{r}$ for all k of the Newton method and it converges (Q-)quadratically:

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|_2 \leq \frac{M\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2}{2(\ell - M\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2)}.$$

Proof:

Let $r_k = \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2$. From Lemma 3.8 and the assumption, we have for $k = 0$,

$$\boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}_0) \succeq \boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}^*) - M r_0 \boldsymbol{I} \succeq (\ell - M r_0)\boldsymbol{I}. \tag{8}$$

Since $r_0 < \bar{r} = \frac{2\ell}{3M} < \frac{\ell}{M}$, we have $\ell - Mr_0 > 0$ and therefore, $\boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}_0)$ is invertible.

Consider the Newton method for $k = 0$, $\boldsymbol{x}_1 = \boldsymbol{x}_0 - [\boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}_0)]^{-1} \boldsymbol{\nabla} \boldsymbol{f}(\boldsymbol{x}_0)$.

Then

$$\begin{aligned}
\boldsymbol{x}_1 - \boldsymbol{x}^* &= \boldsymbol{x}_0 - \boldsymbol{x}^* - [\boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}_0)]^{-1} \boldsymbol{\nabla} \boldsymbol{f}(\boldsymbol{x}_0) \\
&= \boldsymbol{x}_0 - \boldsymbol{x}^* - [\boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}_0)]^{-1} \int_0^1 \boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}^* + \tau(\boldsymbol{x}_0 - \boldsymbol{x}^*))(\boldsymbol{x}_0 - \boldsymbol{x}^*) d\tau \\
&= [\boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}_0)]^{-1} \boldsymbol{G}_0 (\boldsymbol{x}_0 - \boldsymbol{x}^*)
\end{aligned}$$

where $\boldsymbol{G}_0 = \int_0^1 [\boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}_0) - \boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}^* + \tau(\boldsymbol{x}_0 - \boldsymbol{x}^*))] d\tau$.

Then

$$\begin{aligned}
\|\boldsymbol{G}_0\|_2 &= \left\| \int_0^1 [\boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}_0) - \boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}^* + \tau(\boldsymbol{x}_0 - \boldsymbol{x}^*))] d\tau \right\|_2 \\
&\leq \int_0^1 \|\boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}_0) - \boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}^* + \tau(\boldsymbol{x}_0 - \boldsymbol{x}^*))\|_2 d\tau \\
&\leq \int_0^1 M|1-\tau| r_0 d\tau = \frac{r_0}{2} M.
\end{aligned}$$

From (8),

$$\|[\boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}_0)]^{-1}\|_2 \leq (\ell - M r_0)^{-1}.$$

Then

$$r_1 \leq \frac{M r_0^2}{2(\ell - M r_0)}.$$

Since $r_0 < \bar{r} = \frac{2\ell}{3M}$, $\frac{M r_0}{2(\ell - M r_0)} < 1$, and $r_1 < r_0$.

One can see now that the same argument is valid for all k's. ∎

- Comparing this result with the rate of convergence of the steepest descent, we see that the Newton method is much faster.
- Surprisingly, the region of *quadratic convergence* of the Newton method is almost the same as the region of the *linear convergence* of the gradient method.

$$\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2 < \frac{2\ell}{M} \quad \text{(steepest descent method)} \quad \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2 < \frac{2\ell}{3M} \quad \text{(Newton method)}$$

- This justifies a standard recommendation to use the steepest descent method only at the initial stage of the minimization process in order to get close to a local minimum and then perform the Newton method to refine.

4.4.3 The Conjugate Gradient Methods

The conjugate gradient methods were initially proposed for minimizing convex quadratic functions. Consider the problem

$$\boxed{\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x})}$$

with $f(\boldsymbol{x}) = \alpha + \langle \boldsymbol{a}, \boldsymbol{x} \rangle + \frac{1}{2}\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} \rangle$ and $\boldsymbol{A} \succ \boldsymbol{O}$. Since its minimal solution is $\boldsymbol{x}^* = -\boldsymbol{A}^{-1}\boldsymbol{a}$, we can rewrite $f(\boldsymbol{x})$ as:

$$\begin{aligned}
f(\boldsymbol{x}) &= \alpha - \langle \boldsymbol{A}\boldsymbol{x}^*, \boldsymbol{x} \rangle + \frac{1}{2}\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} \rangle \\
&= \alpha - \frac{1}{2}\langle \boldsymbol{A}\boldsymbol{x}^*, \boldsymbol{x}^* \rangle + \frac{1}{2}\langle \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^* \rangle.
\end{aligned}$$

Thus, $f(\boldsymbol{x}^*) = \alpha - \frac{1}{2}\langle \boldsymbol{A}\boldsymbol{x}^*, \boldsymbol{x}^* \rangle$ and $\boldsymbol{\nabla} \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}^*)$.

Definition 4.22 Given a starting point $\boldsymbol{x}_0$, the linear *Krylov subspaces* is defined as

$$\mathcal{L}_k := \text{span}\{\boldsymbol{A}(\boldsymbol{x}_0 - \boldsymbol{x}^*), \ldots, \boldsymbol{A}^k(\boldsymbol{x}_0 - \boldsymbol{x}^*)\}, \quad k \geq 1,$$

where $\text{span}\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_p\}$ is the linear subspace of $\mathbb{R}^n$ spanned by the vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_p \in \mathbb{R}^n$.

We claim temporarily that the sequence of points generated by a *conjugate gradient method* is defined as follows:

$$\boxed{\boldsymbol{x}_k := \arg\min\{f(\boldsymbol{x}) \mid \boldsymbol{x} \in \boldsymbol{x}_0 + \mathcal{L}_k\},\ k \geq 1.}$$

Lemma 4.23 For any $k \geq 1$, $\mathcal{L}_k = \text{span}\{\boldsymbol{\nabla f}(\boldsymbol{x}_0), \ldots, \boldsymbol{\nabla f}(\boldsymbol{x}_{k-1})\}$.

Proof:

Let us prove by induction hypothesis.

For $k = 1$, the statement is true since $\boldsymbol{\nabla f}(\boldsymbol{x}_0) = \boldsymbol{A}(\boldsymbol{x}_0 - \boldsymbol{x}^*)$.

Suppose the claim is true for some $k \geq 1$. Then from the definition of the conjugate gradient method,

$$\boldsymbol{x}_k = \boldsymbol{x}_0 + \sum_{i=1}^{k} \lambda_i \boldsymbol{A}^i(\boldsymbol{x}_0 - \boldsymbol{x}^*)$$

with some $\lambda_i \in \mathbb{R}, \quad i = 1, \ldots, k$. Therefore,

$$\boldsymbol{\nabla f}(\boldsymbol{x}_k) = \boldsymbol{A}(\boldsymbol{x}_0 - \boldsymbol{x}^*) + \sum_{i=1}^{k} \lambda_i \boldsymbol{A}^{i+1}(\boldsymbol{x}_0 - \boldsymbol{x}^*) = \boldsymbol{A}(\boldsymbol{x}_0 - \boldsymbol{x}^*) + \sum_{i=1}^{k-1} \lambda_i \boldsymbol{A}^{i+1}(\boldsymbol{x}_0 - \boldsymbol{x}^*) + \lambda_k \boldsymbol{A}^{k+1}(\boldsymbol{x}_0 - \boldsymbol{x}^*).$$

The first two terms of the last expression belongs to $\mathcal{L}_k$ from the definition. And then,

$$\text{span}\{\mathcal{L}_k, \boldsymbol{\nabla f}(\boldsymbol{x}_k)\} \subseteq \text{span}\{\mathcal{L}_k, \boldsymbol{A}^{k+1}(\boldsymbol{x}_0 - \boldsymbol{x}^*)\} = \mathcal{L}_{k+1}.$$

There are two ways to show that the equality holds. Assume that $\boldsymbol{A}^{k+1}(\boldsymbol{x}_0 - \boldsymbol{x}^*) \in \mathcal{L}_k$. Then it is obvious and $\mathcal{L}_k = \mathcal{L}_{k+1}$. If $\boldsymbol{A}^{k+1}(\boldsymbol{x}_0 - \boldsymbol{x}^*) \notin \mathcal{L}_k$, the equality holds unless $\lambda_k = 0$. However, this possibility implies that $\boldsymbol{x}_k \in \mathcal{L}_{k-1}$, $\boldsymbol{x}_{k-1} = \boldsymbol{x}_k$ and therefore, $\mathcal{L}_{k-1} = \mathcal{L}_k = \mathcal{L}_{k+1}$ again.

An alternative way is to use contradiction. If the equality does not hold, $\boldsymbol{\nabla f}(\boldsymbol{x}_k) \in \mathcal{L}_k$ implies $\boldsymbol{A}^{k+1}(\boldsymbol{x}_0 - \boldsymbol{x}^*) \in \mathcal{L}_k$, which again implies the equality, or $\lambda_k = 0$, which implies that $\boldsymbol{x}_k = \boldsymbol{x}_{k-1}$ (algorithm terminated). ∎

Lemma 4.24 For any $k, \ell \geq 0,\ k \neq \ell$, we have $\langle \boldsymbol{\nabla f}(\boldsymbol{x}_k), \boldsymbol{\nabla f}(\boldsymbol{x}_\ell) \rangle = 0$.

Proof:

Let $k \geq i$, and consider

$$\phi(\boldsymbol{\lambda}) = f\left(\boldsymbol{x}_0 + \sum_{j=1}^{k} \lambda_j \boldsymbol{\nabla f}(\boldsymbol{x}_{j-1})\right).$$

From the previous lemma, there is a $\boldsymbol{\lambda}^*$ such that $\boldsymbol{x}_k = \boldsymbol{x}_0 + \sum_{j=1}^{k} \lambda_j^* \boldsymbol{\nabla f}(\boldsymbol{x}_{j-1})$. Moreover, $\boldsymbol{\lambda}^*$ is the minimum of the function $\phi(\boldsymbol{\lambda})$. Therefore,

$$\frac{\partial \phi}{\partial \lambda_i}(\boldsymbol{\lambda}^*) = \langle \boldsymbol{\nabla f}(\boldsymbol{x}_k), \boldsymbol{\nabla f}(\boldsymbol{x}_{i-1}) \rangle = 0.$$

∎

Corollary 4.25 The sequence generated by the conjugate gradient method for the convex quadratic function is finite.

Proof:

Since the number of orthogonal directions in $\mathbb{R}^n$ cannot exceed n. ∎

Let us define $\boldsymbol{\delta}_i = \boldsymbol{x}_{i+1} - \boldsymbol{x}_i$. It is clear that $\mathcal{L}_k = \text{span}\{\boldsymbol{\delta}_0, \boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{k-1}\}$ (Exercise 10).

Lemma 4.26 For any $k, \ell \geq 0$, $k \neq \ell$, $\langle \boldsymbol{A}\boldsymbol{\delta}_k, \boldsymbol{\delta}_\ell \rangle = 0$.

Proof:
Left for exercise. ∎

The vectors $\{\boldsymbol{\delta}_i\}$ are called *conjugate* with respect to matrix $\boldsymbol{A}$.

Now, let us be more precise with the conjugate gradient method. We will define the next iterations as follows:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - h_k \boldsymbol{\nabla f}(\boldsymbol{x}_k) + \sum_{j=0}^{k-1} \lambda_j \boldsymbol{\delta}_j$$

Using the previous properties, we arrive that (see Exercise 11)

$$\lambda_j = 0, \quad (j = 0, 1, \ldots, k-2), \quad \lambda_{k-1} = \frac{h_k \|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2}{\langle \boldsymbol{\nabla f}(\boldsymbol{x}_k) - \boldsymbol{\nabla f}(\boldsymbol{x}_{k-1}), \boldsymbol{\delta}_{k-1}\rangle}. \tag{9}$$

Thus

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - h_k \boldsymbol{p}_k$$

where

$$\boldsymbol{p}_k = \boldsymbol{\nabla f}(\boldsymbol{x}_k) - \frac{\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2 \boldsymbol{p}_{k-1}}{\langle \boldsymbol{\nabla f}(\boldsymbol{x}_k) - \boldsymbol{\nabla f}(\boldsymbol{x}_{k-1}), \boldsymbol{p}_{k-1}\rangle}.$$

Finally, we can present the Conjugate Gradient Method

Conjugate Gradient Method	
Step 0:	Let $\boldsymbol{x}_0 \in \mathbb{R}^n$, compute $f(\boldsymbol{x}_0)$, $\boldsymbol{\nabla f}(\boldsymbol{x}_0)$ and set $\boldsymbol{p}_0 := \boldsymbol{\nabla f}(\boldsymbol{x}_0)$, $k := 0$
Step 1:	Find $\boldsymbol{x}_{k+1} := \boldsymbol{x}_k - h_k \boldsymbol{p}_k$ by "approximate line search" on the scalar h_k
Step 2:	Compute $f(\boldsymbol{x}_{k+1})$ and $\boldsymbol{\nabla f}(\boldsymbol{x}_{k+1})$
Step 3:	Compute the coefficient β_{k+1}
Step 4:	Set $p_{k+1} := \boldsymbol{\nabla f}(\boldsymbol{x}_{k+1}) - \beta_{k+1}\boldsymbol{p}_k$, $k := k+1$ and go to Step 1

The most popular choices for the coefficient β_k are:

1. *Hestenes-Stiefel (1952):* $\beta_{k+1} = \frac{\langle \boldsymbol{\nabla f}(\boldsymbol{x}_{k+1}), \boldsymbol{\nabla f}(\boldsymbol{x}_{k+1}) - \boldsymbol{\nabla f}(\boldsymbol{x}_k)\rangle}{\langle \boldsymbol{\nabla f}(\boldsymbol{x}_{k+1}) - \boldsymbol{\nabla f}(\boldsymbol{x}_k), \boldsymbol{p}_k\rangle}$.
2. *Fletcher-Reeves (1964):* $\beta_{k+1} = \frac{\|\boldsymbol{\nabla f}(\boldsymbol{x}_{k+1})\|_2^2}{\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2}$.
3. *Polak-Ribière (1969):* $\beta_{k+1} = \frac{\langle \boldsymbol{\nabla f}(\boldsymbol{x}_{k+1}), \boldsymbol{\nabla f}(\boldsymbol{x}_{k+1}) - \boldsymbol{\nabla f}(\boldsymbol{x}_k)\rangle}{\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2}$.
4. *Polak-Ribière plus:* $\beta_{k+1} = \max\left\{0, \frac{\langle \boldsymbol{\nabla f}(\boldsymbol{x}_{k+1}), \boldsymbol{\nabla f}(\boldsymbol{x}_{k+1}) - \boldsymbol{\nabla f}(\boldsymbol{x}_k)\rangle}{\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2}\right\}$.
5. *Dai-Yuan (1999):* $\beta_{k+1} = \frac{\|\boldsymbol{\nabla f}(\boldsymbol{x}_{k+1})\|_2^2}{\langle \boldsymbol{\nabla f}(\boldsymbol{x}_{k+1}) - \boldsymbol{\nabla f}(\boldsymbol{x}_k), \boldsymbol{p}_k\rangle}$.

Among them, Hestenes-Stiefel and Polak-Ribière are empirically preferred.