Then

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

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

Then

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

From (8),

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

Then

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

Since $r_0 < \bar{r} = \frac{2\ell}{3M}$, $\quad \frac{M r_0}{2(\ell - M r_0)} < 1$, 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 $\boldsymbol{\nabla f}(\boldsymbol{x}) = \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}^*)$.

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.