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.

5.5 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 5.16 Given a starting point $\boldsymbol{x}_0$, the linear *Krylov subspaces* is defined as

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

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

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,

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

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 5.18 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 5.19 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 $\mathcal{L}_k = \text{Lin}\{\boldsymbol{\delta}_0, \boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{k-1}\}$ (Exercise 6).

Lemma 5.20 For any $k, \ell \geq 0$, $k \neq \ell$, $\langle \boldsymbol{A\delta}_k, \boldsymbol{\delta}_\ell \rangle = 0$.

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:

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

Using the previous properties, we arrive that (see Exercise 7)

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

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.

5.6 Quasi-Newton Methods

The basic idea of quasi-Newton methods is to approximate the Hessian matrix (or its inverse) which we need to compute in the Newton method. There are of course infinitely many ways to do so, but we choose the ones which satisfy the *secant equation*:

$$\boldsymbol{H}_{k+1}\boldsymbol{y}_k = \boldsymbol{s}_k$$

where $\boldsymbol{y}_k = \boldsymbol{\nabla f}(\boldsymbol{x}_{k+1}) - \boldsymbol{\nabla f}(\boldsymbol{x}_k)$, $\boldsymbol{s}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k$.

The general scheme of the quasi-Newton method is as follows.

Quasi-Newton Method	
Step 0:	Let $\boldsymbol{x}_0 \in \mathbb{R}^n$, $\boldsymbol{H}_0 := \boldsymbol{I}$, $k := 0$. Compute $f(\boldsymbol{x}_0), \boldsymbol{\nabla f}(\boldsymbol{x}_0)$
Step 1:	Set $\boldsymbol{p}_k := \boldsymbol{H}_k\boldsymbol{\nabla f}(\boldsymbol{x}_k)$
Step 2:	Find $\boldsymbol{x}_{k+1} := \boldsymbol{x}_k - h_k\boldsymbol{p}_k$ by "approximate line search" on the scalar h_k
Step 3:	Compute $f(\boldsymbol{x}_{k+1})$ and $\boldsymbol{\nabla f}(\boldsymbol{x}_{k+1})$
Step 4:	Compute $\boldsymbol{H}_{k+1}$ from $\boldsymbol{H}_k$, $k := k+1$ and go to Step 1

The most popular updates for $\boldsymbol{H}_{k+1}$ are:

1. *BFGS (Broyden-Fletcher-Goldfarb-Shanno)*

$$\boldsymbol{H}_{k+1} := \left(\boldsymbol{I} - \frac{\boldsymbol{s}_k(\boldsymbol{y}_k)^T}{\langle \boldsymbol{s}_k, \boldsymbol{y}_k\rangle}\right)\boldsymbol{H}_k\left(\boldsymbol{I} - \frac{\boldsymbol{y}_k(\boldsymbol{s}_k)^T}{\langle \boldsymbol{s}_k, \boldsymbol{y}_k\rangle}\right) + \frac{\boldsymbol{s}_k(\boldsymbol{s}_k)^T}{\langle \boldsymbol{s}_k, \boldsymbol{y}_k\rangle}$$

2. *DFP (Davidon-Fletcher-Powell)*

$$\boldsymbol{H}_{k+1} := \boldsymbol{H}_k + \frac{\boldsymbol{s}_k(\boldsymbol{s}_k)^T}{\langle \boldsymbol{y}_k, \boldsymbol{s}_k\rangle} - \frac{\boldsymbol{H}_k\boldsymbol{y}_k(\boldsymbol{y}_k)^T\boldsymbol{H}_k}{\langle \boldsymbol{y}_k, \boldsymbol{H}_k\boldsymbol{y}_k\rangle}$$

3. *Symmetric-Rank-One*

$$\boldsymbol{H}_{k+1} := \boldsymbol{H}_k + \frac{(\boldsymbol{s}_k - \boldsymbol{H}_k\boldsymbol{y}_k)(\boldsymbol{s}_k - \boldsymbol{H}_k\boldsymbol{y}_k)^T}{\langle \boldsymbol{s}_k - \boldsymbol{H}_k\boldsymbol{y}_k, \boldsymbol{y}_k\rangle}$$

In the same way for the conjugate gradient method, we can show that the quasi-Newton method converges in finite number of iterations for a strictly convex quadratic function. Moreover, under some strict convexity conditions at the neighborhood of the local minimum, it is possible to show that its iterates converge super-linearly [Nocedal].

5.7 Exercises

1. Give a geometric interpretation of the following step-size strategies:

   Let $0 < c_1 < c_2 < 1$,

   - Wolfe condition

$$\begin{aligned} f(\boldsymbol{x}_k - h\boldsymbol{\nabla f}(\boldsymbol{x}_k)) &\leq f(\boldsymbol{x}_k) - c_1 h\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2, \\ \langle \boldsymbol{\nabla f}(\boldsymbol{x}_k - h\boldsymbol{\nabla f}(\boldsymbol{x}_k)), \boldsymbol{\nabla f}(\boldsymbol{x}_k)\rangle &\leq c_2\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2. \end{aligned}$$

   - Strong Wolfe condition

$$\begin{aligned} f(\boldsymbol{x}_k - h\boldsymbol{\nabla f}(\boldsymbol{x}_k)) &\leq f(\boldsymbol{x}_k) - c_1 h\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2, \\ |\langle \boldsymbol{\nabla f}(\boldsymbol{x}_k - h\boldsymbol{\nabla f}(\boldsymbol{x}_k)), \boldsymbol{\nabla f}(\boldsymbol{x}_k)\rangle| &\leq c_2\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2. \end{aligned}$$

2. Consider a sequence $\{\beta_k\}_{k=0}^{\infty}$ which converges to zero.

   The sequence is said to converge *Q-linearly* if there exists a scalar $\rho \in (0,1)$ such that

$$\left|\frac{\beta_{k+1}}{\beta_k}\right| \leq \rho,$$

   for all k sufficiently large. *Q-superlinear* convergence occurs when we have

$$\lim_{k\to\infty} \frac{\beta_{k+1}}{\beta_k} = 0,$$

   while the convergence is *Q-quadratic* if there is a constant C such that

$$\frac{|\beta_{k+1}|}{\beta_k^2} \leq C$$

   for all k sufficiently large. *Q-superquadratic* convergence is indicated by

$$\lim_{k\to\infty} \frac{\beta_{k+1}}{\beta_k^2} = 0.$$

   (a) Show that the following implications are valid: Q-superquadratic ⇒ Q-quadratic ⇒ Q-superlinear ⇒ Q-linear.

   (b) Give examples of sequences which do not imply the opposite directions in the three cases above.

   A zero converging sequence $\{\beta_k\}_{k=0}^{\infty}$ is said to converge *R-linearly* if it is dominated by a Q-linearly converging sequence. That is, if there is a Q-linearly converging sequence $\{\hat{\beta}_k\}_{k=0}^{\infty}$ such that $0 \leq |\beta_k| \leq \hat{\beta}_k$.

   (c) Give a sequence which is R-linearly converging but not Q-linearly converging.

3. Let $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x}$ such that $\boldsymbol{Q}$ is symmetric, and indefinite. Apply the steepest descent method with constant step. Show that if the starting point $\boldsymbol{x}_0$ belongs to the space spanned by the negative eigenvectors, the sequence generated by the steepest descent method diverges.

4. Prove Lemma 5.20.

5. In light of Theorem 5.15, show that under Assumption 5.14, if we want to obtain $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2 < \varepsilon$, we need an order of $\ln(\ln \varepsilon^{-1})$ iterations for the Newton method.

6. In the Section 5.5, show that $\mathcal{L}_k = \{\boldsymbol{\delta}_0, \boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_{k-1}\}$.

7. In the same section, arrive at the expression (9) for a strictly convex quadratic function.

8. Show that the secant equation is valid for BFGS, DFP and symmetric-rank-one formulae.

9. Given $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ and a non-singular matrix $\boldsymbol{M} \in \mathbb{R}^{n\times n}$, if $1+\boldsymbol{v}^T\boldsymbol{M}^{-1}\boldsymbol{u} \neq 0$, then the following formula is valid:

$$(\boldsymbol{M} + \boldsymbol{u}\boldsymbol{v}^T)^{-1} = \boldsymbol{M}^{-1} - \frac{\boldsymbol{M}^{-1}\boldsymbol{u}\boldsymbol{v}^T\boldsymbol{M}^{-1}}{1+\boldsymbol{v}^T\boldsymbol{M}^{-1}\boldsymbol{u}}. \quad \text{(Sherman-Morrison formula)}$$

   Apply this formula to compute the inverses $\boldsymbol{B}_{k+1}$ of $\boldsymbol{H}_{k+1}$ for BFGS, DFP and symmetric-rank-one formulae.

10. Apply the quasi-Newton method with BFGS, DFP, and Symmetric-Rank-One updates for the strictly convex function $f(\boldsymbol{x}) = \alpha + \langle \boldsymbol{a}, \boldsymbol{x}\rangle + \frac{1}{2}\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}\rangle$ with $\boldsymbol{A} \succ \boldsymbol{O}$.