Let us evaluate the result of one step of the gradient method.

Consider $\boldsymbol{y} = \boldsymbol{x} - hf'(\boldsymbol{x})$. From Lemma 1.4.4,

$$\begin{aligned} f(\boldsymbol{y}) &\leq f(\boldsymbol{x}) + \langle f'(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 \\ &= f(\boldsymbol{x}) - h\|f'(\boldsymbol{x})\|^2 + \frac{h^2 L}{2}\|f'(\boldsymbol{x})\|^2 \\ &= f(\boldsymbol{x}) - h\left(1 - \frac{h}{2}L\right)\|f'(\boldsymbol{x})\|^2. \end{aligned} \tag{1.3}$$

Thus, one step of the gradient method decreases the value of the objective function at least as follows for $h^* = 1/L$.

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) - \frac{1}{2L}\|f'(\boldsymbol{x})\|^2.$$

Now, since $f(\boldsymbol{x}_{k+1}) = f(\boldsymbol{x}_k - h_k f'(\boldsymbol{x}_k))$, consider the **Goldstein-Armijo Rule** previously described.

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \leq \beta h_k \|f'(\boldsymbol{x}_k)\|^2,$$

and from (1.3)

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq h_k\left(1 - \frac{h_k}{2}L\right)\|f'(\boldsymbol{x}_k)\|^2.$$

Therefore, $h_k \geq 2(1-\beta)/L$.

Also, substituting in

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq \alpha h_k \|f'(\boldsymbol{x}_k)\|^2 \geq \frac{2}{L}\alpha(1-\beta)\|f'(\boldsymbol{x}_k)\|^2.$$

Thus, in all three step-size strategies mentioned here, we can say that

$$\boxed{f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq \frac{\omega}{L}\|f'(\boldsymbol{x}_k)\|^2}$$

for some positive constant ω.

Summing up the above inequality we have:

$$\frac{\omega}{L}\sum_{k=0}^{N}\|f'(\boldsymbol{x}_k)\|^2 \leq f(\boldsymbol{x}_0) - f(\boldsymbol{x}_{N+1}) \leq f(\boldsymbol{x}_0) - f^*$$

where f^* is the optimal value of the problem.

As a simple consequence we have

$$\|f'(\boldsymbol{x}_k)\| \to 0 \quad \text{as} \quad k \to \infty.$$

Finally,

$$g_N^* \equiv \min_{0 \leq k \leq N}\|f'(\boldsymbol{x}_k)\| \leq \frac{1}{\sqrt{N+1}}\left[\frac{1}{\omega}L(f(\boldsymbol{x}_0) - f^*)\right]^{1/2}. \tag{1.4}$$

Remark 1.5.1 $g_N^* \to 0$, but we cannot say anything about the rate of convergence of the sequence $\{f(\boldsymbol{x}_k)\}$ or $\{\boldsymbol{x}_k\}$.

Example 1.5.2 Consider the function $f(x,y) = \frac{1}{2}x^2 + \frac{1}{4}y^4 - \frac{1}{2}y^2$. $(0,-1)^T$ and $(0,1)^T$ are local minimal solutions, but $(0,0)^T$ is a stationary point.

If we start the gradient method from $(1,0)^T$, we will only converge to the stationary point.

We focus now on the following problem class:

Model:	1. Unconstrained minimization
	2. $f \in C_L^{1,1}(\mathbb{R}^n)$
	3. $f(\boldsymbol{x})$ is bounded from below
Oracle:	First-order black box
ε**-solution:**	$f(\bar{\boldsymbol{x}}) \leq f(\boldsymbol{x}_0), \quad \|f'(\bar{\boldsymbol{x}})\| < \epsilon$

From (1.4), we have

$$g_N^* < \varepsilon \quad \text{if} \quad N+1 > \frac{L}{\omega\varepsilon^2}(f(\boldsymbol{x}_0) - f^*).$$

Remark 1.5.3 This is much better than the result of Theorem 1.2.3, since *it does not depend on n*.

Finally, consider the following problem under Assumption 1.5.4.

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

Assumption 1.5.4

1. $f \in C_M^{2,2}(\mathbb{R}^n)$.
2. There is a local minimum of the function $f(\boldsymbol{x})$ at which its Hessian is positive definite.
3. We know some bound $0 < \ell \leq L < \infty$ for the Hessian at $\boldsymbol{x}^*$:
$$\ell \boldsymbol{I} \preceq f''(\boldsymbol{x}^*) \preceq L\boldsymbol{I}.$$
4. Our starting point $\boldsymbol{x}_0$ is close enough to $\boldsymbol{x}^*$.

Theorem 1.5.5 Let $f(\boldsymbol{x})$ satisfy our assumptions above and let the starting point $\boldsymbol{x}_0$ be close enough to a local minimum:

$$r_0 = \|\boldsymbol{x}_0 - \boldsymbol{x}^*\| < \bar{r} = \frac{2\ell}{M}.$$

Then, the gradient method with step-size $h^* = 2/(L+\ell)$ converges as follows:

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\| \leq \frac{\bar{r}r_0}{\bar{r} - r_0}\left(1 - \frac{2\ell}{L+3\ell}\right)^k.$$

This rate of convergence is called (R-)*linear*.

Proof: In the gradient method, the iterates are $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - h_k f'(\boldsymbol{x}_k)$.
Since $f'(\boldsymbol{x}^*) = 0$,

$$f'(\boldsymbol{x}_k) = f'(\boldsymbol{x}_k) - f'(\boldsymbol{x}^*) = \int_0^1 f''(\boldsymbol{x}^* + \tau(\boldsymbol{x}_k - \boldsymbol{x}^*))(\boldsymbol{x}_k - \boldsymbol{x}^*)d\tau = \boldsymbol{G}_k(\boldsymbol{x}_k - \boldsymbol{x}^*),$$

and therefore,

$$\boldsymbol{x}_{k+1} - \boldsymbol{x}^* = \boldsymbol{x}_k - \boldsymbol{x}^* - h_k \boldsymbol{G}_k(\boldsymbol{x}_k - \boldsymbol{x}^*) = (\boldsymbol{I} - h_k \boldsymbol{G}_k)(\boldsymbol{x}_k - \boldsymbol{x}^*).$$

Let $r_k = \|\boldsymbol{x}_k - \boldsymbol{x}^*\|$. From Lemma 1.4.6,

$$f''(\boldsymbol{x}^*) - \tau M r_k \boldsymbol{I} \preceq f''(\boldsymbol{x}^* + \tau(\boldsymbol{x}_k - \boldsymbol{x}^*)) \preceq f''(\boldsymbol{x}^*) + \tau M r_k \boldsymbol{I}.$$

Integrating all parts from 0 to 1 and using our hypothesis,

$$(\ell - \frac{r_k}{2} M)\boldsymbol{I} \preceq \boldsymbol{G}_k \preceq (L + \frac{r_k}{2} M)\boldsymbol{I}.$$

Therefore,

$$\left(1 - h_k(L + \frac{r_k}{2} M)\right) \boldsymbol{I} \preceq \boldsymbol{I} - h_k \boldsymbol{G}_k \preceq \left(1 - h_k(\ell - \frac{r_k}{2} M)\right) \boldsymbol{I}.$$

We arrive at

$$\|\boldsymbol{I} - h_k \boldsymbol{G}_k\|_2 \leq \max\{|a_k(h_k)|, |b_k(h_k)|\}$$

where $a_k(h) = 1 - h(\ell - \frac{r_k}{2} M)$ and $b_k(h) = h(L + \frac{r_k}{2} M) - 1$.

Notice that $a_k(0) = 1$ and $b_k(0) = -1$.

Now, let us use our hypothesis that $r_0 < \bar{r}$.

When $a_k(h) = b_k(h)$, we have $1 - h(\ell - \frac{r_k}{2} M) = h(L + \frac{r_k}{2} M) - 1$, and therefore

$$h_k^* = \frac{2}{L + \ell}.$$

(Surprisingly, it does not depend neither on M nor r_k). Finally,

$$r_{k+1} = \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| \leq \left(1 - \frac{2}{L + \ell}\left(\ell - \frac{r_k}{2} M\right)\right) \|\boldsymbol{x}_k - \boldsymbol{x}^*\|.$$

That is,

$$r_{k+1} \leq \left(\frac{L - \ell}{L + \ell} + \frac{r_k M}{L + \ell}\right) r_k.$$

and $r_{k+1} < r_k < \bar{r}$.

Now, let us analyze the rate of convergence. Multiplying the above inequality by $M/(L+\ell)$,

$$\frac{M r_{k+1}}{L + \ell} \leq \frac{M(L - \ell)}{(L + \ell)^2} r_k + \frac{M^2 r_k^2}{(L + \ell)^2}.$$

Calling $\alpha_k = \frac{M r_k}{L + \ell}$ and $q = \frac{2\ell}{L + \ell}$, we have

$$\alpha_{k+1} \leq (1 - q)\alpha_k + \alpha_k^2 = \alpha_k(1 + \alpha_k - q) = \frac{\alpha_k(1 - (\alpha_k - q)^2)}{1 - (\alpha_k - q)}. \tag{1.5}$$

Now, since $r_k < \frac{2\ell}{M}$, $\alpha_k - q = \frac{Mr_k}{L+\ell} - \frac{2\ell}{L+\ell} < 0$, and $1 + (\alpha_k - q) = \frac{L-\ell}{L+\ell} + \frac{Mr_k}{L+\ell} > 0$. Therefore, $-1 < \alpha_k - q < 0$, and (1.5) becomes $\leq \frac{\alpha_k}{1+q-\alpha_k}$.

$$\frac{1}{\alpha_{k+1}} \geq \frac{1+q}{\alpha_k} - 1.$$

$$\frac{q}{\alpha_{k+1}} - 1 \geq \frac{q(1+q)}{\alpha_k} - q - 1 = (1+q)\left(\frac{q}{\alpha_k} - 1\right).$$

and then,

$$\frac{q}{\alpha_k} - 1 \geq (1+q)^k \left(\frac{q}{\alpha_0} - 1\right) = (1+q)^k \left(\frac{2\ell}{L+\ell}\frac{L+\ell}{Mr_0} - 1\right) = (1+q)^k \left(\frac{\bar{r}}{r_0} - 1\right).$$

Finally, we arrive at

$$r_k = \|\boldsymbol{x}_k - \boldsymbol{x}^*\| \leq \frac{\bar{r}r_0}{\bar{r} - r_0}\left(1 - \frac{2\ell}{L+3\ell}\right)^k.$$

∎

1.6 The Newton method

Example 1.6.1 Let us apply the Newton method to find the root of the following function

$$\phi(t) = \frac{t}{\sqrt{1+t^2}}.$$

Clearly $t^* = 0$.

The Newton method will give:

$$t_{k+1} = t_k - \frac{\phi(t_k)}{\phi'(t_k)} = t_k - t_k(1+t_k^2) = -t_k^3.$$

Therefore, the method converges if $|t_0| < 1$, it oscillates if $|t_0| = 1$, and finally, diverges if $|t_0| > 1$.

Assumption 1.6.2

1. $f \in C_M^{2,2}(\mathbb{R}^n)$.

2. There is a local minimum of the function $f(\boldsymbol{x})$ at which its Hessian is positive definite:

$$f''(\boldsymbol{x}^*) \succeq \ell \boldsymbol{I}, \quad \ell > 0.$$

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

Theorem 1.6.3 Let the function $f(\boldsymbol{x})$ satisfy the above assumptions. Suppose that the initial starting point $\boldsymbol{x}_0$ is close enough to $\boldsymbol{x}^*$:

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

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

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

Proof: Consider the Newton method $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - [f''(\boldsymbol{x}_k)]^{-1} f'(\boldsymbol{x}_k)$.

Then

$$\begin{aligned}
\boldsymbol{x}_{k+1} - \boldsymbol{x}^* &= \boldsymbol{x}_k - \boldsymbol{x}^* - [f''(\boldsymbol{x}_k)]^{-1} f'(\boldsymbol{x}_k) \\
&= \boldsymbol{x}_k - \boldsymbol{x}^* - [f''(\boldsymbol{x}_k)]^{-1} \int_0^1 f''(\boldsymbol{x}^* + \tau(\boldsymbol{x}_k - \boldsymbol{x}^*))(\boldsymbol{x}_k - \boldsymbol{x}^*) d\tau \\
&= [f''(\boldsymbol{x}_k)]^{-1} \boldsymbol{G}_k (\boldsymbol{x}_k - \boldsymbol{x}^*)
\end{aligned}$$

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

Let $r_k = \|\boldsymbol{x}_k - \boldsymbol{x}^*\|$. Then

$$\begin{aligned}
\|\boldsymbol{G}_k\| &= \| \int_0^1 [f''(\boldsymbol{x}_k) - f''(\boldsymbol{x}^* + \tau(\boldsymbol{x}_k - \boldsymbol{x}^*))] d\tau \| \\
&\le \int_0^1 \|f''(\boldsymbol{x}_k) - f''(\boldsymbol{x}^* + \tau(\boldsymbol{x}_k - \boldsymbol{x}^*))\| d\tau \\
&\le \int_0^1 M|1-\tau| r_k d\tau = \frac{r_k}{2} M.
\end{aligned}$$

From Lemma 1.4.6 and from the hypothesis

$$f''(\boldsymbol{x}_k) \succeq f''(\boldsymbol{x}^*) - M r_k \boldsymbol{I} \succeq (\ell - M r_k)\boldsymbol{I}.$$

For $r_0 < \bar{r} = \frac{2\ell}{3M} < \frac{\ell}{M}$,

$$\|[f''(\boldsymbol{x}_0)]^{-1}\| \le (\ell - M r_0)^{-1}.$$

Then

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

Since $r_0 < \bar{r}$, $\frac{M r_0}{2(\ell - M r_0)} < \frac{\ell}{3(\ell - M r_0)} < 1$, and $r_1 < r_0$. This argument is valid for all k's. ∎

- Comparing this result with the rate of convergence of the gradient method, 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.

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

- This justifies a standard recommendation to use the gradient 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.

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

Definition 1.7.1 Given a starting point $\boldsymbol{x}_0$, the linear *Krylov subspaces* is defined as

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

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

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

Lemma 1.7.2 For any $k \geq 1$, $\mathcal{L}_k = \mathrm{Lin}\{f'(\boldsymbol{x}_0), \ldots, f'(\boldsymbol{x}_{k-1})\}$.

Proof: Let us prove by induction hypothesis.

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

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,

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

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

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

If the equality does not hold, $f'(\boldsymbol{x}_k) \in \mathcal{L}_k$ 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). ∎