2.5 The gradient method for smooth and strongly convex functions

Let us consider the gradient method with constant step h.

Theorem 2.5.1 Let $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$, and $0 < h < \frac{2}{L}$. The gradient method with constant step generates a sequence which converges as follows:

$$f(\boldsymbol{x}_k) - f^* \leq \frac{2(f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*))\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2}{2\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 + kh(2 - Lh)(f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*))}.$$

Proof: Denote $r_k = \|\boldsymbol{x}_k - \boldsymbol{x}^*\|$. Then

$$\begin{aligned} r_{k+1}^2 &= \|\boldsymbol{x}_k - \boldsymbol{x}^* - hf'(\boldsymbol{x}_k)\|^2 \\ &= r_k^2 - 2h\langle f'(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{x}^*\rangle + h^2\|f'(\boldsymbol{x}_k)\|^2 \\ &= r_k^2 - 2h\langle f'(\boldsymbol{x}_k) - f'(\boldsymbol{x}^*), \boldsymbol{x}_k - \boldsymbol{x}^*\rangle + h^2\|f'(\boldsymbol{x}_k)\|^2 \\ &\leq r_k^2 - h\left(\frac{2}{L} - h\right)\|f'(\boldsymbol{x}_k)\|^2, \end{aligned}$$

where the last inequality follows from Theorem 2.1.8.

Therefore, $r_{k+1} < r_k < \cdots < r_0$.

Now

$$\begin{aligned} f(\boldsymbol{x}_{k+1}) &\leq f(\boldsymbol{x}_k) + \langle f'(\boldsymbol{x}_k), \boldsymbol{x}_{k+1} - \boldsymbol{x}_k\rangle + \frac{L}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2 \\ &= f(\boldsymbol{x}_k) - \omega\|f'(\boldsymbol{x}_k)\|^2 < f(\boldsymbol{x}_k), \end{aligned} \tag{2.1}$$

where $\omega = h(1 - \frac{L}{2}h)$. Denoting by $\Delta_k = f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)$, from the convexity of f,

$$\Delta_k = f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \langle f'(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{x}^*\rangle \leq \|f'(\boldsymbol{x}_k)\| r_k \leq \|f'(\boldsymbol{x}_k)\| r_0. \tag{2.2}$$

Combining (2.1) and (2.2),

$$\Delta_{k+1} \leq \Delta_k - \frac{\omega}{r_0^2}\Delta_k^2.$$

Thus dividing by $\Delta_k\Delta_{k+1}$,

$$\frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{\omega}{r_0^2}\frac{\Delta_k}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{\omega}{r_0^2}.$$

Summing up these inequalities we get

$$\frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_0} + \frac{\omega}{r_0^2}(k+1).$$

∎

To obtain the optimal step size, it is sufficient to find the maximum of the function $\omega = \omega(h) = h(1 - \frac{L}{2}h)$ which is $h^* = 1/L$.

Corollary 2.5.2 If $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$, the gradient method with constant step $h = 1/L$ yields

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \frac{2L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2}{k+4}.$$

Proof: Left for exercise. ∎

Theorem 2.5.3 Let $f \in \mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$, and $0 < h \leq \frac{2}{\mu+L}$. The gradient method with constant step generates a sequence which converges as follows:

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 \leq \left(1 - \frac{2h\mu L}{\mu + L}\right)^k \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2.$$

If $h = \frac{2}{\mu+L}$, then

$$\begin{aligned} \|\boldsymbol{x}_k - \boldsymbol{x}^*\| &\leq \left(\frac{L/\mu - 1}{L/\mu + 1}\right)^k \|\boldsymbol{x}_0 - \boldsymbol{x}^*\| \\ f(\boldsymbol{x}_k) - f^* &\leq \frac{L}{2}\left(\frac{L/\mu - 1}{L/\mu + 1}\right)^{2k} \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2. \end{aligned}$$

Proof: Denote $r_k = \|\boldsymbol{x}_k - \boldsymbol{x}^*\|$. Then

$$\begin{aligned} r_{k+1}^2 &= \|\boldsymbol{x}_k - \boldsymbol{x}^* - hf'(\boldsymbol{x}_k)\|^2 \\ &= r_k^2 - 2h\langle f'(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{x}^*\rangle + h^2\|f'(\boldsymbol{x}_k)\|^2 \\ &= r_k^2 - 2h\langle f'(\boldsymbol{x}_k) - f'(\boldsymbol{x}^*), \boldsymbol{x}_k - \boldsymbol{x}^*\rangle + h^2\|f'(\boldsymbol{x}_k)\|^2 \\ &\leq r_k^2 - 2h\left(\frac{\mu L}{\mu + L} r_k^2 + \frac{1}{\mu + L}\|f'(\boldsymbol{x}_k) - f'(\boldsymbol{x}^*)\|^2\right) + h^2\|f'(\boldsymbol{x}_k)\|^2 \\ &= \left(1 - \frac{2h\mu L}{\mu + L}\right) r_k^2 + h\left(h - \frac{2}{\mu + L}\right)\|f'(\boldsymbol{x}_k)\|^2 \end{aligned}$$

from Theorem 2.3.7, and proves the first two inequalities.

Now, from Theorem 2.1.8,

$$\begin{aligned} f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) - \langle f'(\boldsymbol{x}^*), \boldsymbol{x}_k - \boldsymbol{x}^*\rangle &\leq \frac{L}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 \\ &\leq \frac{L}{2}\left(\frac{L/\mu - 1}{L/\mu + 1}\right)^{2k} r_0^2. \end{aligned}$$

∎

Theorem 2.5.4 In the special case of a strongly convex quadratic function $f(\boldsymbol{x}) = \frac{1}{2}\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}\rangle + \langle \boldsymbol{a}, \boldsymbol{x}\rangle + \alpha$ with $\lambda_1(\boldsymbol{A}) = L \geq \lambda_n(\boldsymbol{A}) = \mu > 0$, we can obtain

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\| \leq \left(\frac{L/\mu - 1}{L/\mu + \sqrt{\frac{\mu}{2L}}}\right)^k \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|$$

for the gradient method with exact line search.

Proof: See [YUAN2010]. ▮

- Note that the previous result for the gradient method Theorem 1.5.5 was only a local result.

- Comparing the rate of convergence of the gradient method for the classes $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$ and $\mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$, Theorems 2.5.1 (Corollary 2.5.2) and 2.5.3 with their lower complexity bounds, Theorems 2.2.1 and 2.4.1, respectively, we possible have a huge gap.

2.6 The optimal gradient method

Definition 2.6.1 A pair of sequences $\{\phi_k(\boldsymbol{x})\}_{k=0}^{\infty}$ and $\{\lambda_k\}_{k=0}^{\infty}$ with $\lambda_k \geq 0$ is called an *estimate sequence* of the function $f(\boldsymbol{x})$ if

$$\lambda_k \to 0,$$

and for any $\boldsymbol{x} \in \mathbb{R}^n$ and any $k \geq 0$, we have

$$\phi_k(\boldsymbol{x}) \leq (1-\lambda_k) f(\boldsymbol{x}) + \lambda_k \phi_0(\boldsymbol{x}).$$

Lemma 2.6.2 Given an estimate sequence $\{\phi_k(\boldsymbol{x})\}_{k=0}^{\infty}$, $\{\lambda_k\}_{k=0}^{\infty}$, and if for some sequence $\{\boldsymbol{x}_k\}_{k=1}^{\infty}$ we have

$$f(\boldsymbol{x}_k) \leq \phi_k^* \equiv \min_{\boldsymbol{x} \in \mathbb{R}^n} \phi_k(\boldsymbol{x})$$

then $f(\boldsymbol{x}_k) - f^* \leq \lambda_k(\phi_0(\boldsymbol{x}^*) - f(\boldsymbol{x}^*)) \to 0$.

Proof: It follows from the definition. ▮

Lemma 2.6.3 Assume that

1. $f \in \mathcal{S}_\mu^1(\mathbb{R}^n)$, possible with $\mu = 0$ (which means that $f \in \mathcal{F}^1(\mathbb{R}^n)$).

2. $\phi_0(\boldsymbol{x})$ is an arbitrary function on $\mathbb{R}^n$.

3. $\{\boldsymbol{y}_k\}_{k=0}^{\infty}$ is an arbitrary sequence in $\mathbb{R}^n$.

4. $\{\alpha_k\}_{k=0}^{\infty}$ is an arbitrary sequence such that $\alpha_k \in (0,1)$, $\sum_{k=0}^{\infty} \alpha_k = \infty$, and $\alpha_{-1} = 0$.

Then the pair of sequences $\{\Pi_{i=-1}^{k-1}(1-\alpha_i)\}_{k=0}^{\infty}$ and $\{\phi_k(\boldsymbol{x})\}_{k=0}^{\infty}$ recursively defined as

$$\phi_{k+1}(\boldsymbol{x}) \quad = \quad (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k \left[f(\boldsymbol{y}_k) + \langle f'(\boldsymbol{y}_k), \boldsymbol{x} - \boldsymbol{y}_k \rangle + \frac{\mu}{2} \|\boldsymbol{x} - \boldsymbol{y}_k\|^2 \right]$$

is an estimate sequence.

Proof: Let us prove by induction on k. For $k = 0$, $\phi_0(\boldsymbol{x}) = (1 - (1 - \alpha_{-1}))\, f(\boldsymbol{x}) + (1 - \alpha_{-1})\phi_0(\boldsymbol{x})$ since $\alpha_{-1} = 1$. Suppose that the induction hypothesis is valid for k. Since $f \in \mathcal{S}^1_\mu(\mathbb{R}^n)$,

$$\begin{aligned}
\phi_{k+1}(\boldsymbol{x}) &= (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k \left[f(\boldsymbol{y}_k) + \langle f'(\boldsymbol{y}_k), \boldsymbol{x} - \boldsymbol{y}_k\rangle + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}_k\|^2 \right] \\
&\leq (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k f(\boldsymbol{x}) \\
&= \left(1 - (1-\alpha_k)\Pi_{i=-1}^{k-1}(1-\alpha_i)\right) f(\boldsymbol{x}) + (1-\alpha_k)\left(\phi_k(\boldsymbol{x}) - (1 - \Pi_{i=-1}^{k-1}(1-\alpha_i)) f(\boldsymbol{x})\right) \\
&\leq \left(1 - (1-\alpha_k)\Pi_{i=-1}^{k-1}(1-\alpha_i)\right) f(\boldsymbol{x}) + (1-\alpha_k)\Pi_{i=-1}^{k-1}(1-\alpha_i)\phi_0(\boldsymbol{x}) \\
&= \left(1 - \Pi_{i=-1}^{k}(1-\alpha_i)\right) f(\boldsymbol{x}) + \Pi_{i=-1}^{k}(1-\alpha_i)\phi_0(\boldsymbol{x}).
\end{aligned}$$

The remaining part is left for exercise. ∎

Lemma 2.6.4 Let $\gamma_0, \phi_0^* \in \mathbb{R}$, $\mu \in \mathbb{R}$ (possible with $\mu = 0$), $\boldsymbol{v}_0 \in \mathbb{R}^n$, and $\{\boldsymbol{y}_k\}_{k=0}^\infty$ a given arbitrarily sequence. Define $\phi_0(\boldsymbol{x}) = \phi_0^* + \frac{\gamma_0}{2}\|\boldsymbol{x} - \boldsymbol{v}_0\|^2$. If we define recursively $\phi_{k+1}(\boldsymbol{x})$ such as the previous lemma:

$$\phi_{k+1}(\boldsymbol{x}) = (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k \left[f(\boldsymbol{y}_k) + \langle f'(\boldsymbol{y}_k), \boldsymbol{x} - \boldsymbol{y}_k\rangle + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}_k\|^2 \right],$$

for an arbitrary sequence $\{\alpha_k\}_{k=0}^\infty$ such that $\alpha_k \in (0,1)$ and $\sum_{k=0}^\infty \alpha_k = \infty$. Then $\phi_{k+1}(\boldsymbol{x})$ preserve the canonical form

$$\phi_{k+1}(\boldsymbol{x}) = \phi_{k+1}^* + \frac{\gamma_{k+1}}{2}\|\boldsymbol{x} - \boldsymbol{v}_{k+1}\|^2 \tag{2.3}$$

for

$$\begin{aligned}
\gamma_{k+1} &= (1-\alpha_k)\gamma_k + \alpha_k\mu, \\
\boldsymbol{v}_{k+1} &= \frac{1}{\gamma_{k+1}}[(1-\alpha_k)\gamma_k\boldsymbol{v}_k + \alpha_k\mu\boldsymbol{y}_k - \alpha_k f'(\boldsymbol{y}_k)], \\
\phi_{k+1}^* &= (1-\alpha_k)\phi_k^* + \alpha_k f(\boldsymbol{y}_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|f'(\boldsymbol{y}_k)\|^2 \\
&\quad + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_k\|^2 + \langle f'(\boldsymbol{y}_k), \boldsymbol{v}_k - \boldsymbol{y}_k\rangle\right).
\end{aligned}$$

Proof: We will use again the induction hypothesis in k. Note that $\phi_0''(\boldsymbol{x}) = \gamma_0\boldsymbol{I}$. Now, for any $k \geq 0$,

$$\phi_{k+1}''(\boldsymbol{x}) = (1-\alpha_k)\phi_k''(\boldsymbol{x}) + \alpha_k\mu\boldsymbol{I} = ((1-\alpha_k)\gamma_k + \alpha_k\mu)\,\boldsymbol{I} = \gamma_{k+1}\boldsymbol{I}.$$

Therefore, $\phi_{k+1}(\boldsymbol{x})$ is a quadratic function of the form (2.3). From the first-order optimality condition

$$\begin{aligned}
\phi_{k+1}'(\boldsymbol{x}) &= (1-\alpha_k)\phi_k'(\boldsymbol{x}) + \alpha_k f'(\boldsymbol{y}_k) + \alpha_k\mu(\boldsymbol{x} - \boldsymbol{y}_k) \\
&= (1-\alpha_k)\gamma_k(\boldsymbol{x} - \boldsymbol{v}_k) + \alpha_k f'(\boldsymbol{y}_k) + \alpha_k\mu(\boldsymbol{x} - \boldsymbol{y}_k) = 0.
\end{aligned}$$