4.4 Smooth convex optimization

All convergence results presented so far have a local nature and do not guarantee *global* convergence to a *global* minimum. More can be said if the target functional is convex. This section mainly follows Chapter 2 from [N] but it also includes some of the discussion from `https://blogs.princeton.edu/imabandit/orf523-the-complexities-of-optimization/`.

4.4.1 Convex functions

Definition 4.18 *A function $f : \mathbb{R}^n \to \mathbb{R}$ is called* convex *if*

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y}) \tag{4.32}$$

holds for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\alpha \in [0,1]$.

It is often quite tedious to verify the condition (4.32). It is much easier to check convexity for (twice) continuously differentiable functions.

Lemma 4.19 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. f is convex if and only if one of the following three conditions holds for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:*

1.
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}), \tag{4.33}$$

2.
$$(\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}))^T(\mathbf{y}-\mathbf{x}) \geq 0, \tag{4.34}$$

3.
$$H(\mathbf{x}) \geq 0, \tag{4.35}$$

where we additionally assume for (4.35) that f is twice continuously differentiable.

Proof. In the following, we set $\mathbf{x}_\alpha = \alpha\mathbf{x} + (1-\alpha)\mathbf{y}$.

1. (4.32) $\Rightarrow$ (4.33): Rearranging (4.32) gives for any $\alpha \in [0,1[$:

$$f(\mathbf{y}) \geq \frac{1}{1-\alpha}\big(f(\mathbf{x}_\alpha) - \alpha f(\mathbf{x})\big) = f(\mathbf{x}) + \frac{1}{1-\alpha}\big(f(\mathbf{x} + (1-\alpha)(\mathbf{y}-\mathbf{x})) - f(\mathbf{x})\big).$$

Letting $\alpha \to 1$ implies (4.33).

(4.33) $\Rightarrow$ (4.32). Applying (4.33) to $\mathbf{y}, \mathbf{x}_\alpha$ and $\mathbf{x}, \mathbf{x}_\alpha$, respectively, gives

$$f(\mathbf{y}) \geq f(\mathbf{x}_\alpha) + \nabla f(\mathbf{x}_\alpha)^T(\mathbf{y}-\mathbf{x}_\alpha),$$
$$f(\mathbf{x}) \geq f(\mathbf{x}_\alpha) + \nabla f(\mathbf{x}_\alpha)^T(\mathbf{x}-\mathbf{x}_\alpha).$$

Multiplying these inequalities with α and $1-\alpha$, respectively, and adding the results up yields (4.32).

2. (4.33) $\Rightarrow$ (4.34): Reversing the roles of $\mathbf{x}$ and $\mathbf{y}$ in (4.33) gives the ineqality $-f(\mathbf{y}) \geq -f(\mathbf{x}) - \nabla f(\mathbf{y})^T(\mathbf{y}-\mathbf{x})$, which – when added to (4.33) – yields (4.34).

(4.34) ⇒ (4.33): This implication follows from

$$\begin{aligned} f(\mathbf{y}) - f(\mathbf{x}) &= \int_0^1 \nabla f(\mathbf{x}_\alpha)^T(\mathbf{y}-\mathbf{x})\,\mathrm{d}\alpha \\ &= \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \int_0^1 \big(\nabla f(\mathbf{x}_\alpha) - \nabla f(\mathbf{x})\big)^T(\mathbf{y}-\mathbf{x})\,\mathrm{d}\alpha \\ &= \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{1}{1-\alpha}\int_0^1 \big(\nabla f(\mathbf{x}_\alpha) - \nabla f(\mathbf{x})\big)^T(\mathbf{x}_\alpha-\mathbf{x})\,\mathrm{d}\alpha \\ &\geq \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}). \end{aligned}$$

3. (4.34) ⇒ (4.35): Let $\mathbf{x}_\tau = \mathbf{x} + \tau\mathbf{s}$ for an arbitrary vector $\mathbf{s} \in \mathbb{R}^n$ and $\tau > 0$. Then (4.34) yields

$$0 \leq \frac{1}{\tau^2}\big(\nabla f(\mathbf{x}_\tau) - \nabla f(\mathbf{x})\big)^T(\mathbf{x}_\tau - \mathbf{x}) = \frac{1}{\tau}\big(\nabla f(\mathbf{x}_\tau) - \nabla f(\mathbf{x})\big)^T\mathbf{s} \overset{\tau\to 0}{\longrightarrow} \mathbf{s}^T H(\mathbf{x})\mathbf{s},$$

which implies positive semidefiniteness of $H(\mathbf{x})$.

4. (4.35) ⇒ (4.34): This implication follows from the Taylor expansion, with an integral representation of the remainder term. □

Examples of convex functions:

- The univariate functions e^x, $|x|^p$ for $p \geq 1$, $\frac{x^2}{1-|x|}$, $|x| - \log(1+|x|)$ are convex.
- The norm properties imply that *any* vector norm on $\mathbb{R}^n$ is convex. In particular, this holds for $\|\mathbf{x}\|_2$ and $\|\mathbf{x}\|_1$.
- The function $f(\mathbf{x}) = \alpha + \mathbf{b}^T\mathbf{x} + \frac{1}{2}\mathbf{x}^T A\mathbf{x}$ is convex if A is symmetric positive semidefinite. In particular, any linear function $\alpha + \mathbf{b}^T\mathbf{x}$ is convex.
- If f_1, f_2 are convex functions then the functions $f_1(\mathbf{x}) + f_2(\mathbf{x})$ and $\max\{f_1(\mathbf{x}), f_2(\mathbf{x})\}$ are convex as well.
- If f is convex then the function $\varphi(\mathbf{x}) := f(A\mathbf{x} + \mathbf{b})$ is convex as well for any matrix A and vector $\mathbf{b}$ of suitable size.

The following result is one of the main reasons for the importance of convex functions.

Theorem 4.20 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and continuously differentiable. Then x^* is a global minimizer for f if and only if $\nabla f(\mathbf{x}^*) = 0$.*

Proof. One direction follows immediately from Theorem 4.1. For the other direction, suppose that $\nabla f(\mathbf{x}^*) = 0$. Then (4.33) immediately implies $f(\mathbf{x}) \geq f(\mathbf{x}^*)$ for any $\mathbf{x} \in \mathbb{R}^n$ and thus $\mathbf{x}^*$ is a global minimizer. □

4.4.2 Strongly convex functions

We already know that the local convergence rate of gradient descent methods depends on the condition number of $H(\mathbf{x}^*)$. Since convexity does not necessarily imply that this condition number is finite; we need to introduce a stronger concept to ensure fast convergence of gradient descent.

Definition 4.21 *A continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is called* strongly convex *if there is $\mu > 0$ such that*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(x)^T(\mathbf{y} - \mathbf{x}) + \frac{1}{2}\mu\|\mathbf{y} - \mathbf{x}\|_2^2$$

holds for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

Later on, the value of μ will play an important role and we will sometimes say *μ-strongly convex function*. When adding an μ_1-strongly convex function with an μ_2-strongly convex function, one obtains an $(\mu_1 + \mu_2)$-strongly convex function.

An immediate consequence of Definition 4.21, we have

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) + \frac{1}{2}\mu\|\mathbf{x} - \mathbf{x}^*\|_2^2$$

at a minimizer $\mathbf{x}^*$. Thus, the minimizer $\mathbf{x}^*$ is uniquely determined.

The following lemma extends Lemma 4.19 and can be proven in a similar manner.

Lemma 4.22 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. f is μ-strongly convex if and only if one of the following three conditions holds for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:*

1. $f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y}) - \alpha(1-\alpha)\frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 \quad \forall \alpha \in [0,1]$,
2. $(\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}))^T(\mathbf{y} - \mathbf{x}) \geq \mu\|\mathbf{x} - \mathbf{y}\|_2^2$,
3. $H(\mathbf{x}) \geq \mu I_n$,

where we additionally assume for the last case that f is twice continuously differentiable.

4.4.3 Nesterov's accelerated gradient descent method

The purpose of the line search strategies from Section 4.2 was to ensure global convergence to a stationary point. This is much less of a concern for convex functions, for which a constant step size is sufficient.

The following theorem summarizes the results from Section 2.1.5 of [N]. Note that we call a function *L-smooth* if it is continously differentiable and its gradient is Lipschitz continuous with Lipschitz constant L:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2 \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

If f is twice continuously differentiable, this is equivalent to $\|H(\mathbf{x})\|_2 \leq L$ for all $\mathbf{x} \in \mathbb{R}^n$.

Theorem 4.23 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex L-smooth function with minimum $f^* = f(\mathbf{x}^*)$. Then the sequence generated by $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)$ for some initial vector $\mathbf{x}_0 \in \mathbb{R}^n$ has the following properties:*

1. *if $\alpha = \frac{1}{L}$ then $f(\mathbf{x}_k) - f^* \leq \frac{2L\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{k+4}$;*
2. *if f is μ-strongly convex and $\alpha = \frac{2}{\mu + L}$ then*

$$\|\mathbf{x}_k - \mathbf{x}_0\|_2^2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|\mathbf{x}_0 - \mathbf{x}_k\|_2^2,$$

$$f(\mathbf{x}_k) - f^* \leq \frac{L}{2}\left(\frac{\kappa - 1}{\kappa + 1}\right)^{2k} \|\mathbf{x}_0 - \mathbf{x}_k\|_2^2,$$

where $\kappa = L/\mu$.

The convergence predicted in the first part of the theorem is algebraic of order 1: $O(1/k)$. This is not optimal and can be improved with an adapted choice of α, as we will see below. The convergence predicted in the second part of the theorem is very similar to Theorem 4.9, with a difference of major importance: Theorem 4.9 is a statement about local convergence, while the second part of Theorem 4.23 ensures global (exponential) convergence. It turns out that the convergence rates predicted by Theorem 4.23 can be improved significantly by adjusting the step size.