# Lecture 9. Descent Method

## 9.1 Unconstrained optimization problems

We now study the general convex optimization problems. First, we consider the easiest case: no constraints. Namely, the optimization problem is

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

where $f(x)$ is a convex function.

Recall that, the optimality condition for convex functions is

**Theorem**

Suppose $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ is a convex function. Then $x^*$ is a *global* minimum point of $f$ iff

$$\forall\, y \in D, \quad \nabla f(x^*)^\mathsf{T}(y - x^*) \geq 0\,.$$

In particular, if $D = \mathbb{R}^n$, then $x^*$ is a global minimum point iff $\nabla f(x^*) = \mathbf{0}$.

For convenience, we assume that $D = \mathbb{R}^n$, the objective function $f(x)$ is differentiable and has a finite minimum point $x^*$ (and the minimum value $f^*$). For some simple cases, we can compute the minimum point by solving the equation $\nabla f(x^*) = 0$. However, in general we cannot expect that closed-form solutions always exist. So we introduce some algorithms to find optimal solutions.

## 9.2 Descent method

Analogously to the simplex method, we would like to move from a solution $x$ to a better "neighbor" $y$. The convexity guarantees that

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

As we hope $y$ is better, i.e., $f(y) < f(x)$, it requires that $\nabla f(x)^\mathsf{T}(y - x) < 0$.

Conversely, we know that if the directional derivative $\nabla f(x)^\top v < 0$, then there exists $\varepsilon > 0$ such that $f(x + \varepsilon v) < f(x)$. So $\nabla f(x)^\top v < 0$ is a reasonable requirement for the moving direction $v$.

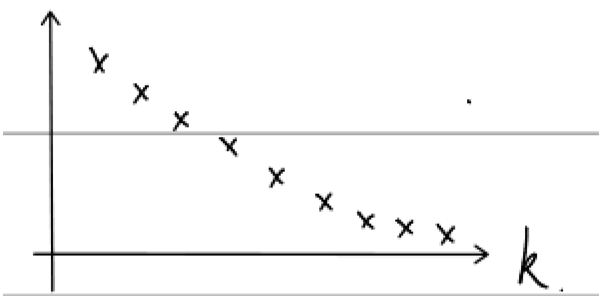
This inspired the so-called *descent method*: start from a solution $\boldsymbol{x}_0$ and move to $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + t_k \boldsymbol{v}_k$ iteratively, where $t_k$ is the *step size* to be determined and $\boldsymbol{v}_k$ is the moving direction satisfying $\nabla f(\boldsymbol{x}_k)^\top \boldsymbol{v}_k < 0$.

The first question is when we can stop? Of course the ideal stopping criterion is $\nabla f(\boldsymbol{x}_k) = \mathbf{0}$ for some $k$. If so, we know that $x_k$ is indeed a minimum point. However, in practice, we cannot expect this happens. So we usually use stopping criteria such as $\|\nabla f(\boldsymbol{x})\| < \delta$, $|f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k)| < \delta$, or 1000 iterations.

```
given a starting point x_0
repeat
    choose a proper step size t_k
    x_{k+1} ← x_k + t_k v_k where ∇f(x_k)^T v_k < 0
    k ← k + 1
until ||∇f(x_k)|| ≤ δ for some sufficiently small δ
```

The next question is, does this algorithm converge to an optimal solution? In fact, we claim that if we assume that $t_k, \boldsymbol{v}_k$ only depend on $\boldsymbol{x}_k$, and the choice of $t_k$ satisfies $f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$ for every $\boldsymbol{x}_k \notin \arg\min f(\boldsymbol{x})$ (note that the optimal solution may not be unique), then the value of objective functions $\{f(\boldsymbol{x}_k)\}$ generated by the descent method converge to the minimum value $f^*$. (However, $\{\boldsymbol{x}_k*\}$ may not converge and we will give an example later.)

We assume $f(\boldsymbol{x})$ has a finite minimum value $f^*$, and $f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$. So $f(\boldsymbol{x}_k)$ has a limit as $k$ goes to infinity. Now we would like to show that the limit is $f^*$.




Let $c = \lim_{k\to\infty} f(\boldsymbol{x}_k)$. Intuitively, if $c > f^*$, as we hope $f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$ as long as $f(\boldsymbol{x}_k) \neq f^*$, we can argue that $f(\boldsymbol{x}_{k+1})$ still decreases too fast even if $f(\boldsymbol{x}_k)$ is sufficiently close to $c$.

Rigorously, let $S = \{\boldsymbol{x} \mid c \le f(\boldsymbol{x}) \le f(\boldsymbol{x}_0)\}$. Then $S$ is a *compact set*, if we assume $f(\infty) = \infty$ for convenience (otherwise $S$ may not be necessarily bounded). Let $g(x) : \mathbb{R}^n \to \mathbb{R}$ be a function defined by

$$g(\boldsymbol{y}) = f(\boldsymbol{y} + t_{\boldsymbol{y}}\boldsymbol{v}_{\boldsymbol{y}}) - f(\boldsymbol{y})\,,$$

where $t_{\boldsymbol{y}}, \boldsymbol{v}_{\boldsymbol{y}}$ are the step size and the direction we choose if $\boldsymbol{x}_k = \boldsymbol{y}$. That is, $g(\boldsymbol{y})$ measures the difference between $f(\boldsymbol{x}_{k+1})$ and $f(\boldsymbol{x}_k)$ if we set $\boldsymbol{x}_k = \boldsymbol{y}$.

By our assumption $f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$ as long as $f(\boldsymbol{x}_k) \ne f^*$, and noting that $x \in S$ if $f(\boldsymbol{x}) \ge c > f^*$, we conclude that $g(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in S$. Applying the extreme value theorem, there exists

$$\delta = \min_{\boldsymbol{x} \in S} g(\boldsymbol{x}) > 0\,,$$

which implies that $f(\boldsymbol{x}_{k+1}) \le f(\boldsymbol{x}_k) - \delta$ for every $\boldsymbol{x}_k \in S$.This contradicts our assumption that there exists $\{\boldsymbol{x}_k\}$ such that $f(\boldsymbol{x}_k) \downarrow c$, and thus completes the proof.

**Tip**

In fact, it is not necessary to define $g$ as the difference between the function values. Analogously to the *amortised analysis* for some data structures, we may define $g$ to measure the difference between some *potential function*. So this argument above is a simplified result of the *Lyapunov's global stability theorem in discrete time*.
Suppose $d_{k+1} = \rho(d_k)$ where $\rho : \mathbb{R}^n \to \mathbb{R}$ is a continuous function and $\rho(0) = 0$. If there exists a continuous (Lyapunov) function $\ell : \mathbb{R}^n \to \mathbb{R}$ such that

1. $\ell(0) = 0$, $\ell(x) > 0$ for all $x \ne 0$, (*positivity*)
2. $\ell(x) \to \infty$ as $\|x\| \to \infty$, (*radical unboundedness*)
3. $\ell(\rho(x)) < \ell(x)$ for all $x \ne 0$. (*strict decrease*)

Then for all $d_0 \in \mathbb{R}^n$, we have $d_k \to 0$ as $k \to \infty$.
For our setting, just select an optimal solution $\boldsymbol{x}^*$, and set $d_k = \boldsymbol{x}_k - \boldsymbol{x}^*$, $\rho(d_k) = d_k + t_k\boldsymbol{v}_k$ and $\ell(d_k) = f(d_k + \boldsymbol{x}^*) - f^*$.

## 9.3 Gradient descent

We now consider a specific descent method, the *gradient descent*, where we select $v_k = -\nabla f(x_k)$. Then trivially $\nabla f(x_k)^\mathsf{T} v_k < 0$.

There is an advantage to choose $-\nabla f(x_k)$ since it is the direction of *steepest descent*, namely, the value of $f$ decreases most rapidly: For any *unit* length vector $v$, the directional derivative $\nabla f(x)^\top v$ satisfies

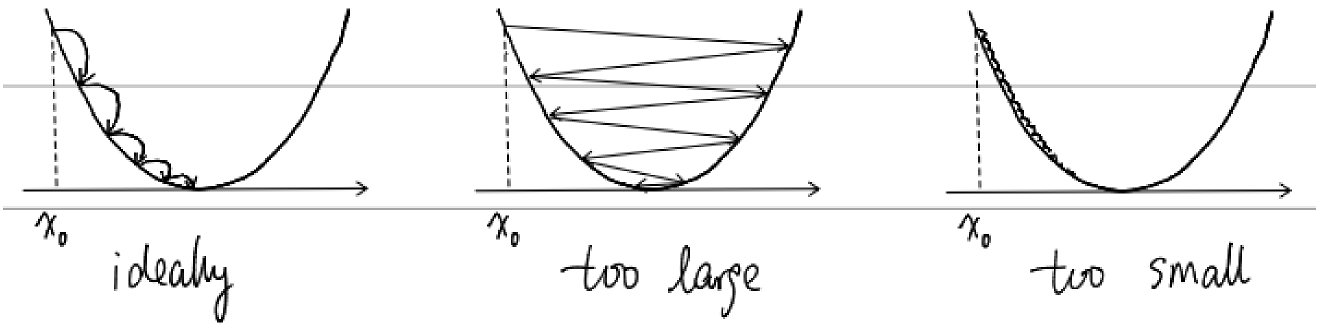
$$-\|v\| \cdot \|\nabla f(x)\| \leq \nabla f(x)^\top v \leq \|v\| \cdot \|\nabla f(x)\|$$

by the Cauchy-Schwarz inequality, and the equality holds iff $v = \pm\nabla f(x)/\|\nabla f(x)\|$.

Applying this choice of directions, we obtain the *gradient descent method*:

given a starting point $x_0$
repeat
    choose a proper step size $t_k$
    $x_{k+1} \leftarrow x_k - t_k \nabla f(x_k)$
    $k \leftarrow k + 1$
until $\|\nabla f(x_k)\| \leq \delta$ for some sufficiently small $\delta$

We now consider how to choose the step size $t_k$. Intuitively, the choice of step size can effect the converge rate of the algorithm.




Let's start from an easy example: $f(x) = ax^2$ where $a > 0$. Since we hope $f(x_{k+1}) < f(x_k)$, it requires that $|x_{k+1}| < |x_k|$, which is equivalent to

$$|(1 - 2at_k)x_k| < |x_k| \,.$$

So $t_k < 1/a$ suffices.

Next, consider the multivariate function $f(\boldsymbol{x}) = \boldsymbol{x}^\top Q\boldsymbol{x}$ where $Q \succeq 0$. Now $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - 2t_k Q\boldsymbol{x}_k$. So

$$f(\boldsymbol{x}_{k+1}) = \boldsymbol{x}_k^\top Q\boldsymbol{x}_k + 4t_k^2(Q\boldsymbol{x}_k)^\top Q(Q\boldsymbol{x}_k) - 4t_k(Q\boldsymbol{x}_k)^\top(Q\boldsymbol{x}_k)\,.$$

It is sufficient to find a value of $t_k$ such that for all $\boldsymbol{v} \in \mathbb{R}^n$, $t_k\boldsymbol{v}^\top Q\boldsymbol{v} < \boldsymbol{v}^\top\boldsymbol{v}$. We need the following lemma.

**Lemma (*Rayleigh quotient*)**

Let $Q \succeq 0$ be a positive semi-definite matrix, and $\lambda_{\min}$ and $\lambda_{\max}$ be its minimum and maximum eigenvalues, respectively. Then for all $x \in \mathbb{R}^n$, we have

$$\lambda_{\min}\|x\|_2^2 \leq x^\top Q x \leq \lambda_{\max}\|x\|_2^2$$

**Proof**

Since $Q \in \mathbb{R}^{n\times n}$ is symmetric, consider its eigen-decomposition $Q = U\Lambda U^\top$, where $\Lambda = \operatorname{diag}\{\lambda_1, \dots, \lambda_n\}$ is the diagonal matrix consisting of $Q$'s eigenvalues, and $U = (u_1, \dots, u_n)$ consists of corresponding unit-length eigenvectors. It easy to see that $UU^\top = I$.
Assume $\boldsymbol{x} = U\boldsymbol{y}$ (i.e. $\boldsymbol{y} = U^{-1}\boldsymbol{x} = U^\top\boldsymbol{x}$). Then

$$\boldsymbol{x}^\top Q\boldsymbol{x} = \boldsymbol{y}^\top U^\top QU\boldsymbol{y} = \boldsymbol{y}^\top U^\top U\Lambda U^\top U\boldsymbol{y} = \boldsymbol{y}^\top\Lambda\boldsymbol{y} = \sum_{i=1}^{n}\lambda_i y_i^2\,.$$

So clearly we have $\lambda_{\min}\|\boldsymbol{y}\|^2 \leq \boldsymbol{x}^\top Q\boldsymbol{x} \leq \lambda_{\max}\|\boldsymbol{y}\|^2$. Moreover, we have

$$\|\boldsymbol{y}\|^2 = \boldsymbol{y}^\top\boldsymbol{y} = \boldsymbol{x}^\top U^\top U\boldsymbol{x} = \boldsymbol{x}^\top\boldsymbol{x} = \|\boldsymbol{x}\|^2\,,$$

which completes the proof.

Note that in this proof we do not really need $Q \succeq 0$. This lemma holds for all symmetric $Q$. Applying this lemma, it gives that $t_k < 1/\lambda_{\max}$ suffices in the gradient descent method for quadratic functions.

However, for general cases, we cannot expect a universal condition for $t_k$. For example, consider the function $f(x) = |x|$. If we choose $t_k$ to be a constant $t > 0$, no matter what value $t$ is, the algorithm does not work as long as $|x_k| < t$.

**Question**

Under which assumptions can we choose a constant as the step size?

## 9.4 $L$-smooth functions

We would like to avoid functions similar to $|x|$, where $\nabla f(x)$ changes too drastically near $x^*$.

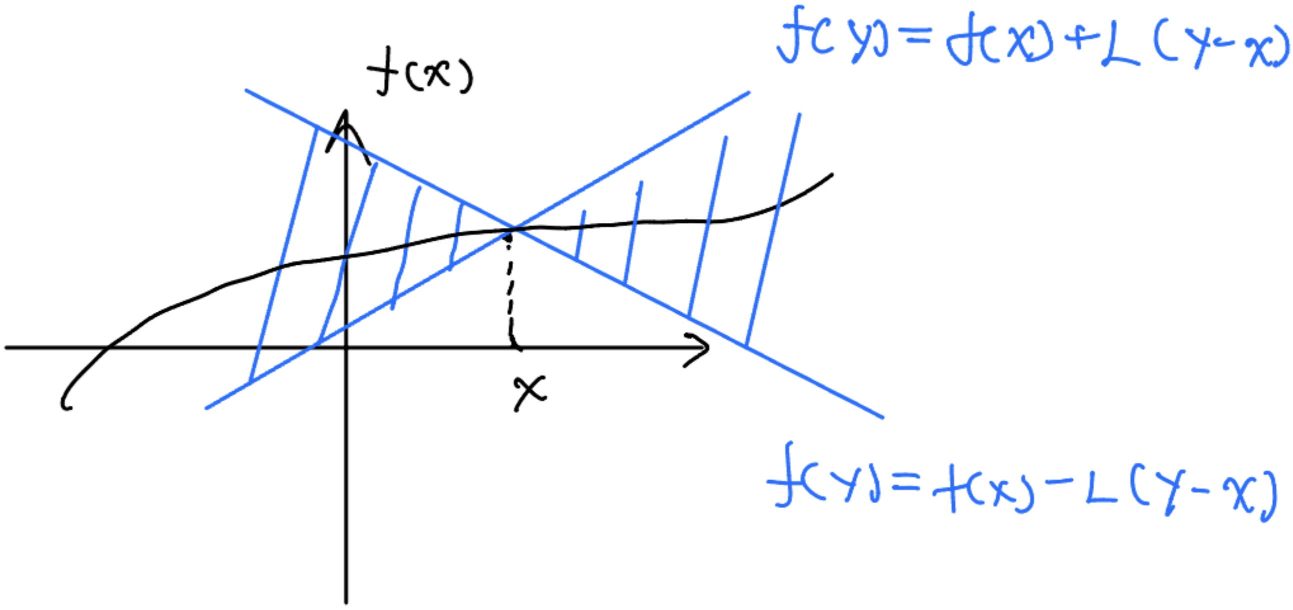
**Definition (*Lipschitz continuity*)**

A function $f : \mathbb{R}^n \to \mathbb{R}$ is $L$-Lipschitz, if for all $x, y \in \operatorname{dom} f$,

$$\|f(x) - f(y)\| \le L\|x - y\| .$$

We usually use $L^2$-norm, unless otherwise specified.

An $L$-Lipschitz function is continuous, but may not be differentiable. Intuitively, for a Lipschitz continuous function, there exists a double cone (white) whose origin can be moved along the graph so that the whole graph always stays outside the double cone.




**Example**

- $f(x) = kx$ where $x \in \mathbb{R}$ is $|k|$-Lipschitz.
- $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$ where $\boldsymbol{x} \in \mathbb{R}^n$ is $\|\boldsymbol{w}\|$-Lipschitz
- $f(\boldsymbol{x}) = Q\boldsymbol{x}$ where $\boldsymbol{x} \in \mathbb{R}^n$ is $\lambda_{\max}(Q^\top Q)^{1/2}$-Lipschitz, since

$$\begin{aligned}\|f(\boldsymbol{x}) - f(\boldsymbol{y})\| = \|Q(\boldsymbol{x} - \boldsymbol{y})\| &= \left((\boldsymbol{x} - \boldsymbol{y})^\top Q^\top Q(\boldsymbol{x} - \boldsymbol{y})\right)^{1/2} \\ &\le \lambda_{\max}(Q^\top Q)^{1/2}\|\boldsymbol{x} - \boldsymbol{y}\|\end{aligned}$$

  by the bound for the Rayleigh quotient. In particular, if $Q$ is symmetric,

$$\lambda_{\max}(Q^\top Q)^{1/2} = \max\{|\lambda_{\min}(Q)|, |\lambda_{\max}(Q)|\} .$$

Recall that we hope $\nabla f(x)$ does not change rapidly. So we define the following notion of "smoothness".

**Definition (*Smoothness*)**

A function $f : \mathbb{R}^n \to \mathbb{R}$ is $L$-smooth if $\nabla f$ if $L$-Lipschitz, i.e., for all $x, y$,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| .$$

**Example**

$f(\boldsymbol{x}) = \boldsymbol{x}^\top Q\boldsymbol{x}$ with $Q \succeq 0$ is $2\lambda_{\max}(Q)$-smooth ($\nabla f(\boldsymbol{x}) = 2Q\boldsymbol{x}$).

We use the notation $A \succeq B$ if $A - B \succeq 0$. Then we have the following equivalent definitions.

**Lemma**

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a twice differentiable function. Then $f$ is $L$-smooth iff $-L\boldsymbol{I}_n \preceq \nabla^2 f(\boldsymbol{x}) \preceq L\boldsymbol{I}_n$ for all $\boldsymbol{x} \in \mathbb{R}^n$, where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. Namely, for all $\boldsymbol{x} \in \mathbb{R}^n$, $\left|\lambda_i(\nabla^2 f(\boldsymbol{x}))\right| \leq L$, where $\lambda_1, \ldots, \lambda_n$ are $n$ eigenvalues.

Note that if $f : \mathbb{R} \to \mathbb{R}$, we can easily prove the " $\Longleftarrow$ " direction since the mean value theorem gives that $f'(x) - f'(y) = f''(z)(x - y)$ for some $z$. However, there is no such theorem for vector-valued functions.

**Proof**

- " $\Longleftarrow$ " direction. We would like to restrict the vector-valued function $\nabla f$ to a line. Fix any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Let $\varphi : [0, 1] \to \mathbb{R}$ be a function defined by

$$\varphi(t) = \left\langle \nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}), \nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))\right\rangle .$$

Then, $\varphi(1) = \langle \nabla f(\boldsymbol{y}), \nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\rangle$ and $\varphi(0) = \langle \nabla f(\boldsymbol{x}), \nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\rangle$. By the mean value theorem, there exists $t \in [0, 1]$ such that $\varphi(1) - \varphi(0) = \varphi'(t)$. Note that

$$\begin{aligned}\varphi'(t) &= \left\langle \nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}),\ \nabla^2 f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))(\boldsymbol{y} - \boldsymbol{x})\right\rangle \\ &\leq \|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\| \cdot \|\nabla^2 f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))(\boldsymbol{y} - \boldsymbol{x})\|\end{aligned}$$

by the Cauchy-Schwarz inequality. It implies that

$$\begin{aligned}\|\nabla f(\boldsymbol{y})-\nabla f(\boldsymbol{x})\|^2 &= \varphi(1)-\varphi(0)\\ &\le \|\nabla f(\boldsymbol{y})-\nabla f(\boldsymbol{x})\|\cdot\|\nabla^2 f(\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x}))(\boldsymbol{y}-\boldsymbol{x})\|\,,\end{aligned}$$

which further gives that

$$\|\nabla f(\boldsymbol{y})-\nabla f(\boldsymbol{x})\| \le \|\nabla^2 f(\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x}))(\boldsymbol{y}-\boldsymbol{x})\| \le L\|\boldsymbol{y}-\boldsymbol{x}\|\,.$$

The last inequality follows from the third example of Lipschitz functions.

- " $\Longrightarrow$ " direction. Fix any $\boldsymbol{x},\boldsymbol{v}\in\mathbb{R}^n$. Let $\psi:\mathbb{R}_{\ge 0}\to\mathbb{R}$ be a function defined by

$$\psi(t)=\langle\nabla f(\boldsymbol{x}+t\boldsymbol{v}),\,\boldsymbol{v}\rangle\,.$$

Then, by the Cauchy-Schwarz inequality and the $L$-smoothness, we have

$$\begin{aligned}|\psi(t)-\psi(0)| &= \left|\langle\nabla f(\boldsymbol{x}+t\boldsymbol{v})-\nabla f(\boldsymbol{x}),\,\boldsymbol{v}\rangle\right|\\ &\le \|\nabla f(\boldsymbol{x}+t\boldsymbol{v})-\nabla f(\boldsymbol{x})\|\cdot\|v\|\\ &\le tL\|v\|^2\,,\end{aligned}$$

which further gives that $\left|\frac{\psi(t)-\psi(0)}{t}\right|\le L\|\boldsymbol{v}\|^2$. Taking the limit $t\to 0$ on both sides, and applying the chain rule, we obtain that

$$\left|\boldsymbol{v}^\top\nabla^2 f(\boldsymbol{x})\boldsymbol{v}\right| = \left|\psi'(0)\right| \le L\|\boldsymbol{v}\|^2\,.$$

Thus, $-L\boldsymbol{I}_n\preceq\nabla^2 f(\boldsymbol{x})\preceq L\boldsymbol{I}_n$.

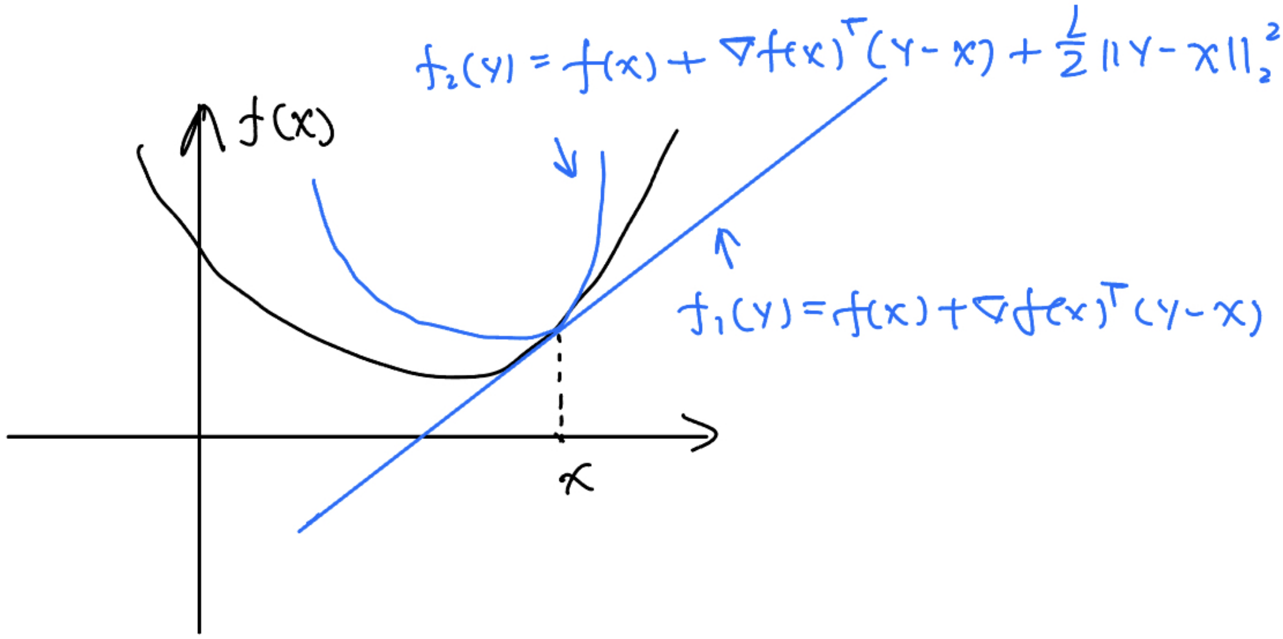
An $L$-smooth functions may be not convex. If $f$ is further convex, all absolute values are not necessary.

**Lemma**

Suppose $f:\mathbb{R}^n\to\mathbb{R}$ is a differentiable function. Then $f$ is $L$-smooth iff for all $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n$,

$$\left|f(\boldsymbol{y})-f(\boldsymbol{x})-\langle\nabla f(\boldsymbol{x}),\,\boldsymbol{y}-\boldsymbol{x}\rangle\right| \le \frac{L}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2\,.$$

Recall that $f$ is convex iff $f(\boldsymbol{y})-f(\boldsymbol{x})-\langle\nabla f(\boldsymbol{x}),\,\boldsymbol{y}-\boldsymbol{x}\rangle\ge 0$, which shows that $f$ is underestimated by an affine function. Now, if $f$ is $L$-smooth, it is overestimated by a quadratic function.




**Proof**

- " $\Longleftarrow$ " direction. Fix $\boldsymbol{x} \in \mathbb{R}^n$. Define

$$g_1(\boldsymbol{y}) = f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}),\, \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2\,,$$
$$g_2(\boldsymbol{y}) = f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}),\, \boldsymbol{y} - \boldsymbol{x} \rangle - \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2\,.$$

Note that for all $\boldsymbol{y} \in \mathbb{R}^n$, $g_2(\boldsymbol{y}) \le 0 \le g_1(\boldsymbol{y})$, and $g_1(\boldsymbol{x}) = g_2(\boldsymbol{x}) = 0$. So $\boldsymbol{x}$ is a local minimum point of $g_1$, which gives that $\nabla^2 g_1(\boldsymbol{x}) \succeq 0$. Since $\nabla^2 g_1(\boldsymbol{y}) = \nabla^2 f(\boldsymbol{y}) + L\boldsymbol{I}_n$, we conclude that $\nabla^2 f(\boldsymbol{x}) \succeq -L\boldsymbol{I}_n$. Similarly, $\boldsymbol{x}$ is a local maximum point of $g_2$, and thus $\nabla^2 g_2(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x}) - L\boldsymbol{I}_n \preceq 0$.

- " $\Longrightarrow$ " direction. Fix $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Let

$$h(\theta) = f(\boldsymbol{x} + \theta(\boldsymbol{y} - \boldsymbol{x}))\,.$$

It is clear that $h'(\theta) = \langle \nabla f(\boldsymbol{x} + \theta(\boldsymbol{y} - \boldsymbol{x})), \boldsymbol{y} - \boldsymbol{x} \rangle$, and

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) = h(1) - h(0) = \int_0^1 h'(\theta)\,\mathrm{d}\theta\,.$$

Moreover, $\langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle = h'(0) = \int_0^1 h'(0)\,\mathrm{d}\theta$. Therefore, it holds that

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}),\, \boldsymbol{y} - \boldsymbol{x} \rangle = \int_0^1 h'(\theta) - h'(0)\,\mathrm{d}\theta\,.$$

Note that

$$\begin{aligned} \left|h'(\theta) - h'(0)\right| &= |\nabla f(\boldsymbol{x} + \theta(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x})), \boldsymbol{y} - \boldsymbol{x}| \\ &\le \|\nabla f(\boldsymbol{x} + \theta(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x})\| \cdot \|\boldsymbol{y} - \boldsymbol{x}\| \\ &\le \theta L \|\boldsymbol{y} - \boldsymbol{x}\|^2 . \end{aligned}$$

We now have

$$\begin{aligned} \left|\langle f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle\right| &\le \int_0^1 \left|h'(\theta) - h'(0)\right| \mathrm{d}\theta \\ &\le \int_0^1 \theta L \|\boldsymbol{y} - \boldsymbol{x}\|^2 \,\mathrm{d}\theta = \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 , \end{aligned}$$

which completes the proof.

Recall that, we hope to find the value of the step size $t$ such that $f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$. Now we assume that $f$ is $L$-smooth. Then

$$\begin{aligned} f(\boldsymbol{x}_{k+1}) &= f(\boldsymbol{x}_k - t \cdot \nabla f(\boldsymbol{x}_k)) \\ &\le f(\boldsymbol{x}_k) - \langle \nabla f(\boldsymbol{x}_k), t \cdot \nabla f(\boldsymbol{x}_k)\rangle + \frac{L}{2}\|t \cdot \nabla f(\boldsymbol{x}_k)\|^2 \\ &= f(\boldsymbol{x}_k) - t\left(1 - \frac{Lt}{2}\right)\|\nabla f(\boldsymbol{x}_k)\|^2 \\ &< f(\boldsymbol{x}_k) \end{aligned}$$

if we set $t < 2/L$. In particular, if we choose $t \le 1/L$, it gives the following *descent lemma*.

**Lemma (*Descent lemma*)**

For an $L$-smooth differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ (not necessarily convex), and $t \le 1/L$, we have

$$f(\boldsymbol{x}_{k+1}) \le f(\boldsymbol{x}_k) - \frac{t}{2}\|\nabla f(\boldsymbol{x}_k)\|^2 .$$