**AM 221: Advanced Optimization** Spring 2016

*Prof. Yaron Singer* *Lecture 8 — February 22nd*

# 1 Overview

In the previous lecture we saw characterizations of optimality in linear optimization, and we reviewed the simplex method for finding optimal solutions. In this lecture we return to convex optimization and give a characterization of optimality. The main theorem we will prove today shows that for a convex function $f : S \to \mathbb{R}$ defined over a convex set $S$, a stationary point (the point $\bar{\mathbf{x}}$ for which $\nabla f(\bar{\mathbf{x}})$) is a global minimum. We we will review basic definitions and properties from multivariate calculus, prove some important properties of convex functions, and obtain various characterizations of optimality.

# 2 A Characterization of Convex Functions

Recall the definitions we introduced in the second lecture of convex sets and convex functions:

**Definition.** *A set $S$ is called a **convex set** if any two points in $S$ contain their line, i.e. for any $\mathbf{x}_1, \mathbf{x}_2 \in S$ we have that $\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2 \in S$ for any $\lambda \in [0,1]$.*

**Definition.** *For a convex set $S \subseteq R^n$, we say that a function $f : S \to \mathbb{R}$ is **convex on** $S$ if for any two points $\mathbf{x}_1, \mathbf{x}_2 \in S$ and any $\lambda \in [0,1]$ we have that:*

$$f\left(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2\right) \leq \lambda f(\mathbf{x}_1) + \Big(1-\lambda\Big) f(\mathbf{x}_2).$$

We will first give an important characterization of convex function. To so, we need to characterize multivariate functions via their Taylor expansion.

## 2.1 First-order Taylor approximation

**Definition.** *The **gradient** of a function $f : \mathbb{R}^n \to \mathbb{R}$ at point $\mathbf{x} \in \mathbb{R}^n$ denoted $\nabla f(\bar{\mathbf{x}})$ is:*

$$\nabla f(x) := \left[\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n}\right]^\intercal$$

**First-order Taylor expansion.** The first order Taylor expansion of a function is:

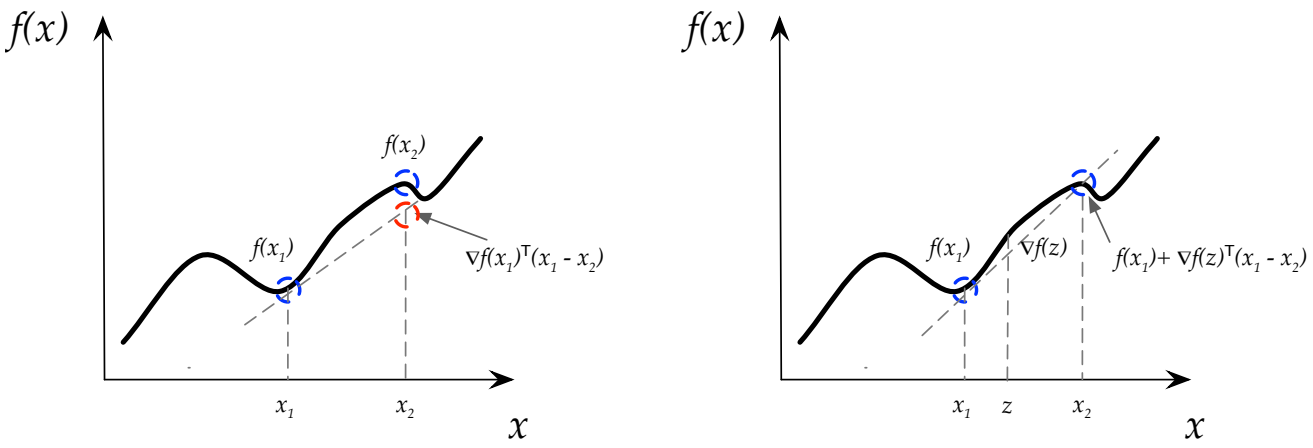


Figure 1: Depiction of first-order taylor expansion.

- For a function $f : \mathbb{R} \to \mathbb{R}$ of a single variable, differentiable at $\bar{\mathbf{x}} \in \mathbb{R}$, we write:

$$\forall x \in \mathbb{R},\ f(x) = f(\bar{\mathbf{x}}) + f'(\bar{\mathbf{x}})(x - \bar{\mathbf{x}}) + o(||x - \bar{x}||)$$

  where by definition:

$$\lim_{\bar{x} \to x} \frac{o(||x - \bar{x}||)}{||x - \bar{x}||} = 0$$

- Similarly, when $f : \mathbb{R}^n \to \mathbb{R}$ is a multivariate function, differentiable at $\bar{\mathbf{x}} \in \mathbb{R}^n$, we write:

$$\forall \mathbf{x} \in \mathbb{R}^n,\ f(x) = f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^\intercal(\mathbf{x} - \bar{\mathbf{x}}) + o(\|x - \bar{\mathbf{x}}\|)$$

- If $f$ is continuous over $[\bar{\mathbf{x}}, \mathbf{x}]$ and differentiable over $(\bar{\mathbf{x}}, \mathbf{x})$, then:

$$f(\mathbf{x}) = f(\bar{\mathbf{x}}) + \nabla f(\mathbf{z})^\intercal(\mathbf{x} - \bar{\mathbf{x}}), \quad \text{for some } \mathbf{z} \in [\bar{\mathbf{x}}, \mathbf{x}].$$

This characterization implies that the function $f$ can be approximated by the affine function $l : x \mapsto f(\bar{\mathbf{x}}) + f'(\bar{\mathbf{x}})(x - \bar{\mathbf{x}})$ when $x$ is "close to" $\bar{\mathbf{x}}$.

## 2.2 Directional derivatives

**Definition.** *Let $f : \mathbb{R}^n \to R$ be a function differentiable at $x \in \mathbb{R}^n$ and let us consider $d \in \mathbb{R}^n$ with $\|d\| = 1$. We define the* ***derivative of $f$ at $\mathbf{x}$ in direction $\mathbf{d}$*** *as:*

$$f'(\mathbf{x}, \mathbf{d}) = \lim_{\mathbf{x} \to \mathbf{0}} \frac{f(\mathbf{x} + \lambda \mathbf{d}) - f(\mathbf{x})}{\lambda}$$

**Claim 1.** $f'(\mathbf{x}, \mathbf{d}) = \nabla f(\mathbf{x})^\intercal \mathbf{d}$.

*Proof.* Using the first order expansion of $f$ at $\mathbf{x}$:

$$f(\mathbf{x} + \lambda \mathbf{d}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal(\lambda \mathbf{d}) + o(\|\lambda \mathbf{d}\|)$$

hence, dividing by $\lambda$ (and noticing that $\|\lambda \mathbf{d}\| = \lambda^2 \|\mathbf{d}\|$):

$$\frac{f(\mathbf{x} + \lambda \mathbf{d}) - f(\mathbf{x})}{\lambda} = \nabla f(\mathbf{x})^\intercal \mathbf{d} + o(\lambda \|\mathbf{d}\|)$$

letting $\lambda$ go to 0 concludes the proof. □

### 2.3 Lower bounding convex functions with affine functions

In order to prove the characterization of convex functions in the next section we will need the following lemma. This lemma says that any convex function can essentially be underestimated by an affine function.

**Theorem 2.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $S$ be a convex subset of $\mathbb{R}^n$. Then, $f$ is convex if and only if for any $\mathbf{x}, \mathbf{y} \in S$ we have that $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal (\mathbf{y} - \mathbf{x})$.*

*Proof.* [ $\Longrightarrow$ ] Assume $f$ is convex, and let $\mathbf{z} = \lambda \mathbf{y} + (1 - \lambda)\mathbf{x}$ for some $\mathbf{x}, \mathbf{y} \in S$ and $\lambda \in [0, 1]$. From convexity of $f$ we have that:

$$f(\mathbf{z}) = f(\lambda \mathbf{y} + (1 - \lambda)\mathbf{x}) \leq \lambda f(\mathbf{y}) + (1 - \lambda) f(\mathbf{x})$$

and therefore by subtracting $f(\mathbf{x})$ from both sides we get:

$$\begin{aligned} f\left(\mathbf{x} + \lambda(\mathbf{y} - \mathbf{x})\right) - f(\mathbf{x}) &= f\left(\lambda \mathbf{y} + (1 - \lambda)\mathbf{x}\right) - f(\mathbf{x}) \\ &= f(\mathbf{z}) - f(\mathbf{x}) \\ &\leq \lambda f(\mathbf{y}) + (1 - \lambda) f(\mathbf{x}) - f(\mathbf{x}) \\ &= \lambda f(\mathbf{y}) - \lambda f(\mathbf{x}). \end{aligned}$$

Thus we get that (for $\lambda > 0$):

$$\frac{f\left(\mathbf{x} + \lambda(\mathbf{y} - \mathbf{x})\right) - f(\mathbf{x})}{\lambda} \leq f(\mathbf{y}) - f(\mathbf{x})$$

Applying Claim 1 on $\mathbf{d} = \mathbf{x} - \mathbf{y}$ we have that:

$$\nabla f(\mathbf{x})^\intercal (\mathbf{y} - \mathbf{x}) = \lim_{\lambda \to 0^+} \frac{f(\mathbf{x} + \lambda(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})}{\lambda} \leq f(\mathbf{y}) - f(\mathbf{x}).$$

[ $\Longleftarrow$ ] Assume that $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal (\mathbf{y} - \mathbf{x})$ for any $\mathbf{x}, \mathbf{y} \in S$ and show that $f$ is convex. Let $\mathbf{x}, \mathbf{y} \in S$ and $\mathbf{z} = \lambda \mathbf{y} + (1 - \lambda)\mathbf{x}$. By our assumption we have that:

$$f(\mathbf{y}) \geq f(\mathbf{z}) + \nabla f(\mathbf{z})^\intercal (\mathbf{y} - \mathbf{z}) \tag{1}$$

$$f(\mathbf{x}) \geq f(\mathbf{z}) + \nabla f(\mathbf{z})^\intercal (\mathbf{x} - \mathbf{z}) \tag{2}$$

Multiplying (1) by $\lambda$ and adding $(1 - \lambda) \times (2)$ gives us:

$$\begin{aligned} \lambda f(\mathbf{y}) + (1 - \lambda) f(\mathbf{x}) &\geq \lambda f(\mathbf{z}) + \lambda \nabla f(\mathbf{z})^\intercal (\mathbf{y} - \mathbf{z}) + (1 - \lambda) f(\mathbf{z}) + (1 - \lambda) \nabla f(\mathbf{z})^\intercal (\mathbf{x} - \mathbf{z}) \\ &= f(\mathbf{z}) + \nabla f(\mathbf{z})^\intercal (\lambda \mathbf{y} - \lambda \mathbf{z}) + \nabla f(\mathbf{z})^\intercal \Big( (1 - \lambda)\mathbf{x} - (1 - \lambda)\mathbf{z} \Big) \\ &= f(\mathbf{z}) + \nabla f(\mathbf{z})^\intercal \Big( \lambda \mathbf{y} + (1 - \lambda)\mathbf{x} - \mathbf{z} \Big) \\ &= f(\lambda \mathbf{y} + (1 - \lambda)\mathbf{x}) \end{aligned}$$

since $\lambda \mathbf{y} + (1 - \lambda)\mathbf{x} - \mathbf{z} = 0$. This is exactly the definition of convexity. □

# 3 Conditions for optimality

**Definition.** *A point $\bar{\mathbf{x}} \in \mathbb{R}^n$ at which $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$ is called a **stationary point**.*

## 3.1 Necessary conditions

We now want to find necessary and sufficient conditions for local optimality.

**Claim 3.** *If $\bar{\mathbf{x}}$ is a local extremum of a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, then $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$.*

*Proof.* Let us assume that $\bar{\mathbf{x}}$ is a local minimum for $f$. Then for all $\mathbf{d} \in \mathbb{R}^n$, $f(\bar{\mathbf{x}}) \leq f(\bar{\mathbf{x}} + \lambda \mathbf{d})$ for $\lambda$ small enough. Hence:

$$0 \leq f(\bar{\mathbf{x}} + \lambda \mathbf{d}) - f(\bar{\mathbf{x}}) = \lambda \nabla f(\bar{\mathbf{x}})^\intercal \mathbf{d} + o(\|\lambda \mathbf{d}\|)$$

dividing by $\lambda > 0$ and letting $\lambda \to 0^+$, we obtain $0 \leq \nabla f(\bar{\mathbf{x}})^\intercal \mathbf{d}$. Similarly, dividing by $\lambda < 0$ and letting $\lambda \to 0^-$, we obtain $0 \geq \nabla f(\bar{\mathbf{x}})^\intercal \mathbf{d}$. As a consequence, $\nabla f(\bar{\mathbf{x}})^\intercal \mathbf{d} = 0$ for all $\mathbf{d} \in \mathbb{R}^n$. This implies that $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$.

The case where $\bar{\mathbf{x}}$ is a local maximum can be dealt with similarly. □

The above claim states that a a stationary point can either be a minimum, maximum, or a saddle point of the function, and we will see in the following section that the Hessian of the function can be used to indicate which one exactly. For convex functions however it turns out that a stationary point necessarily implies that the function is at its minimum. Note that with the above Claim, this says that for a convex function a point is optimal if and only if it is stationary.

**Proposition 4.** *Let $S \subseteq \mathbb{R}^n$ be a convex set and $f : S \to \mathbb{R}^n$ be a convex function. If $\bar{\mathbf{x}}$ a stationary point then $\bar{\mathbf{x}}$ is a global minimum.*

*Proof.* From Theorem 2 we know that for any $\mathbf{x}, \mathbf{y} \in S$ : $f(\mathbf{y}) \geq f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})(\mathbf{y} - \mathbf{x})$. Since $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$ this implies that $f(\mathbf{y}) \geq f(\bar{\mathbf{x}})$. As this holds for any $\mathbf{y} \in S$, $\bar{\mathbf{x}}$ is a global minimum. □

# 4 Characterizations of Convexity and Optimality via the Hessian

We will now use the Hessian of a function to obtain characterizations of convexity and optimality. We will begin by defining the *Hessian* of a function and then see that it plays a role in approximating a function via a second-order Taylor expansion. We will then use *semi-definiteness* of the Hessian matrix to characterize both conditions of optimality as well as the convexity of a function.

**Definition.** *Given a function $f : \mathbb{R}^n \to \mathbb{R}$ its **Hessian** matrix at point $\mathbf{x} \in \mathbb{R}^n$ denoted $H_f(\mathbf{x})$ (also sometimes denoted $\nabla^2 f(\mathbf{x})$) is:*

$$H_f(\mathbf{x}) := \begin{bmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{bmatrix}$$

**Second-order Taylor expansion.** When $f$ is twice differentiable it is possible to obtain an approximation of $f$ by quadratic functions.

- Assume that $f : \mathbb{R}^n \to R$ is twice differentiable at $\bar{\mathbf{x}} \in \mathbb{R}^n$, then:

  $$\forall \mathbf{x} \in \mathbb{R}^n, f(\mathbf{x}) = f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^\intercal (x - \bar{\mathbf{x}}) + \frac{1}{2}(x - \bar{\mathbf{x}})^\intercal H_f(\bar{\mathbf{x}})(x - \bar{\mathbf{x}}) + o(\|x - \bar{\mathbf{x}}\|^2)$$

  where by definition:

  $$\lim_{x \to \bar{\mathbf{x}}} \frac{o(\|x - \bar{\mathbf{x}}\|^2)}{\|x - \bar{\mathbf{x}}\|^2} = 0$$

- Similarly to the first-order alternative expansion, we have, when $f$ is of class $C^1$ over $[\bar{\mathbf{x}}, \mathbf{x}]$ and twice differentiable over this interval:

  $$\forall \mathbf{x} \in \mathbb{R}^n, f(\mathbf{x}) = f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^\intercal (\mathbf{x} - \bar{\mathbf{x}}) + \frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}})^\intercal H_f(\mathbf{z})(\mathbf{x} - \bar{\mathbf{x}}) \quad \text{for some } \mathbf{z} \in [\bar{\mathbf{x}}, \mathbf{x}].$$

**Semi-definiteness of a matrix.** The following classification of symmetric matrices will be useful.

**Definition.** *Let $A$ by a symmetric matrix in $\mathbb{R}^{n \times n}$. We say that $A$ is:*

1. positive definite *iff* $\mathbf{x}^\intercal A \mathbf{x} > 0$ *for all* $x \in \mathbb{R}^n \setminus \{0\}$;
2. negative definite *iff* $\mathbf{x}^\intercal A \mathbf{x} < 0$ *for all* $x \in \mathbb{R}^n \setminus \{0\}$;
3. positive semidefinite *iff* $\mathbf{x}^\intercal A \mathbf{x} \geq 0$ *for all* $x \in \mathbb{R}^n$;
4. negative semidefinite *iff* $\mathbf{x}^\intercal A \mathbf{x} \leq 0$ *for all* $x \in \mathbb{R}^n$;
5. *Otherwise we say that $A$ is* indefinite.

**Example: indefinite matrix.** Consider the following matrix $A$:

$$A := \begin{bmatrix} +4 & -1 \\ -1 & -2 \end{bmatrix}$$

Then we have $\mathbf{x}^\intercal A \mathbf{x} = 4x_1^2 - 2x_1 x_2 - 2x_2^2$. For $\mathbf{x} = (1\ 0)$, we have $x^\intercal A \mathbf{x} = 4 > 0$. For $\mathbf{x} = (0\ 1)$ we have $\mathbf{x}^\intercal A \mathbf{x} = -2 < 0$. $A$ is therefore indefinite.

The following theorem gives a useful characterization of (semi)definite matrices.

**Theorem 5.** *Let $A$ be a symmetric matrix in $\mathbb{R}^{n\times n}$.*

1. *$A$ is positive (negative) definite iff all its eigenvalues are positive (negative);*
2. *$A$ is positive (negative) semidefinite iff all its eigenvalues are non-negative (non-positive);*
3. *Otherwise $A$ is indefinite.*

*Proof.* The spectral theorem tells us that any real symmetric matrix is diagonalizable. This allows us to diagonalize $A$: we can write $A = P^\intercal DP$ where $D$ is a diagonal matrix. Then one can write $\mathbf{x}^\intercal A\mathbf{x} = \mathbf{x}^\intercal P^\intercal DP\mathbf{x} = (P\mathbf{x})^\intercal D(P\mathbf{x})$. Let us define $\mathbf{y} := P\mathbf{x}$, then we have:

$$\mathbf{x}^\intercal A\mathbf{x} = \sum_{i=1}^{n} \lambda_i y_i^2$$

where $\lambda_1, \ldots, \lambda_n$ are the diagonal elements of $D$, the eigenvalues of $A$. Also note that since $P$ is invertible, $\mathbf{y} = 0$ iff $\mathbf{x} = \mathbf{0}$. If $\lambda_i > 0$ for all $i$, then we see that $\mathbf{x}^\intercal A\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$. For the other direction, choosing $\mathbf{y}$ such that $y_i = 1$ and $y_j = 0$ for $j \neq i$ yields $\lambda_i > 0$.

The other cases can be dealt with similarly. □

### 4.1 Necessary and sufficient conditions for local extrema

We can now give the second-order necessary conditions for local extrema via the Hessian.

**Theorem 6.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function twice differentiable at $\bar{\mathbf{x}} \in \mathbb{R}^n$.*

1. *If $\bar{\mathbf{x}}$ is a local minimum, then $H_f(\bar{\mathbf{x}})$ is positive semidefinite.*
2. *If $\bar{\mathbf{x}}$ is a local maximum, then $H_f(\bar{\mathbf{x}})$ is negative semidefinite.*

*Proof.* Let us assume that $\bar{\mathbf{x}}$ is a local minimum. We know from Theorem 3 that $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$, hence the second-order expansion at $\bar{\mathbf{x}}$ takes the form:

$$f(\bar{\mathbf{x}} + \mathbf{d}) = f(\bar{\mathbf{x}}) + \frac{1}{2} d^\intercal H_f(\bar{\mathbf{x}})\mathbf{d} + o(\|d\|^2)$$

Because $\bar{\mathbf{x}}$ is a local minimum, when $\|\mathbf{d}\|$ is small enough:

$$0 \leq f(\bar{\mathbf{x}} + \mathbf{d}) - f(\bar{\mathbf{x}}) \sim_{\|d\|\to 0} \frac{1}{2}\mathbf{d}^\intercal H_f(\bar{\mathbf{x}})\mathbf{d}$$

So $\mathbf{d}^\intercal H_f(\bar{\mathbf{x}})$ must be non-negative when $\|d\|$ is small enough. This is true for any such $\mathbf{d}$, hence $H_f(\bar{\mathbf{x}})$ is positive semidefinite. □

Combining the conditions of Theorem 3 (stationary point) and Theorem 5 is almost enough to obtain a sufficient condition for local extrema. We only to slightly strengthen the conditions of Theorem 5.

**Theorem 7.** *Let $f$ be a function twice differentiable at a stationary point $\bar{\mathbf{x}}$:*

1. *if $H_f(\bar{\mathbf{x}})$ is positive definite then $\bar{\mathbf{x}}$ is a local minimum.*
2. *if $H_f(\bar{\mathbf{x}})$ is negative definite then $\bar{\mathbf{x}}$ is a local maximum.*

*Proof.* Let us assume that $H_f(\bar{\mathbf{x}})$ is positive definite. We know that $\bar{\mathbf{x}}$ is a stationary point. We can write the second-order expansion at $\bar{\mathbf{x}}$:

$$f(\bar{\mathbf{x}} + \mathbf{d}) = f(\bar{\mathbf{x}}) + \frac{1}{2} d^\intercal H_f(x)\mathbf{d} + o(\|\mathbf{d}\|^2)$$

When $\|\mathbf{d}\|$ is small enough, $f(\bar{\mathbf{x}} + \mathbf{d}) - f(\bar{\mathbf{x}})$ has the same sign as $\frac{1}{2} d^\intercal H_f(x)\mathbf{d}$. But because $H_f(\bar{\mathbf{x}})$ is positive definite, for any $\mathbf{d} \neq \mathbf{0}$ small enough, $f(\bar{\mathbf{x}} + \mathbf{d}) - f(\bar{\mathbf{x}}) > 0$. This proves that $\bar{\mathbf{x}}$ is a local minimum.

We can make a similar proof when $H_f(\bar{\mathbf{x}})$ is negative definite. □

*Remark* 8. When $H_f(\bar{\mathbf{x}})$ is indefinite at a stationary point $\bar{\mathbf{x}}$, we have what is known as a *saddle point*: $\bar{\mathbf{x}}$ will be a minimum along the eigenvectors of $H_f(\bar{\mathbf{x}})$ for which the eigenvalues are positive and a maximum along the eigenvectors of $H_f(\bar{\mathbf{x}})$ for which the eigenvalues are negative.

## 4.2 Characterization of convexity

**Theorem 9.** *Let $S \subseteq \mathbb{R}^n$ be convex, and $f : S \to \mathbb{R}$ twice continuous differentiable on $S$.*

1. *If $H_f(\mathbf{x})$ is positive semi-definite for any $\mathbf{x} \in S$ then $f$ is convex on $S$.*
2. *If $H_f(\mathbf{x})$ is positive definite for any $\mathbf{x} \in S$ then $f$ is **strictly** convex on $S$.*
3. *If $S$ is open and $f$ is convex, then $H_f(\mathbf{x})$ is positive semi-definite $\forall \mathbf{x} \in S$.*

*Proof.*

1. From the tailor expansion of $f$ we know that for some $\mathbf{z} \in [\mathbf{x}, \mathbf{y}]$:

   $$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{1}{2}\Big((\mathbf{y} - \mathbf{x})^T H_f(\mathbf{z})(\mathbf{y} - \mathbf{x})\Big)$$

   If $H_f(\mathbf{z})$ is positive semi-definite, this necessarily implies that:

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

   and from Lemma 2 we get that $f$ is convex.

2. if $H_f(\mathbf{x})$ is positive definite, we have that:

   $$f(\mathbf{y}) > f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}).$$

   Applying the same idea as in Lemma 2 we can show that in this case $f$ is **strictly** convex.

3. Let $f$ be a convex function and assume $S$ is open. For $\mathbf{x} \in S$, and some small $\lambda > 0$, for any $\mathbf{d} \in \mathbb{R}^n$ we have that $\mathbf{x} + \lambda\mathbf{d} \in S$. From the Taylor expansion of $f$ we get:

$$f(\mathbf{x} + \lambda\mathbf{d}) = f(\mathbf{x}) + \lambda\nabla f(\mathbf{x})^T\mathbf{d} + \frac{\lambda^2}{2}\mathbf{d}^T H_f(\mathbf{x})\mathbf{d} + o(||\lambda\mathbf{d}||^2)$$

From Lemma 2 we get that if $f$ is convex then:

$$f(\mathbf{x} + \lambda\mathbf{d}) \geq f(\mathbf{x}) + \lambda\nabla f(\mathbf{x})^T\mathbf{d}.$$

Therefore, we have that for any $\mathbf{d} \in \mathbb{R}^n$:

$$\frac{\lambda^2}{2}\mathbf{d}^T H_f(\mathbf{x})\mathbf{d} + o(||\lambda\mathbf{d}||^2) \geq 0$$

Dividing by $\lambda^2$ and taking $\lambda \to 0^+$ gives us that for any $\mathbf{d} \in \mathbb{R}^n$: $\mathbf{d}^T H_f(\mathbf{x})\mathbf{d} \geq 0$. □

*Remark* 10. It is important to note that if $S$ is open and $f$ is strongly convex, then $H_f(\mathbf{x})$ is still (only) positive semi-definite $\forall \mathbf{x} \in S$. Consider $f(x) = x^4$ which is strongly convex, then the Hessian is $H_f(x) = 12x^2$ which equals 0 at $x = 0$.