CHAPTER 6

Optimality Conditions for Unconstrained Problems

1. Unconstrained Optimization

1.1. Existence. Consider the problem of minimizing the function $f : \mathbb{R}^n \to \mathbb{R}$ where f is continuous on all of $\mathbb{R}^n$:

$$\mathcal{P} \qquad \min_{x\in\mathbb{R}^n} f(x).$$

As we have seen, there is no guarantee that f has a minimum value, or if it does, it may not be attained. To clarify this situation, we examine conditions under which a solution is guaranteed to exist. Recall that we already have at our disposal a rudimentary existence result for constrained problems. This is the Weierstrass Extreme Value Theorem.

THEOREM 1.1. *(WEIERSTRASS EXTREME VALUE THEOREM) Every continuous function on a compact set attains its extreme values on that set.*

We now build a basic existence result for unconstrained problems based on this theorem. For this we make use of the notion of a coercive function.

DEFINITION 1.1. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be coercive if for every sequence $\{x^\nu\} \subset \mathbb{R}^n$ for which $\|x^\nu\| \to \infty$ it must be the case that $f(x^\nu) \to +\infty$ as well.*

Continuous coercive functions can be characterized by an underlying compactness property on their lower level sets.

THEOREM 1.2. *(Coercivity and Compactness) Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuous on all of $\mathbb{R}^n$. The function f is coercive if and only if for every $\alpha \in \mathbb{R}$ the set $\{x \,|\, f(x) \le \alpha\}$ is compact.*

PROOF. We first show that the coercivity of f implies the compactness of the sets $\{x \,|\, f(x) \le \alpha\}$. We begin by noting that the continuity of f implies the closedness of the sets $\{x \,|\, f(x) \le \alpha\}$. Thus, it remains only to show that any set of the form $\{x \,|\, f(x) \le \alpha\}$ is bounded. We show this by contradiction. Suppose to the contrary that there is an $\alpha \in \mathbb{R}^n$ such that the set $S = \{x \,|\, f(x) \le \alpha\}$ is unbounded. Then there must exist a sequence $\{x^\nu\} \subset S$ with $\|x^\nu\| \to \infty$. But then, by the coercivity of f, we must also have $f(x^\nu) \to \infty$. This contradicts the fact that $f(x^\nu) \le \alpha$ for all $\nu = 1, 2, \ldots$. Therefore the set S must be bounded.

Let us now assume that each of the sets $\{x \,|\, f(x) \le \alpha\}$ is bounded and let $\{x^\nu\} \subset \mathbb{R}^n$ be such that $\|x^\nu\| \to \infty$. Let us suppose that there exists a subsequence of the integers $J \subset \mathbb{N}$ such that the set $\{f(x^\nu)\}_J$ is bounded above. Then there exists $\alpha \in \mathbb{R}^n$ such that $\{x^\nu\}_J \subset \{x \,|\, f(x) \le \alpha\}$. But this cannot be the case since each of the sets $\{x \,|\, f(x) \le \alpha\}$ is bounded while every subsequence of the sequence $\{x^\nu\}$ is unbounded by definition. Therefore, the set $\{f(x^\nu)\}_J$ cannot be bounded, and so the sequence $\{f(x^\nu)\}$ contains no bounded subsequence, i.e. $f(x^\nu) \to \infty$. $\square$

This result in conjunction with Weierstrass's Theorem immediately yields the following existence result for the problem $\mathcal{P}$.

THEOREM 1.3. *(Coercivity implies existence) Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuous on all of $\mathbb{R}^n$. If f is coercive, then f has at least one global minimizer.*

PROOF. Let $\alpha \in \mathbb{R}$ be chosen so that the set $S = \{x \,|\, f(x) \le \alpha\}$ is non-empty. By coercivity, this set is compact. By Weierstrass's Theorem, the problem $\min\{f(x) \,|\, x \in S\}$ has at least one global solution. Obviously, the set of global solutions to the problem $\min\{f(x) \,|\, x \in S\}$ is a global solution to $\mathcal{P}$ which proves the result. $\square$

REMARK 1.1. *It should be noted that we only need to know that the coercivity hypothesis is stronger than is strictly required in order to establish the existence of a solution. Indeed, a global minimizer must exist if there exist* *one non-empty compact lower level set. We do not need all of them to be compact. However, in practice, coercivity is easy to check.*

1.2. First-Order Optimality Conditions. This existence result can be quite useful, but unfortunately it does not give us a constructive test for optimality. That is, we may know a solution exists, but we still do not have a method for determining whether any given point may or may not be a solution. We now present such a test using the derivatives of the objective function f. For this we will assume that f is twice continuously differentiable on $\mathbb{R}^n$ and develop constructible first- and second-order necessary and sufficient conditions for optimality.

The optimality conditions we consider are built up from those developed in first term calculus for functions mapping from $\mathbb{R}$ to $\mathbb{R}$. The reduction to the one dimensional case comes about by considering the functions $\phi : \mathbb{R} \to \mathbb{R}$ given by

$$\phi(t) = f(x + td)$$

for some choice of x and d in $\mathbb{R}^n$. The key variational object in this context is the directional derivative of f at a point x in the direction d given by

$$f'(x; d) = \lim_{t \downarrow 0} \frac{f(x + td) - f(x)}{t}.$$

When f is differentiable at the point $x \in \mathbb{R}^n$, then

$$f'(x; d) = \nabla f(x)^T d = \phi'(0).$$

Note that if $f'(x; d) < 0$, then there must be a $\bar{t} > 0$ such that

$$\frac{f(x + td) - f(x)}{t} < 0 \quad \text{whenever} \quad 0 < t < \bar{t}\ .$$

In this case, we must have

$$f(x + td) < f(x) \quad \text{whenever} \quad 0 < t < \bar{t}\ .$$

That is, we can always reduce the function value at x by moving in the direction d an arbitrarily small amount. In particular, if there is a direction d such that $f'(x; d)$ exists with $f'(x; d) < 0$, then x cannot be a local solution to the problem $\min_{x \in \mathbb{R}^n} f(x)$. Or equivalently, if x is a local to the problem $\min_{x \in \mathbb{R}^n} f(x)$, then $f'(x; d) \geq 0$ whenever $f'(x; d)$ exists. We state this elementary result in the following lemma.

LEMMA 1.1 (Basic First-Order Optimality Result). *Let $f : \mathbb{R}^n \to \mathbb{R}$ and let $\overline{x} \in \mathbb{R}^n$ be a local solution to the problem $\min_{x \in \mathbb{R}^n} f(x)$. Then*

$$f'(x; d) \geq 0$$

for every direction $d \in \mathbb{R}^n$ for which $f'(x; d)$ exists.

We now apply this result to the case in which $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable.

THEOREM 1.4. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable at a point $\overline{x} \in \mathbb{R}^n$. If $\overline{x}$ is a local minimum of f, then $\nabla f(\overline{x}) = 0$.*

PROOF. By Lemma 1.1 we have

$$0 \leq f'(\overline{x}; d) = \nabla f(\overline{x})^T d \quad \text{for all} \quad d \in \mathbb{R}^n\ .$$

Taking $d = -\nabla f(\overline{x})$ we get

$$0 \leq -\nabla f(\overline{x})^T \nabla f(\overline{x}) = -\left\|\nabla f(\overline{x})\right\|^2 \leq 0.$$

Therefore, $\nabla f(\overline{x}) = 0$. □

When $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, any point $x \in \mathbb{R}^n$ satisfying $\nabla f(x) = 0$ is said to be a stationary (or, equivalently, a critical) point of f. In our next result we link the notions of coercivity and stationarity.

THEOREM 1.5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable on all of $\mathbb{R}^n$. If f is coercive, then f has at least one global minimizer these global minimizers can be found from among the set of critical points of f.*

PROOF. Since differentiability implies continuity, we already know that f has at least one global minimizer. Differentiabilty implies that this global minimizer is critical. □

This result indicates that one way to find a global minimizer of a coercive differentiable function is to first find all critical points and then from among these determine those yielding the smallest function value.

1.3. Second-Order Optimality Conditions. To obtain second-order conditions for optimality we must first recall a few properties of the Hessian matrix $\nabla^2 f(x)$. The calculus tells us that if f is twice continuously differentiable at a point $x \in \mathbb{R}^n$, then the hessian $\nabla^2 f(x)$ is a symmetric matrix. Symmetric matrices are orthogonally diagonalizable. That is, there exists and orthonormal basis of eigenvectors of $\nabla^2 f(x)$, $v^1, v^2, \ldots, v^n \in \mathbb{R}^n$ such that

$$\nabla^2 f(x) = V \begin{bmatrix} \lambda_1 & 0 & 0 & \ldots & 0 \\ 0 & \lambda_2 & 0 & \ldots & 0 \\ \vdots & & \ddots & \vdots & \\ 0 & 0 & \ldots & \ldots & \lambda_n \end{bmatrix} V^T$$

where $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of $\nabla^2 f(x)$ and V is the matrix whose columns are given by their corresponding vectors $v^1, v^2, \ldots, v^n$:

$$V = \left[v^1, v^2, \ldots, v^n\right] \ .$$

It can be shown that $\nabla^2 f(x)$ is positive semi-definite if and only if $\lambda_i \geq 0,\ i = 1, 2, \ldots, n$, and it is positive definite if and only if $\lambda_i > 0,\ i = 1, 2, \ldots, n$. Thus, in particular, if $\nabla^2 f(x)$ is positive definite, then

$$d^T \nabla^2 f(x) d \geq \lambda_{\min} \|d\|^2 \quad \text{for all } d \in \mathbb{R}^n,$$

where $\lambda_{\min}$ is the smallest eigenvalue of $\nabla^2 f(x)$.

We now give our main result on second-order necessary and sufficient conditions for optimality in the problem $\min_{x \in \mathbb{R}^n} f(x)$. The key tools in the proof are the notions of positive semi-definiteness and definiteness along with the second-order Taylor series expansion for f at a given point $\overline{x} \in \mathbb{R}^n$:

$$f(x) = f(\overline{x}) + \nabla f(\overline{x})^T (x - \overline{x}) + \frac{1}{2}(x - \overline{x})^T \nabla^2 f(\overline{x})(x - \overline{x}) + o(\|x - \overline{x}\|^2) \tag{67}$$

where

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

THEOREM 1.6. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable at the point $\overline{x} \in \mathbb{R}^n$.*

1. *(Necessity) If $\overline{x}$ is a local minimum of f, then $\nabla f(\overline{x}) = 0$ and $\nabla^2 f(\overline{x})$ is positive semi-definite.*
2. *(Sufficiency) If $\nabla f(\overline{x}) = 0$ and $\nabla^2 f(\overline{x})$ is positive definite, then there is an $\alpha > 0$ such that $f(x) \geq f(\overline{x}) + \alpha \|x - \overline{x}\|^2$ for all x near $\overline{x}$.*

PROOF. (1) We make use of the second-order Taylor series expansion (67) and the fact that $\nabla f(\overline{x}) = 0$ by Theorem 1.4. Given $d \in \mathbb{R}^n$ and $t > 0$ set $x := \overline{x} + td$, plugging this into (67) we find that

$$0 \leq \frac{f(\overline{x} + td) - f(\overline{x})}{t^2} = \frac{1}{2} d^T \nabla^2 f(\overline{x}) d + \frac{o(t^2)}{t^2}$$

since $\nabla f(\overline{x}) = 0$ by Theorem 1.4. Taking the limit as $t \to 0$ we get that

$$0 \leq d^T \nabla^2 f(\overline{x}) d.$$

Since d was chosen arbitrarily, $\nabla^2 f(\overline{x})$ is positive semi-definite.

(2) The Taylor expansion (67) and the hypothesis that $\nabla f(\overline{x}) = 0$ imply that

$$\frac{f(x) - f(\overline{x})}{\|x - \overline{x}\|^2} = \frac{1}{2} \frac{(x - \overline{x})^T}{\|x - \overline{x}\|} \nabla^2 f(\overline{x}) \frac{(x - \overline{x})}{\|x - \overline{x}\|} + \frac{o(\|x - \overline{x}\|^2)}{\|x - \overline{x}\|^2}. \tag{68}$$

If $\lambda_{\min} > 0$ is the smallest eigenvalue of $\nabla^2 f(\overline{x})$, choose $\epsilon > 0$ so that

$$\left| \frac{o(\|x - \overline{x}\|^2)}{\|x - \overline{x}\|^2} \right| \leq \frac{\lambda_{\min}}{4} \tag{69}$$

whenever $\|x - \overline{x}\| < \epsilon$. Then, for all $\|x - \overline{x}\| < \epsilon$, we have from (68) and (69) that

$$\begin{aligned} \frac{f(x) - f(\overline{x})}{\|x - \overline{x}\|^2} &\geq \frac{1}{2}\lambda_{\min} + \frac{o(\|x - \overline{x}\|^2)}{\|x - \overline{x}\|^2} \\ &\geq \frac{1}{4}\lambda_{\min}. \end{aligned}$$

Consequently, if we set $\alpha = \frac{1}{4}\lambda_{\min}$, then

$$f(x) \geq f(\overline{x}) + \alpha \|x - \overline{x}\|^2$$

whenever $\|x - \overline{x}\| < \epsilon$.

□

In order to apply the second-order sufficient condition one must be able to check that a symmetric matrix is positive definite. As we have seen, this can be done by computing the eigenvalues of the matrix and checking that they are all positive. But there is another approach that is often easier to implement using the *principal minors* of the matrix.

THEOREM 1.7. *Let $H \in \mathbb{R}^{n\times n}$ be symmetric. We define the kth principal minor of H, denoted $\Delta_k(H)$, to be the determinant of the upper-left $k \times k$ submatrix of H. Then*

(1) H is positive definite if and only if $\Delta_k(H) > 0,\ k = 1, 2, \ldots, n$.
(2) H is negative definite if and only if $(-1)^k\Delta_k(H) > 0,\ k = 1, 2, \ldots, n$.

DEFINITION 1.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable at $\overline{x}$. If $\nabla f(\overline{x}) = 0$, but $\overline{x}$ is neither a local maximum or a local minimum, we call $\overline{x}$ a saddle point for f.*

THEOREM 1.8. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable at $\overline{x}$. If $\nabla f(\overline{x}) = 0$ and $\nabla^2 f(\overline{x})$ has both positive and negative eigenvalues, then $\overline{x}$ is a saddle point of f.*

THEOREM 1.9. *Let $H \in \mathbb{R}^{n\times n}$ be symmetric. If H is niether positive definite or negative definite and all of its principal minors are non-zero, then H has both positive and negative eigenvalues. In this case we say that H is indefinite.*

EXAMPLE 1.1. *Consider the matrix*

$$H = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 5 & 1 \\ -1 & 1 & 4 \end{bmatrix}.$$

We have

$$\Delta_1(H) = 1, \quad \Delta_2(H) = \begin{vmatrix} 1 & 1 \\ 1 & 5 \end{vmatrix} = 4, \quad \textit{and} \quad \Delta_3(H) = \det(H) = 8.$$

Therefore, H is positive definite.

1.4. Convexity. In the previous section we established first- and second-order optimality conditions. These conditions we based on only local information and so only refer to properties of local extrema. In this section we study the notion of convexity which allows us to provide optimality conditions for global solutions.

DEFINITION 1.3. *(1) A set $C \subset \mathbb{R}^n$ is said to be convex if for every $x, y \in C$ and $\lambda \in [0, 1]$ one has*

$$(1-\lambda)x + \lambda y \in C\ .$$

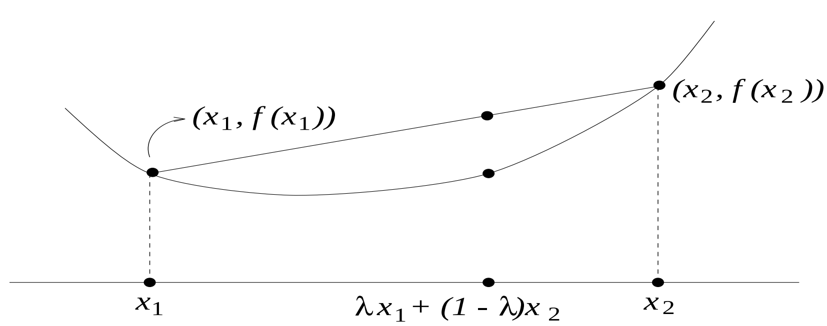
(2) A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be convex if for every two points $x_1, x_2 \in \mathbb{R}^n$ and $\lambda \in [0, 1]$ we have

$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda)f(x_2). \tag{70}$$

The function f is said to be strictly convex if for every two distinct points $x_1, x_2 \in \mathbb{R}^n$ and $\lambda \in (0, 1)$ we have

$$f(\lambda x_1 + (1-\lambda)x_2) < \lambda f(x_1) + (1-\lambda)f(x_2). \tag{71}$$

The inequality (70) is equivalent to the statement that the secant line connecting $(x_1, f(x_1))$ and $(x_2, f(x_2))$ lies above the graph of f on the line segment $\lambda x_1 + (1-\lambda)x_2,\ \lambda \in [0, 1]$.



That is, the set

$$\mathrm{epi}\,(f) = \{(x, \mu) \,:\, f(x) \le \mu\},$$

called the *epi-graph* of f is a convex set. Indeed, it can be shown that the convexity of the set $\mathrm{epi}\,(f)$ is equivalent to the convexity of the function f. This observation allows us to extend the definition of the convexity of a function to functions taking potentially infinite values.

DEFINITION 1.4. *A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\} = \bar{\mathbb{R}}$ is said to be convex if the set $epi\,(f) = \{(x, \mu) \,:\, f(x) \le \mu\}$ is a convex set. We also define the essential domain of f to be the set*

$$dom\,(f) = \{x \,:\, f(x) < +\infty\} \;.$$

We say that f is strictly convex if the strict inequality (71) holds whenever $x_1, x_2 \in dom\,(f)$ are distinct.

EXAMPLE 1.2. $c^T x$, $\|x\|$, e^x, x^2

The role of convexity in linking the global and the local in optimization theory is illustrated by the following result.

THEOREM 1.10. *Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ be convex. If $\overline{x} \in \mathbb{R}^n$ is a local minimum for f, then $\overline{x}$ is a global minimum for f.*

PROOF. Suppose to the contrary that there is a $\widehat{x} \in \mathbb{R}^n$ with $f(\widehat{x}) < f(\overline{x})$. Since $\overline{x}$ is a local solution, there is an $\epsilon > 0$ such that

$$f(\overline{x}) \le f(x) \quad \text{whenever} \quad \|x - \overline{x}\| \le \epsilon \;.$$

Taking ϵ smaller if necessary, we may assume that

$$\epsilon < 2\|\overline{x} - \widehat{x}\| \;.$$

Set $\lambda := \epsilon(2\|\overline{x} - \widehat{x}\|)^{-1} < 1$ and $x_\lambda := \overline{x} + \lambda(\widehat{x} - \overline{x})$. Then $\|x_\lambda - \overline{x}\| \le \epsilon/2$ and $f(x_\lambda) \le (1-\lambda)f(\overline{x}) + \lambda f(\widehat{x}) < f(\overline{x})$. This contradicts the choice of ϵ and so no such $\widehat{x}$ exists. □

Strict convexity implies the uniqueness of solutions.

THEOREM 1.11. *Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ be strictly convex. If f has a global minimizer, then it is unique.*

PROOF. Let x^1 and x^2 be distinct global minimizers of f. Then, for $\lambda \in (0, 1)$,

$$f((1-\lambda)x^1 + \lambda x^2) < (1-\lambda)f(x^1) + \lambda f(x^2) = f(x^1) \;,$$

which contradicts the assumption that x^1 is a global minimizer. □

If f is a differentiable convex function, much more can be said. We begin with the following lemma.

LEMMA 1.2. *Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ be convex (not necessarilty differentiable).*

(1) Given $x, d \in \mathbb{R}^n$ the difference quotient

$$\frac{f(x + td) - f(x)}{t} \tag{72}$$

is a non-decreasing function of t on $(0, +\infty)$.

(2) For every $x, d \in \mathbb{R}^n$ the directional derivative $f'(x; d)$ always exists and is given by

$$f'(x; d) := \inf_{t>0} \frac{f(x + td) - f(x)}{t}. \tag{73}$$

PROOF. We first assume (1) is true and show (2). Recall that

$$f'(x; d) := \lim_{t \downarrow 0} \frac{f(x + td) - f(x)}{t}. \tag{74}$$

Now if the difference quotient (72) is non-decreasing in t on $(0, +\infty)$, then the limit in (74) is necessarily given by the infimum in (73). This infimum always exists and so $f'(x; d)$ always exists and is given by (73).

We now prove (1). Let $x, d \in \mathbb{R}^n$ and let $0 < t_1 < t_2$. Then

$$\begin{aligned} f(x + t_1 d) &= f\left(x + \left(\tfrac{t_1}{t_2}\right) t_2 d\right) \\ &= f\left[\left(1 - \left(\tfrac{t_1}{t_2}\right)\right) x + \left(\tfrac{t_1}{t_2}\right)(x + t_2 d)\right] \\ &\le \left(1 - \tfrac{t_1}{t_2}\right) f(x) + \left(\tfrac{t_1}{t_2}\right) f(x + t_2 d). \end{aligned}$$

Hence
$$\frac{f(x+t_1d)-f(x)}{t_1} \le \frac{f(x+t_2d)-f(x)}{t_2}.$$
□

A very important consequence of Lemma 1.2 is the *subdifferential inequality.* This inequality is obtained by plugging $t = 1$ and $d = y - x$ into the right hand side of (73) where y is *any* other point in $\mathbb{R}^n$. This substitution gives the inequality

(75) $$f(y) \ge f(x) + f'(x; y - x) \quad \text{for all} \quad y \in \mathbb{R}^n \text{ and } x \in \operatorname{dom} f\ .$$

The subdifferential inequality immediately yields the following result.

THEOREM 1.12 (Convexity and Optimality). *Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ be convex (not necessarilty differentiable) and let $\overline{x} \in \operatorname{dom} f$. Then the following three statements are equivalent.*

(i) $\overline{x}$ is a local solution to $\min_{x\in\mathbb{R}^n} f(x)$.
(ii) $f'(\overline{x}; d) \ge 0$ for all $d \in \mathbb{R}^n$.
(iii) $\overline{x}$ is a global solution to $\min_{x\in\mathbb{R}^n} f(x)$.

PROOF. Lemma 1.1 gives the implication (i)⇒(ii). To see the implication (ii)⇒(iii) we use the subdifferential inequality and the fact that $f'(\overline{x}; y - \overline{x})$ exists for all $y \in \mathbb{R}^n$ to obtain
$$f(y) \ge f(\overline{x}) + f'(\overline{x}; y - \overline{x}) \ge f(\overline{x}) \quad \text{for all } y \in \mathbb{R}^n.$$
The implication (iii)⇒(i) is obvious. □

If it is further assumed that f is differentiable, then we obtain the following elementary consequence of Theorem 1.12.

THEOREM 1.13. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and suppose that $\overline{x} \in \mathbb{R}^n$ is a point at which f is differentiable. Then $\overline{x}$ is a global minimum of f if and only if $\nabla f(\overline{x}) = 0$.*

As Theorems 1.12 and 1.13 demonstrate, convex functions are well suited to optimization theory. Thus, it is important that we be able to recognize when a function is convex. For this reason we give the following result.

THEOREM 1.14. *Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$.*

(1) If f is differentiable on $\mathbb{R}^n$, then the following statements are equivalent:
  (a) f is convex,
  (b) $f(y) \ge f(x) + \nabla f(x)^T(y - x)$ for all $x, y \in \mathbb{R}^n$
  (c) $(\nabla f(x) - \nabla f(y))^T(x - y) \ge 0$ for all $x, y \in \mathbb{R}^n$.
(2) If f is twice differentiable then f is convex if and only if $\nabla^2 f(x)$ is positive semi-definite for all $x \in \mathbb{R}^n$.

REMARK 1.2. *The condition in Part (c) is called monotonicity.*

PROOF. (a) ⇒ (b) If f is convex, then 1.14 holds. By setting $t := 1$ and $d := y - x$ we obtain (b).

(b) ⇒ (c) Let $x, y \in \mathbb{R}^n$. From (b) we have
$$f(y) \ge f(x) + \nabla f(x)^T(y - x)$$
and
$$f(x) \ge f(y) + \nabla f(y)^T(x - y).$$
By adding these two inequalities we obtain (c).

(c) ⇒ (b) Let $x, y \in \mathbb{R}^n$. By the Mean Value Theorem there exists $0 < \lambda < 1$ such that
$$f(y) - f(x) = \nabla f(x_\lambda)^T(y - x)$$
where $x_\lambda := \lambda y + (1 - \lambda)x$. By hypothesis,
$$\begin{aligned} 0 &\le [\nabla f(x_\lambda) - \nabla f(x)]^T(x_\lambda - x) \\ &= \lambda[\nabla f(x_\lambda) - \nabla f(x)]^T(y - x) \\ &= \lambda[f(y) - f(x) - \nabla f(x)^T(y - x)]. \end{aligned}$$
Hence $f(y) \ge f(x) + \nabla f(x)^T(y - x)$.

(b) $\Rightarrow$ (a) Let $x, y \in \mathbb{R}^n$ and set

$$\alpha := \max_{\lambda \in [0,1]} \varphi(\lambda) := [f(\lambda y + (1-\lambda)x) - (\lambda f(y) + (1-\lambda)f(x))].$$

We need to show that $\alpha \leq 0$. Since $[0,1]$ is compact and φ is continuous, there is a $\lambda \in [0,1]$ such that $\varphi(\lambda) = \alpha$. If λ equals zero or one, we are done. Hence we may as well assume that $0 < \lambda < 1$ in which case

$$0 = \varphi'(\lambda) = \nabla f(x_\lambda)^T(y-x) + f(x) - f(y)$$

where $x_\lambda = x + \lambda(y-x)$, or equivalently

$$\lambda f(y) = \lambda f(x) - \nabla f(x_\lambda)^T(x - x_\lambda).$$

But then

$$\begin{array}{rcl} \alpha & = & f(x_\lambda) - (f(x) + \lambda(f(y) - f(x))) \\ & = & f(x_\lambda) + \nabla f(x_\lambda)^T(x - x_\lambda) - f(x) \\ & \leq & 0 \end{array}$$

by (b).

2) Suppose f is convex and let $x, d \in \mathbb{R}^n$, then by (b) of Part (1),

$$f(x + td) \geq f(x) + t\nabla f(x)^T d$$

for all $t \in \mathbb{R}$. Replacing the left hand side of this inequality with its second-order Taylor expansion yields the inequality

$$f(x) + t\nabla f(x)^T d + \frac{t^2}{2} d^T \nabla^2 f(x) d + o(t^2) \geq f(x) + t\nabla f(x)^T d,$$

or equivalently,

$$\frac{1}{2} d^t \nabla^2 f(x) d + \frac{o(t^2)}{t^2} \geq 0.$$

Letting $t \to 0$ yields the inequality

$$d^T \nabla^2 f(x) d \geq 0.$$

Since d was arbitrary, $\nabla^2 f(x)$ is positive semi-definite.

Conversely, if $x, y \in \mathbb{R}^n$, then by the Mean Value Theorem there is a $\lambda \in (0,1)$ such that

$$f(y) = f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(x_\lambda)(y-x)$$

where $x_\lambda = \lambda y + (1-\lambda)_x$. Hence

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

since $\nabla^2 f(x_\lambda)$ is positive semi-definite. Therefore, f is convex by (b) of Part (1).

□

Convexity is also preserved by certain operations on convex functions. A few of these are given below.

Theorem 1.15. *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, $h : \mathbb{R}^s \times \mathbb{R}^k \to \overline{\mathbb{R}}$ and $f_\nu : \mathbb{R}^n \to \overline{\mathbb{R}}$ be convex functions for $\nu \in N$ where N is an arbitrary index set, and let $\nu_i \in N$ and $\lambda_i \geq 0,\ i = 1, \ldots, m$. Then the following functions are also convex.*

(1) $\phi \circ f$, *where* $\phi : \mathbb{R} \to \mathbb{R}$ *is any non-decreasing function on* $\mathbb{R}$.
(2) $f(x) := \sum_{i=1}^m \lambda_i f_{nu_i}(x)$ *(Non-negative linear combinations)*
(3) $f(x) := \max_{\nu \in N} f_\nu(x)$ *(pointwise max)*
(4) $f(x) := \sup\left\{\sum_{i=1}^m f_{\nu_i}(x^i) \,\middle|\, x = \sum_{i=1}^m x^i\right\}$ *(infimal convolution)*
(5) $f^*(y) := \sup_{x \in \mathbb{R}^n}[y^T x - f(x)]$ *(convex conjugation)*
(6) $\psi(y) = \inf_{x \in \mathbb{R}^s} h(x, y)$ *(infimal projection)*

1.4.1. *More on the Directional Derivative.* It is a powerful fact that convex function are directionally differentiable at every point of their domain in every direction. But this is just the beginning of the story. The directional derivative of a convex function possess several other important and surprising properties. We now develop a few of these.

DEFINITION 1.5. *Let $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$. We say that h is positively homogeneous if*

$$h(\lambda x) = \lambda h(x) \quad \text{for all } x \in \mathbb{R} \text{ and } \lambda > 0.$$

We say that h is subadditive if

$$h(x + y) \leq h(x) + h(y) \quad \text{for all } x, y \in \mathbb{R}.$$

Finally, we say that h is sublinear if it is both positively homogeneous and subadditive.

There are numerous important examples of sublinear functions (as we shall soon see), but perhaps the most familiar of these is the norm $\|x\|$. Positive homogeneity is obvious and subadditivity is simply the triangle inequality. In a certain sense the class of sublinear function is a generalization of norms. It is also important to note that sublinear functions are always convex functions. Indeed, given $x, y \in \operatorname{dom} h$ and $0 \leq \lambda \leq 1$,

$$\begin{aligned} h(\lambda x + (1-\lambda)y) &\leq h(\lambda x) + h(1-\lambda)y) \\ &= \lambda h(x) + (1-\lambda)h(y). \end{aligned}$$

THEOREM 1.16. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function. Then at every point $x \in \operatorname{dom} f$ the directional derivative $f'(x; d)$ is a sublinear function of the d argument, that is, the function $f'(x; \cdot) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is sublinear. Thus, in particular, the function $f'(x; \cdot)$ is a convex function.*

REMARK 1.3. *Since f is convex and $x \in \operatorname{dom} f$, $f'(x; d)$ exists for all $d \in \mathbb{R}^n$.*

PROOF. Let $x \in \operatorname{dom} f$, $d \in \mathbb{R}^n$, and $\lambda > 0$. Then

$$\begin{aligned} f'(x; \lambda d) &= \lim_{t \downarrow 0} \frac{f(x + t\lambda d) - f(x)}{t} \\ &= \lim_{t \downarrow 0} \lambda \frac{f(x + t\lambda d) - f(x)}{\lambda t} \\ &= \lambda \lim_{(\lambda t) \downarrow 0} \frac{f(x + (t\lambda)d) - f(x)}{(\lambda t)} \\ &= \lambda f'(x; d), \end{aligned}$$

showing that $f'(x; \cdot)$ is positively homogeneous.

Next let $d_1, d_2 \in \mathbb{R}^n$, Then

$$\begin{aligned} f'(x; d_1 + d_2) &= \lim_{t \downarrow 0} \frac{f(x + t(d_1 + d_2)) - f(x)}{t} \\ &= \lim_{t \downarrow 0} \frac{f(\frac{1}{2}(x + 2td_1) + \frac{1}{2}(x + 2td_2)) - f(x)}{t} \\ &\leq \lim_{t \downarrow 0} \frac{\frac{1}{2}f(x + 2td_1) + \frac{1}{2}f(x + 2td_2) - f(x)}{t} \\ &\leq \lim_{t \downarrow 0} \frac{\frac{1}{2}(f(x + 2td_1) - f(x)) + \frac{1}{2}(f(x + 2td_2) - f(x))}{t} \\ &= \lim_{t \downarrow 0} \frac{f(x + 2td_1) - f(x)}{2t} + \lim_{t \downarrow 0} \frac{f(x + 2td_2) - f(x)}{2t} \\ &= f'(x; d_1) + f'(x; d_2), \end{aligned}$$

showing that $f'(x; \cdot)$ is subadditive and completing the proof. □