Appendix B
Convex Sets and Functions

The notion of convexity arises when we try to formalize the property that "good local decisions lead to globally optimal solutions." Consider a generic unconstrained optimization problem

$$\text{minimize} \quad f(\boldsymbol{x}). \tag{B.0.1}$$

Here $\boldsymbol{x} \in \mathbb{R}^n$ is the variable of optimization, and $f : \mathbb{R}^n \to \mathbb{R}$ is the objective function, which we are trying to make as small as possible using a numerical algorithm. Figure B.1 displays two objective functions f. The one on the right has many peaks and valleys – it may be very difficult to find the lowest valley, corresponding to the global optimum $\boldsymbol{x}_\star$. Moreover, for the function f on the right, local information around a point $\boldsymbol{x}$ is not particularly helpful for determining what direction to move to reach the global optimum. In contrast, the bowl-shaped function on the left is much more amenable to global optimization – a "gradient descent" type algorithm, that simply determined which direction to move by considering the slope of the graph of the function, would easily "ski" down to the global minimum.

The notion of *convexity* formalizes this property. Convexity is a geometric property. It is convenient to first introduce the notion of a convex set, and then extend this definition to functions.

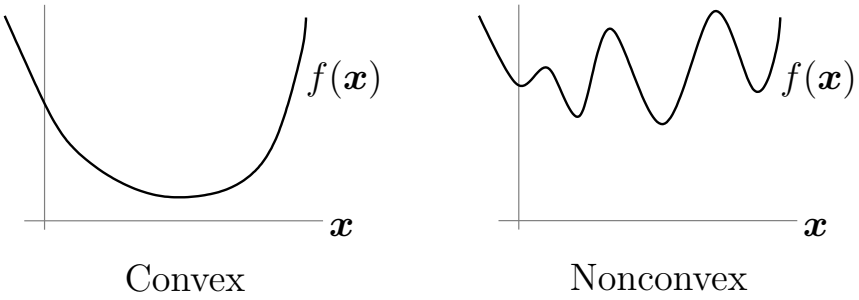

Figure B.1. **Two optimization problems** $\min f(\boldsymbol{x})$. The objective f at left appears to be amenable to global optimization, while the one at right appears to be more challenging.

B.1 Convex Sets

A set C is said to be *closed* if it contains its boundary. More precisely, for any converging sequence of points $\{\boldsymbol{x}_k\}$ in C, we must have:

$$\boldsymbol{x}_k \to \bar{\boldsymbol{x}} \quad \Rightarrow \quad \bar{\boldsymbol{x}} \in C.$$

A set $C \subseteq \mathbb{R}^n$ is *convex* if for every pair of points $\boldsymbol{x}, \boldsymbol{x}' \in C$, the line segment obtained by joining the two points also lies entirely in C:

Definition B.1.1 (Convex set)**.** *$C \subseteq \mathbb{R}^n$ is* convex *if*

$$\forall\, \boldsymbol{x},\, \boldsymbol{x}' \in \mathsf{C}, \quad \alpha \in [0,1], \qquad \alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{x}' \in C. \tag{B.1.1}$$

Figure B.2 gives an example of two sets, one of which is convex and one of which is not.

Example B.1.2 (Convex sets)**.** *Show that the following are convex:*

- *Every affine subspace.*
- *Every norm ball* $\mathfrak{B}_{\|\cdot\|} = \{\boldsymbol{x} \mid \|\boldsymbol{x}\| \leq 1\}$.
- *The empty set.*
- *Any intersection* $C = C_1 \cap C_2$ *of two convex sets* C_1, C_2.

Proposition B.1.3. *1. The intersection of a collection of convex sets* $\bigcap_i C_i$ *is convex.*

2. The image of a convex set under an affine transformation is convex.

Definition B.1.4 (Convex hull)**.** *The* convex hull *of any given set S is the minimal convex set containing S, denoted as conv(S). If S contains a finite number of $S = \{x_i\}_{i=1}^n$ points, we have*

$$\operatorname{conv}(S) = \left\{ \sum_{i=1}^n \alpha_i x_i \;\middle|\; \forall \alpha_i \geq 0 \text{ with } \sum_{i=1}^n \alpha_i = 1. \right\}. \tag{B.1.2}$$

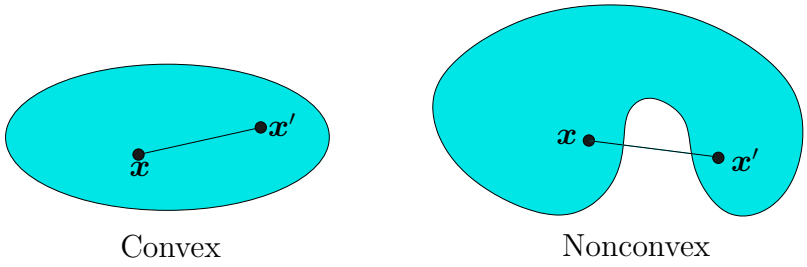

Figure B.2. **Convex and nonconvex sets.** A set is convex if we can select any pair of points $\boldsymbol{x}$, $\boldsymbol{x}'$ in the set, and the line segment joining them lies entirely within the set. The set to the left has this property, while the set to the right does not.

B.2 Convex Functions

For a function $f : \mathcal{D} \to \mathbb{R}$ defined on a (convex) domain $\mathcal{D} \subseteq \mathbb{R}^n$, its *graph* is the set of pairs $(\boldsymbol{x}, f(\boldsymbol{x}))$ that can be generated by evaluating the function f at every point:

$$\text{graph}(f) = \{(\boldsymbol{x}, f(\boldsymbol{x})) \mid \boldsymbol{x} \in \mathcal{D}, f(\boldsymbol{x}) < +\infty\} \subseteq \mathbb{R}^{n+1}. \tag{B.2.1}$$

We give another name to everything that lies above the graph: the *epigraph*:

$$\text{epi}(f) = \{(\boldsymbol{x}, t) \mid \boldsymbol{x} \in \mathcal{D}, t \in \mathbb{R}, f(\boldsymbol{x}) \leq t\} \subseteq \mathbb{R}^{n+1}. \tag{B.2.2}$$

We say that f is a *convex function* if its epigraph is a convex set. Figure B.3 (right) illustrates this property. Figure B.3 (left) suggests an equivalent definition, which is sometimes easier to work with: f is convex if for any pair of points $\boldsymbol{x}$ and $\boldsymbol{x}'$, the line segment joining $(\boldsymbol{x}, f(\boldsymbol{x}))$ and $(\boldsymbol{x}', f(\boldsymbol{x}'))$ lies entirely above the graph of f:

Definition B.2.1 (Convex function)**.** *A function $f : \mathcal{D} \to \mathbb{R}$ is* convex *if for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D}$ and $\alpha \in [0, 1]$,*

$$\alpha f(\boldsymbol{x}) + (1 - \alpha) f(\boldsymbol{x}') \geq f(\alpha \boldsymbol{x} + (1 - \alpha)\boldsymbol{x}'). \tag{B.2.3}$$

Notice that above definitions do not require f is differentiable. If f is differentiable, the notion of convex can be characterized in terms of its derivatives. Since the epigraph is convex, then the tangent plane at each point of the graph should lies beneath the graph. The following statement makes this precise:

Proposition B.2.2 (First-Order Conditions)**.** *Let $f : \mathcal{D} \to \mathbb{R}$ be differentiable. Then f is convex if and only if it satisfies the condition:*

$$f(\boldsymbol{x}') \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T (\boldsymbol{x}' - \boldsymbol{x})$$

for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D}$.

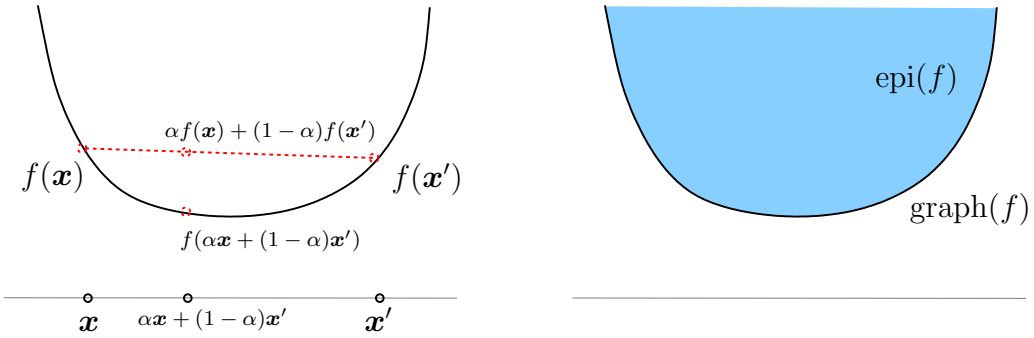

Figure B.3. **Convexity of functions**: a function f is convex if its epigraph $\text{epi}(f) = \{(\boldsymbol{x}, t) \mid t \geq f(\boldsymbol{x})\}$ is a convex set (right). This is true if and only if for every pair of points $\boldsymbol{x}$, $\boldsymbol{x}'$ and scalar $\alpha \in [0, 1]$, $f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{x}') \leq \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{x}')$. The picture at right illustrates this inequality: the segment joining $(\boldsymbol{x}, f(\boldsymbol{x}))$ and $(\boldsymbol{x}', f(\boldsymbol{x}'))$ lies above the graph of f.

This is precisely the geometry of the "nice" function in Figure B.1 (left). From this picture, it is clear that convexity is very favorable for global optimization.[1]

You may also notice that in Figure B.3, the function $f(\boldsymbol{x})$ "curves upward": its second derivative is nonnegative at every point of the domain. For twice differentiable functions, this leads to a simpler condition for con-

[1] Once you've internalized the definition a bit, you may begin to wonder to what extent the implication "convexity $\implies$ easy-to-optimize" is actually true. The convex functions that we encounter in this book will all possess special structure that makes them very amenable to efficient algorithms. However, this is not true of all convex functions – there exist convex functions that are NP-hard to optimize.

There also exist nonconvex functions that are easy to optimize – Chapter 9 provides a brief introduction to this emerging literature. However, if we want to talk about a class of functions, rather than a particular one, then there is a very beautiful motivation for studying convex functions. To appreciate this motivation, we need to first observe a useful fact: if $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ are convex functions, then for any $\alpha, \beta \geq 0$, $h(\boldsymbol{x}) = \alpha f(\boldsymbol{x}) + \beta g(\boldsymbol{x})$ is also convex. If we let $\mathcal{F}$ be the largest class of continuously differentiable functions that satisfy the following three demands:

- Every linear function $\phi(\boldsymbol{x}) = \boldsymbol{a}^* \boldsymbol{x} + b$ is in $\mathcal{F}$,
- Every nonnegative combination $\alpha f_1(\boldsymbol{x}) + \beta f_2(\boldsymbol{x})$ of $f_1, f_2 \in \mathcal{F}$ is in $\mathcal{F}$,
- For every $f \in \mathcal{F}$, the stationarity condition $\nabla f(\boldsymbol{x}_\star) = 0$ implies that $\boldsymbol{x}_\star$ is a global optimizer of f,

it turns out that the $\mathcal{F}$ is precisely the class of **convex**, continuously differentiable functions. You can interpret this as suggesting that for *global* solutions, convex functions really are the right general class of functions to study. For more details, see the book of Nesterov [Nesterov, 2003a].

vexity: the function is convex if and only if its second derivative at any point, and in any direction is positive. The following makes this precise:

Proposition B.2.3 (Second-Order Conditions)**.** *Let $f : \mathcal{D} \to \mathbb{R}$ be twice differentiable. Then f is convex if and only if its Hessian is positive semidefinite:*

$$\nabla^2 f(\boldsymbol{x}) \succeq 0$$

for all $\boldsymbol{x} \in \mathcal{D}$.

The class of convex functions includes important examples such as linear functions and norms:

Example B.2.4 (Convex functions)**.** *Show that the following are convex functions:*

- *Every affine function $f(\boldsymbol{x}) = \boldsymbol{a}^* \boldsymbol{x} + b$.*
- *Every norm $f(\boldsymbol{x}) = \|\boldsymbol{x}\|$.*
- *Every semidefinite quadratic $f(\boldsymbol{x}) = \boldsymbol{x}^* \boldsymbol{P} \boldsymbol{x}$, with $\boldsymbol{P} \succeq \boldsymbol{0}$.*
- *The maximum eigenvalue $\lambda_{\max}(\boldsymbol{X})$ is a convex function over the symmetric matrices $\boldsymbol{X}$.*

Before continuing, we note one nice property of convex functions which will be useful for deriving an appropriate tractable replacement for the ℓ^0 norm.

Definition B.2.5 (Convex combination)**.** *A convex combination of points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ is an expression of the form $\lambda_1 \boldsymbol{x}_1 + \cdots + \lambda_k \boldsymbol{x}_k$, with $\lambda_i \geq 0$ for each i and $\sum_i \lambda_i = 1$.*

Lemma B.2.6 (Jensen's inequality)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. For any k, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \mathbb{R}^n$, $\lambda_1, \ldots, \lambda_k \in \mathbb{R}_+$, with $\sum_i \lambda_i = 1$,*

$$f\left(\sum_i \lambda_i \boldsymbol{x}_i\right) \;\leq\; \sum_i \lambda_i f(\boldsymbol{x}_i). \tag{B.2.4}$$

Proof. The proof is by induction on k. For $k = 1$, there is nothing to show. Now suppose the claim is true for $1, \ldots, k-1$. Then

$$\begin{aligned} f\left(\sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i\right) &\leq \left(\sum_{i=1}^{k-1} \lambda_i\right) f\left(\frac{\sum_{i=1}^{k-1} \lambda_i \boldsymbol{x}_i}{\sum_{i=1}^{k-1} \lambda_i}\right) + \lambda_k f(\boldsymbol{x}_k) && \text{(B.2.5)} \\ &\leq \sum_{i=1}^{k} \lambda_i f(\boldsymbol{x}_i) && \text{(B.2.6)} \end{aligned}$$

as desired. Above, the first step uses the definition of convexity, and the second uses the inductive hypothesis. □

With this lemma, it is easy to show that any α-sublevel set of a convex function $f : \mathcal{D} \to \mathbb{R}$:

$$C_\alpha = \{\boldsymbol{x} \in \mathcal{D} \mid f(\boldsymbol{x}) \leq \alpha\} \tag{B.2.7}$$

is a convex set. However, a function with all its sublevel sets being convex is not necessarily a convex function![2] A function is said to be a *closed* function, if each sublevel set is a closed set. We typically only consider closed convex functions, unless otherwise stated.

Proposition B.2.7. *We can use convex functions to generate other associated convex functions:*

1. *A function is convex if and only if it is convex when restricted to any line that intersects its domain.*
2. *A weighted sum of convex functions with nonnegative weights is convex.*
3. *If f, g are convex functions and g is non-decreasing in its univariate domain, the $h(\boldsymbol{x}) = g(f(\boldsymbol{x}))$ is convex.*
4. *Given a collection of convex functions $f_\alpha : \mathcal{D} \to \mathbb{R}$, $\alpha \in \mathsf{A}$, their point-wise supremum*

$$f(\boldsymbol{x}) \doteq \sup_{\alpha \in \mathsf{A}} f_\alpha(\boldsymbol{x})$$

is also convex.

Example B.2.8. *The maximal eigenvalue of a symmetric matrix is a (closed) convex function.*

Proof. To see that, the maximal eigenvalue function can be written as

$$\sigma_{\max}(\boldsymbol{X}) = \sup\{\boldsymbol{y}^T \boldsymbol{X} \boldsymbol{y}\}, \quad \|\boldsymbol{y}\|_2 = 1.$$

Since the function is the point-wise supremum of a set of linear functions with respect to $\boldsymbol{X}$, it is a convex function. □

Convex Envelop and Conjugate.

For any non-convex (closed) function $g : \mathcal{D} \to \mathbb{R}$ defined on a convex domain $\mathcal{D}$, it has a naturally associated convex function that bounds it from below:

Definition B.2.9 (Convex envelop)**.** *The convex envelop of a closed function g is defined as*

$$\mathrm{conv} g(\boldsymbol{x}) = \sup\{h(\boldsymbol{x}) \mid h(\boldsymbol{x}) \text{ convex } \& \; h(\boldsymbol{x}) \leq g(\boldsymbol{x}) \;\; \forall \boldsymbol{x} \in \mathcal{D}\}. \tag{B.2.8}$$

[2] Such functions are called *quasi-convex*. Please find an example for yourself.

Let us define the (Fenchel) *conjugate* of a function $g(\boldsymbol{x})$ (not necessarily convex) as:

$$g^*(\boldsymbol{\lambda}) = \max_{\boldsymbol{x}} \boldsymbol{\lambda}^T \boldsymbol{x} - g(\boldsymbol{x}). \tag{B.2.9}$$

The conjugate of a function g is essentially the negated dual function of g that we often see in the method of Largragian multipliers (see Section C.3).

Proposition B.2.10. *Assuming the conjugate is well-defined, we have the following:*

1. *The conjugate $g^*(\boldsymbol{\lambda})$ is always a convex function.*
2. $g^{**}(\boldsymbol{x}) = \text{conv}g(\boldsymbol{x})$.

Strong Convexity.

In this book, we sometimes are interested in stronger notion of convexity.

Definition B.2.11 (Strongly convex function)**.** *A function $f : \mathcal{D} \to \mathbb{R}$ is* strongly convex *if f is convex and for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D}$ and $\alpha \in [0, 1]$,*

$$\alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{x}') \geq f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{x}') + \mu \frac{\lambda(1-\lambda)}{2} \|\boldsymbol{x} - \boldsymbol{x}'\|^2 \tag{B.2.10}$$

for some $\mu > 0$.

Notice that the above defition does not require f to be differentiable. If f is first or second order differentiable, we have the following sufficient conditions for f being strongly convex.

Proposition B.2.12. *For a differentiable convex function f over $\mathcal{D}$, we have f is strongly convex if either of the following conditions hold:*

1. $f(\boldsymbol{x}') \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T (\boldsymbol{x}' - \boldsymbol{x}) + \frac{\mu}{2} \|\boldsymbol{x}' - \boldsymbol{x}\|^2, \ \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D}$;
2. $\nabla^2 f(\boldsymbol{x}) \succeq \frac{\mu}{2} I, \ \forall \boldsymbol{x} \in \mathcal{D}$;

for some $\mu > 0$.

However, as we see in Section 3.3.3, we are interested in strong convexity in a restricted sense.

B.3 Subdifferentials of Nonsmooth Convex Functions

For smooth, convex functions f, the local information encoded in the gradient ∇f and Hessian $\nabla^2 f$ characterize both the local and global behavior of f, allowing us to give optimality conditions and construct minimization algorithms. Familiar, classical algorithms such as gradient ascent, Newton's

Figure B.4. Subgradients and subdifferentials of convex functions.

method, and their variants, are all constructed using differential information. Moreover, as we saw in the previous section, these quantities play a critical role in characterizing convexity for smooth functions f.

It is a curious fact, then, that many of the most useful convex objective functions arising in high-dimensional data analysis are nondifferentiable: *their gradients and Hessians do not exist.* For example, the ℓ^1 norm $\|\boldsymbol{x}\|_1 = \sum_{i=1}^n |x_i|$ is nondifferentiable at any point $\boldsymbol{x} \in \mathbb{R}^n$ with fewer than n nonzero entries. These are precisely the points that we care about for sparse estimation! This nonsmooth behavior is actually desirable from the statistical perspective. However, it forces us to make recourse to analytical tools that are general enough to handle nondifferentiable functions. Fortunately, for convex functions, the nondifferential theory rests on simple, geometrically intuitive ideas, which we describe in this section. For accessible introductions to the general theory of convexity, we recommend [Nemirovski, 1995, Nemirovski, 2007, Nesterov, 2003a, Boyd and Vandenberghe, 2004].

The most important notion is that of a *subgradient* of a convex function, which provides a very satisfactory replacement for the gradient, when the function is not differentiable. Recall from Proposition B.2.2 that for convex, *differentiable* f,

$$f(\boldsymbol{y}) \;\geq\; f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle\,, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}. \tag{B.3.1}$$

This inequality has a simple geometric interpretation, which we visualize in Figure B.4. We visualize the graph of the function $f : \mathbb{R}^n \to \mathbb{R}$. The graph is the collection of points of the form $(\boldsymbol{x}, f(\boldsymbol{x})) \in \mathbb{R}^{n+1}$. The graph of

$$h(\boldsymbol{y}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$$

is a hyperplane, which is tangent to the graph of $\boldsymbol{f}$ at $(\boldsymbol{x}, f(\boldsymbol{x}))$. The inequality (B.3.1) says that at all points $\boldsymbol{y}$ in the domain of the function f this tangent hyperplane lies below (or more precisely, not above) the graph of f.

Figure B.4 (right) visualizes the graph of another convex function f, which is not differentiable at point $\boldsymbol{x}$. The gradient of f *does not* exist at $\boldsymbol{x}$. Nevertheless, we can still define a nonvertical hyperplane $\mathcal{H} \subseteq \mathbb{R}^{n+1}$ that passes through $(\boldsymbol{x}, f(\boldsymbol{x}))$, and lies below the graph of f. This hyperplane has normal vector $(\boldsymbol{v}, -1)$, and can be expressed in notation as

$$\mathcal{H} = \{(\boldsymbol{y}, t) \mid t = f(\boldsymbol{x}) + \langle \boldsymbol{v}, \boldsymbol{y} - \boldsymbol{x} \rangle\}\,. \tag{B.3.2}$$

We say that $\boldsymbol{v} \in \mathbb{R}^n$ is a *subgradient* of f at $\boldsymbol{x}$ if it defines a hyperplane that supports the graph of f at $\boldsymbol{x}$, and lies below the graph everywhere:

Definition B.3.1. *Let $f : \mathcal{D} \subset \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function. A vector $\boldsymbol{v}$ is a* subgradient *of f at $\boldsymbol{x} \in \mathcal{D}$ if for all $\boldsymbol{y} \in \mathcal{D}$,*

$$f(\boldsymbol{y}) \;\geq\; f(\boldsymbol{x}) + \langle \boldsymbol{v}, \boldsymbol{y} - \boldsymbol{x} \rangle \,. \tag{B.3.3}$$

When f is differentiable, from Proposition B.2.2 it is clear that $\boldsymbol{v} = \nabla f(\boldsymbol{x})$ satisfies (B.3.3). When f is *nondifferentiable*, at a given point $\boldsymbol{x}$ there can be multiple distinct hyperplanes that support the graph of f, and hence, there can be multiple subgradients $\boldsymbol{v}$ (see Figure B.4). The collection of all subgradients is called the *subdifferential* of f at $\boldsymbol{x}$, and is denoted $\partial f(\boldsymbol{x})$. Formally:

Definition B.3.2. *Let $f : \mathcal{D} \subseteq \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function. The* subdifferential *$\partial f(\boldsymbol{x})$ is the collection of all subgradients of f at $\boldsymbol{x}$:*

$$\partial f(\boldsymbol{x}) = \{\boldsymbol{v} \mid f(\boldsymbol{y}) \;\geq\; f(\boldsymbol{x}) \;+\; \langle \boldsymbol{v}, \boldsymbol{y} - \boldsymbol{x} \rangle \,,\; \forall \boldsymbol{y} \in \mathcal{D}\} \,. \tag{B.3.4}$$

Notice that if $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at $\boldsymbol{x}$, its subdifferential at $\boldsymbol{x}$ is a singleton: $\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}$. This coincides with the classical definition of differentials.

A number of functions of interest have relatively simple subdifferentials.

Example B.3.3. *As good exercises, the reader may try to verify what the subdifferentials for the following functions:*

1. *The subdifferential for $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$ with $\boldsymbol{x} \in \mathbb{R}^n$.*
2. *The subdifferential for $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_\infty$ with $\boldsymbol{x} \in \mathbb{R}^n$.*
3. *The subdifferential for $f(\boldsymbol{X}) = \sum_{j=1}^n \|\boldsymbol{X}\boldsymbol{e}_j\|_2$ with $\boldsymbol{X}$ a matrix in $\mathbb{R}^{n \times n}$.*
4. *The subdifferential for $f(\boldsymbol{x}) = \|\boldsymbol{X}\|_*$ with $\boldsymbol{X}$ a matrix in $\mathbb{R}^{n \times n}$.*

Below are some basic properties of subdifferentials.

Lemma B.3.4 (Monotonicity Property)**.** *Given a convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and any $\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{v}, \boldsymbol{v}' \in \mathbb{R}^n$ such that $\boldsymbol{v} \in \partial f(\boldsymbol{x})$ and $\boldsymbol{v}' \in \partial f(\boldsymbol{x}')$, we have*

$$\langle \boldsymbol{x} - \boldsymbol{x}', \boldsymbol{v} - \boldsymbol{v}' \rangle \geq 0. \tag{B.3.5}$$

Proof. From the definition of subgradient (B.3.2), we have

$$f(\boldsymbol{x}') \geq f(\boldsymbol{x}) + \langle \boldsymbol{v}, \boldsymbol{x}' - \boldsymbol{x} \rangle, \quad f(\boldsymbol{x}) \geq f(\boldsymbol{x}') + \langle \boldsymbol{v}', \boldsymbol{x} - \boldsymbol{x}' \rangle. \tag{B.3.6}$$

Adding these two inequalities together we obtain:

$$f(\boldsymbol{x}) + f(\boldsymbol{x}') \geq f(\boldsymbol{x}) + f(\boldsymbol{x}') + \langle \boldsymbol{v} - \boldsymbol{v}', \boldsymbol{x}' - \boldsymbol{x} \rangle. \tag{B.3.7}$$

Canceling $f(\boldsymbol{x}) + f(\boldsymbol{x}')$ from both sides obtains the desired result. □

Lemma B.3.5. *If a convex function $f(\boldsymbol{x})$ has Lipschitz continuous gradient with constant L, then for any $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, we have:*

$$\langle \nabla f(\boldsymbol{x}_1) - \nabla f(\boldsymbol{x}_2), \boldsymbol{x}_1 - \boldsymbol{x}_2 \rangle \geq \frac{1}{L} \|\nabla f(\boldsymbol{x}_1) - \nabla f(\boldsymbol{x}_2)\|^2 \geq 0. \tag{B.3.8}$$

Proof. Let us define a function $h(\boldsymbol{z}) \doteq f(\boldsymbol{z}) - \boldsymbol{z}^* \nabla f(\boldsymbol{x})$. Then $h(\boldsymbol{z})$ is convex and is minimized at $\boldsymbol{z} = \boldsymbol{x}$ (as $\nabla h(\boldsymbol{x}) = 0$). Hence for any $\boldsymbol{z}$, we have

$$h(\boldsymbol{x}) \leq h\big(\boldsymbol{z} - \frac{1}{L}\nabla h(\boldsymbol{z})\big) \leq h(\boldsymbol{z}) + \langle \nabla h(\boldsymbol{z}), -\frac{1}{L}\nabla h(\boldsymbol{z}) \rangle + \frac{L}{2}\|\frac{1}{L}\nabla h(\boldsymbol{z})\|^2.$$

The last inequality comes from the fact that the function $f(\boldsymbol{x})$ (and hence $h(\boldsymbol{z})$) has Lipschitz continuous gradient with constant L. This gives

$$h(\boldsymbol{x}) \leq h(\boldsymbol{z}) - \frac{1}{2L}\|\nabla h(\boldsymbol{z})\|^2. \tag{B.3.9}$$

Now applying the inequality to $\boldsymbol{x} = \boldsymbol{x}_1, \boldsymbol{z} = \boldsymbol{x}_2$ as well as the reverse case $\boldsymbol{x} = \boldsymbol{x}_2, \boldsymbol{z} = \boldsymbol{x}_1$, we get

$$\begin{aligned} f(\boldsymbol{x}_1) - \boldsymbol{x}_1^* \nabla f(\boldsymbol{x}_1) &\leq f(\boldsymbol{x}_2) - \boldsymbol{x}_2^* \nabla f(\boldsymbol{x}_1) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|^2, \\ f(\boldsymbol{x}_2) - \boldsymbol{x}_2^* \nabla f(\boldsymbol{x}_2) &\leq f(\boldsymbol{x}_1) - \boldsymbol{x}_1^* \nabla f(\boldsymbol{x}_2) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_1) - \nabla f(\boldsymbol{x}_2)\|^2. \end{aligned}$$

Adding these two together gives the desired bound (B.3.8). □