Convex optimization

Notation

Matrices are written in uppercase: A, vectors are written in lowercase: a. A_{ij} denotes the element of A in position (i, j), A_i denotes the ith column of A (it's a vector!). Beware that x_i may denote the ith entry of a vector x or a the ith vector in a list depending on the context. $\mathcal{I}$ denotes a subvector of x that contains the entries listed in the set $\mathcal{I}$. For example, $x_{1:n}$ contains the first n entries of x.

1 Convexity

1.1 Convex sets

A set is convex if it contains all segments connecting points that belong to it.

Definition 1.1 (Convex set)**.** *A convex set $\mathcal{S}$ is any set such that for any $x, y \in \mathcal{S}$ and $\theta \in (0, 1)$*

$$\theta x + (1 - \theta)\, y \in \mathcal{S}. \tag{1}$$

Figure 1 shows a simple example of a convex and a nonconvex set.

The following lemma establishes that the intersection of convex sets is convex.

Lemma 1.2 (Intersection of convex sets)**.** *Let $\mathcal{S}_1, \ldots, \mathcal{S}_m$ be convex subsets of $\mathbb{R}^n$, $\cap_{i=1}^m \mathcal{S}_i$ is convex.*

Proof. Any $x, y \in \cap_{i=1}^m \mathcal{S}_i$ also belong to S_1. By convexity of $\mathcal{S}_1$ $\theta x + (1 - \theta)\, y$ belongs to S_1 for any $\theta \in (0, 1)$ and therefore also to $\cap_{i=1}^m \mathcal{S}_i$. $\square$

The following theorem shows that projection onto non-empty closed convex sets is unique. The proof is in Section B.1 of the appendix.

Theorem 1.3 (Projection onto convex set)**.** *Let $\mathcal{S} \subseteq \mathbb{R}^n$ be a non-empty closed convex set. The projection of any vector $x \in \mathbb{R}^n$ onto $\mathcal{S}$*

$$\mathcal{P}_{\mathcal{S}}(x) := \arg\min_{s \in \mathcal{S}} \|x - s\|_2 \tag{2}$$

exists and is unique.

Figure 1: An example of a nonconvex set (left) and a convex set (right).

A convex combination of n points is any linear combination of the points with nonnegative coefficients that add up to one. In the case of two points, this is just the segment between the points.

Definition 1.4 (Convex combination)**.** *Given n vectors $x_1, x_2, \ldots, x_n \in \mathbb{R}^n$,*

$$x := \sum_{i=1}^{n} \theta_i x_i \tag{3}$$

is a convex combination of $x_1, x_2, \ldots, x_n$ as along as the real numbers $\theta_1, \theta_2, \ldots, \theta_n$ are non-negative and add up to one,

$$\theta_i \geq 0, \quad 1 \leq i \leq n, \tag{4}$$

$$\sum_{i=1}^{n} \theta_i = 1. \tag{5}$$

The convex hull of a set $\mathcal{S}$ contains all convex combination of points in $\mathcal{S}$. Intuitively, it is the smallest convex set that contains $\mathcal{S}$.

Definition 1.5 (Convex hull)**.** *The convex hull of a set $\mathcal{S}$ is the set of all convex combinations of points in $\mathcal{S}$.*

A justification of why we penalize the ℓ_1-norm to promote sparse structure is that the ℓ_1-norm ball is the convex hull of the intersection between the ℓ_0 "norm" ball and the ℓ_∞-norm ball. The lemma is illustrated in 2D in Figure 2 and proved in Section 1.6 of the appendix.

Lemma 1.6 (ℓ_1-norm ball)**.** *The ℓ_1-norm ball is the convex hull of the intersection between the ℓ_0 "norm" ball and the ℓ_∞-norm ball.*

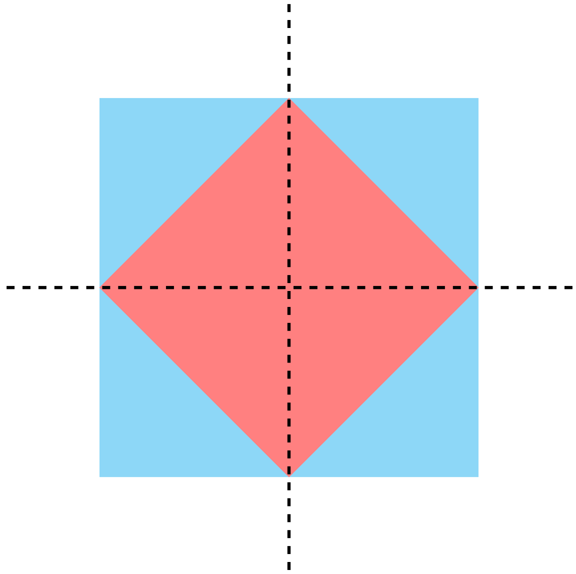

Figure 2: Illustration of Lemma (1.6) The ℓ_0 "norm" ball is shown in black, the ℓ_∞-norm ball in blue and the ℓ_1-norm ball in a reddish color.

1.2 Convex functions

We now define convexity for functions.

Definition 1.7 (Convex function)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for any $x, y \in \mathbb{R}^n$ and any $\theta \in (0, 1)$,*

$$\theta f(x) + (1-\theta) f(y) \geq f(\theta x + (1-\theta) y). \tag{6}$$

The function is strictly convex if the inequality is always strict, i.e. if $x \neq y$ implies that

$$\theta f(x) + (1-\theta) f(y) > f(\theta x + (1-\theta) y). \tag{7}$$

A concave function is a function f such that $-f$ is convex.

Remark 1.8 (Extended-value functions)**.** *We can also consider an arbitrary function f that is only defined in a subset of $\mathbb{R}^n$. In that case f is convex if and only if its extension $\tilde{f}$ is convex, where*

$$\tilde{f}(x) := \begin{cases} f(x) & \text{if } x \in \operatorname{dom}(f), \\ \infty & \text{if } x \notin \operatorname{dom}(f). \end{cases} \tag{8}$$

Equivalently, f is convex if and only if its domain $\operatorname{dom}(f)$ *is convex and any two points in* $\operatorname{dom}(f)$ *satisfy* (6).

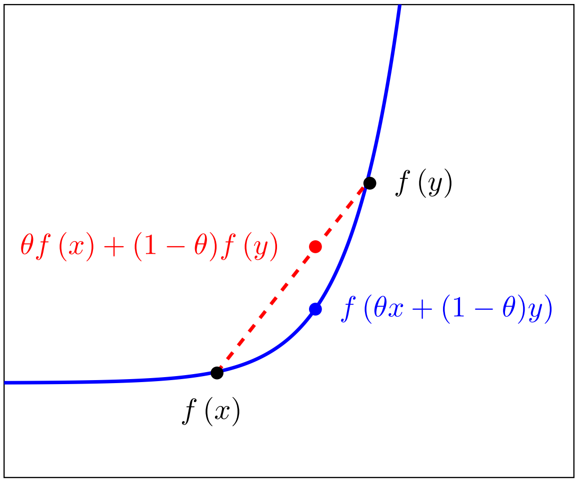

Figure 3: Illustration of condition (6) in Definition 1.7. The curve corresponding to the function must lie below any chord joining two of its points.

Condition (6) is illustrated in Figure 3. The curve corresponding to the function must lie below any chord joining two of its points. It is therefore not surprising that we can determine whether a function is convex by restricting our attention to its behavior along lines in $\mathbb{R}^n$. This is established by the following lemma, which is proved formally in Section B.2 of the appendix.

Lemma 1.9 (Equivalent definition of convex functions)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if for any two points $x, y \in \mathbb{R}^n$ the univariate function $g_{x,y} : [0,1] \to \mathbb{R}$ defined by*

$$g_{x,y}(\alpha) := f(\alpha x + (1-\alpha) y) \tag{9}$$

is convex. Similarly, f is strictly convex if and only if $g_{x,y}$ is strictly convex for any a, b.

Section A in the appendix provides a definition of the norm of a vector and lists the most common ones. It turns out that all norms are convex.

Lemma 1.10 (Norms are convex)**.** *Any valid norm $||\cdot||$ is a convex function.*

Proof. By the triangle inequality inequality and homogeneity of the norm, for any $x, y \in \mathbb{R}^n$ and any $\theta \in (0,1)$

$$||\theta x + (1-\theta) y|| \leq ||\theta x|| + ||(1-\theta) y|| = \theta ||x|| + (1-\theta) ||y|| . \tag{10}$$

□

The ℓ_0 "norm" is not really norm, as explained in Section A, and is not convex either.

Lemma 1.11 (ℓ_0 "norm")**.** *The ℓ_0 "norm" is not convex.*

Proof. We provide a simple counterexample. Let $x := \left(\begin{smallmatrix}1\\0\end{smallmatrix}\right)$ and $y := \left(\begin{smallmatrix}0\\1\end{smallmatrix}\right)$, then for any $\theta \in (0,1)$

$$||\theta x + (1-\theta)\, y||_0 = 2 > 1 = \theta\, ||x||_0 + (1-\theta)\, ||y||_0 \,. \tag{11}$$

□

We end the section by establishing a property of convex functions that is crucial in optimization.

Theorem 1.12 (Local minima are global)**.** *Any local minimum of a convex function $f : \mathbb{R}^n \to \mathbb{R}$ is also a global minimum.*

We defer the proof of the theorem to Section B.4 of the appendix.

1.3 Sublevel sets and epigraph

In this section we define two sets associated to a function that are very useful when reasoning geometrically about convex functions.

Definition 1.13 (Sublevel set)**.** *The γ-sublevel set of a function $f : \mathbb{R}^n \to \mathbb{R}$, where $\gamma \in \mathbb{R}$, is the set of points in $\mathbb{R}^n$ at which the function is smaller or equal to γ,*

$$C_\gamma := \{x \mid f(x) \leq \gamma\} \,. \tag{12}$$

Lemma 1.14 (Sublevel sets of convex functions)**.** *The sublevel sets of a convex function are convex.*

Proof. If $x, y \in \mathbb{R}^n$ belong to the γ-sublevel set of a convex function f then for any $\theta \in (0,1)$

$$f(\theta x + (1-\theta)\, y) \leq \theta f(x) + (1-\theta)\, f(y) \quad \text{by convexity of } f \tag{13}$$

$$\leq \gamma \tag{14}$$

because both x and y belong to the γ-sublevel set. We conclude that any convex combination of x and y also belongs to the γ-sublevel set. □

Recall that the graph of a function $f : \mathbb{R}^n \to \mathbb{R}$ is the curve in $\mathbb{R}^{n+1}$

$$\operatorname{graph}(f) := \{x \mid f(x_{1:n}) = x_{n+1}\} \,, \tag{15}$$

where $x_{1:n} \in \mathbb{R}^n$ contains the first n entries of x. The epigraph of a function is the set in $\mathbb{R}^{n+1}$ that lies above the graph of the function. An example is shown in Figure 4.

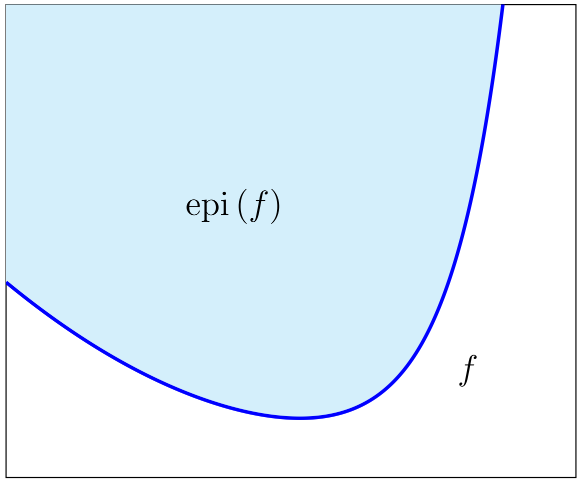

Figure 4: Epigraph of a function.

Definition 1.15 (Epigraph)**.** *The epigraph of a function* $f : \mathbb{R}^n \to \mathbb{R}$ *is defined as*

$$\operatorname{epi}(f) := \{x \mid f(x_{1:n}) \leq x_{n+1}\}. \tag{16}$$

Epigraphs allow to reason geometrically about convex functions. The following basic result is proved in Section B.5 of the appendix.

Lemma 1.16 (Epigraphs of convex functions are convex)**.** *A function is convex if and only if its epigraph is convex.*

1.4 Operations that preserve convexity

It may be challenging to determine whether a function of interest is convex or not by using the definition directly. Often, an easier alternative is to express the function in terms of simpler functions that are known to be convex. In this section we list some operations that preserve convexity.

Lemma 1.17 (Composition of convex and affine function)**.** *If* $f : \mathbb{R}^n \to \mathbb{R}$ *is convex, then for any* $A \in \mathbb{R}^{n \times m}$ *and any* $b \in \mathbb{R}^n$*, the function*

$$h(x) := f(Ax + b) \tag{17}$$

is convex.

Proof. By convexity of f, for any $x, y \in \mathbb{R}^m$ and any $\theta \in (0,1)$

$$
\begin{aligned}
h(\theta x + (1-\theta) y) &= f(\theta(Ax+b) + (1-\theta)(Ay+b)) && (18)\\
&\leq \theta f(Ax+b) + (1-\theta) f(Ay+b) && (19)\\
&= \theta\, h(x) + (1-\theta)\, h(y). && (20)
\end{aligned}
$$

□

Corollary 1.18 (Least squares)**.** *For any $A \in \mathbb{R}^{n\times m}$ and any $y \in \mathbb{R}^n$ the least-squares cost function*

$$||Ax - y||_2 \tag{21}$$

is convex.

Lemma 1.19 (Nonnegative weighted sums)**.** *The weighted sum of m convex functions $f_1, \ldots, f_m$*

$$f := \sum_{i=1}^{m} \alpha_i\, f_i \tag{22}$$

is convex as long as the weights $\alpha_1, \ldots, \alpha \in \mathbb{R}$ are nonnegative.

Proof. By convexity of $f_1, \ldots, f_m$, for any $x, y \in \mathbb{R}^m$ and any $\theta \in (0,1)$

$$
\begin{aligned}
f(\theta x + (1-\theta) y) &= \sum_{i=1}^{m} \alpha_i\, f_i(\theta x + (1-\theta) y) && (23)\\
&\leq \sum_{i=1}^{m} \alpha_i\, (\theta f_i(x) + (1-\theta) f_i(y)) && (24)\\
&= \theta\, f(x) + (1-\theta)\, f(y). && (25)
\end{aligned}
$$

□

Corollary 1.20 (Regularized least squares)**.** *Regularized least-squares cost functions of the form*

$$||Ax - y||_2^2 + ||x||, \tag{26}$$

where $||\cdot||$ is an arbitrary norm, are convex.

Proposition 1.21 (Pointwise maximum/supremum of convex functions)**.** *The pointwise maximum of m convex functions $f_1, \ldots, f_m$ is convex*

$$f_{\max}(x) := \max_{1\leq i\leq m} f_i(x). \tag{27}$$

The pointwise supremum of a family of convex functions indexed by a set $\mathcal{I}$

$$f_{\sup}(x) := \sup_{i \in \mathcal{I}} f_i(x). \tag{28}$$

is convex.

Proof. We prove that the supremum is unique, as it implies the result for the maximum. For any $0 \leq \theta \leq 1$ and any $x, y \in \mathbb{R}$,

$$\begin{aligned}
f_{\sup}(\theta x + (1-\theta) y) &= \sup_{i \in \mathcal{I}} f_i(\theta x + (1-\theta) y) && (29)\\
&\leq \sup_{i \in \mathcal{I}} \theta f_i(x) + (1-\theta) f_i(y) \quad \text{by convexity of the } f_i && (30)\\
&\leq \theta \sup_{i \in \mathcal{I}} f_i(x) + (1-\theta) \sup_{j \in \mathcal{I}} f_j(y) && (31)\\
&= \theta f_{\sup}(x) + (1-\theta) f_{\sup}(y). && (32)
\end{aligned}$$

□

2 Differentiable functions

In this section we characterize the convexity of differentiable functions in terms of the behavior of their first and second order Taylor expansions, or equivalently in terms of their gradient and Hessian.

2.1 First-order conditions

Consider the first-order Taylor expansion of $f : \mathbb{R}^n \to \mathbb{R}$ at x,

$$f_x^1(y) := f(x) + \nabla f(x)(y - x). \tag{33}$$

Note that this first-order approximation is a linear function. The following proposition, proved in Section B.6 of the appendix, establishes that a function f is convex if and only if f_x^1 is a lower bound for f for any $x \in \mathbb{R}^n$. Figure 5 illustrates the condition with an example.

Proposition 2.1 (First-order condition)**.** *A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if for every $x, y \in \mathbb{R}^n$*

$$f(y) \geq f(x) + \nabla f(x)^T (y - x). \tag{34}$$

It is strictly convex if and only if

$$f(y) > f(x) + \nabla f(x)^T (y - x). \tag{35}$$

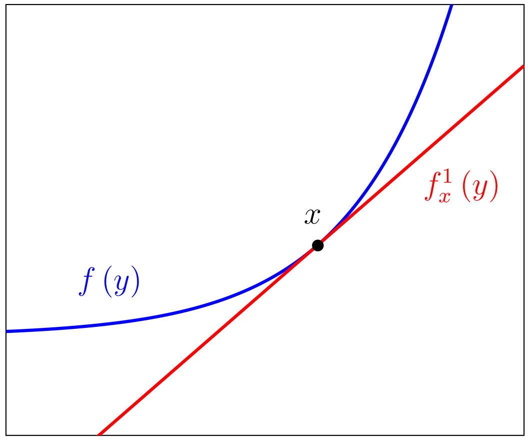

Figure 5: An example of the first-order condition for convexity. The first-order approximation at any point is a lower bound of the function.

An immediate corollary is that for a convex function, any point at which the gradient is zero is a global minimum. If the function is strictly convex, the minimum is unique.

Corollary 2.2. *If a differentiable function f is convex and $\nabla f(x) = 0$, then for any $y \in \mathbb{R}$*

$$f(y) \geq f(x). \tag{36}$$

If f is strictly convex then for any $y \in \mathbb{R}$

$$f(y) > f(x). \tag{37}$$

For any differentiable function f and any $x \in \mathbb{R}^n$ let us define the hyperplane $\mathcal{H}_{f,x} \subset \mathbb{R}^{n+1}$ that corresponds to the first-order approximation of f at x,

$$\mathcal{H}_{f,x} := \left\{ y \mid y_{n+1} = f_x^1(y_{1:n}) \right\}. \tag{38}$$

Geometrically, Proposition 2.1 establishes that $\mathcal{H}_{f,x}$ lies above the epigraph of f. In addition, the hyperplane and $\operatorname{epi}(f)$ intersect at x. In convex analysis jargon, $\mathcal{H}_{f,x}$ is a supporting hyperplane of $\operatorname{epi}(f)$ at x.

Definition 2.3 (Supporting hyperplane)**.** *A hyperplane $\mathcal{H}$ is a supporting hyperplane of a set $\mathcal{S}$ at x if*

- *$\mathcal{H}$ and $\mathcal{S}$ intersect at x,*
- *$\mathcal{S}$ is contained in one of the half-spaces bounded by $\mathcal{H}$.*

The optimality condition has a very intuitive geometric interpretation in terms of the supporting hyperplane $\mathcal{H}_{f,x}$. $\nabla f = 0$ implies that $\mathcal{H}_{f,x}$ is horizontal if the vertical dimension corresponds to the $n+1$th coordinate. Since the epigraph lies above hyperplane, the point at which they intersect must be a minimum of the function.

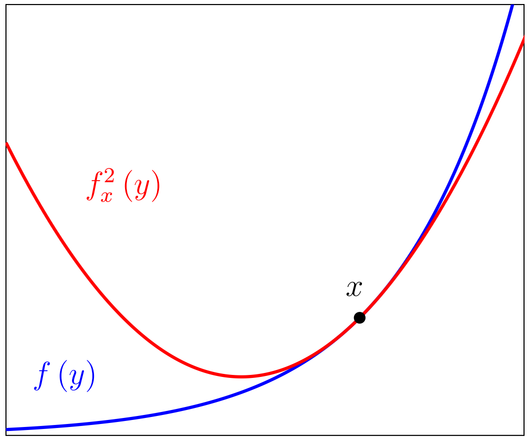

Figure 6: An example of the second-order condition for convexity. The second-order approximation at any point is convex.

2.2 Second-order conditions

For univariate functions that are twice differentiable, convexity is dictated by the curvature of the function. As you might recall from basic calculus, curvature is the rate of change of the slope of the function and is consequently given by its second derivative. The following lemma, proved in Section B.8 of the appendix, establishes that univariate functions are convex if and only if their curvature is always nonnegative.

Lemma 2.4. *A twice-differentiable function $g : \mathbb{R} \to \mathbb{R}$ is convex if and only if $g''(\alpha) \geq 0$ for all $\alpha \in \mathbb{R}$.*

By Lemma 1.9, we can establish convexity in $\mathbb{R}^n$ by considering the restriction of the function along an arbitrary line. By multivariable calculus, the second directional derivative of f at a point x in the direction of a unit vector is equal to $u^T \nabla^2 f(x) u$. As a result, if the Hessian is positive semidefinite, the curvatures is nonnegative in every direction and the function is convex.

Corollary 2.5. *A twice-differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if for every $x \in \mathbb{R}^n$, the Hessian matrix $\nabla^2 f(x)$ is positive semidefinite.*

Proof. By Lemma 1.9 we just need to show that the univariate function $g_{a,b}$ defined by (9) is convex for all $a, b \in \mathbb{R}^n$. By Lemma 2.4 this holds if and only if the second derivative of $g_{a,b}$ is nonnegative. Applying some basic multivariate calculus, we have that for any $\alpha \in (0, 1)$

$$g''_{a,b}(\alpha) := (a - b)^T \nabla^2 f(\alpha a + (1 - \alpha) b)(a - b). \tag{39}$$

This quantity is nonnegative for all $a, b \in \mathbb{R}^n$ if and only if $\nabla^2 f(x)$ is positive semidefinite for any $x \in \mathbb{R}^n$. $\square$

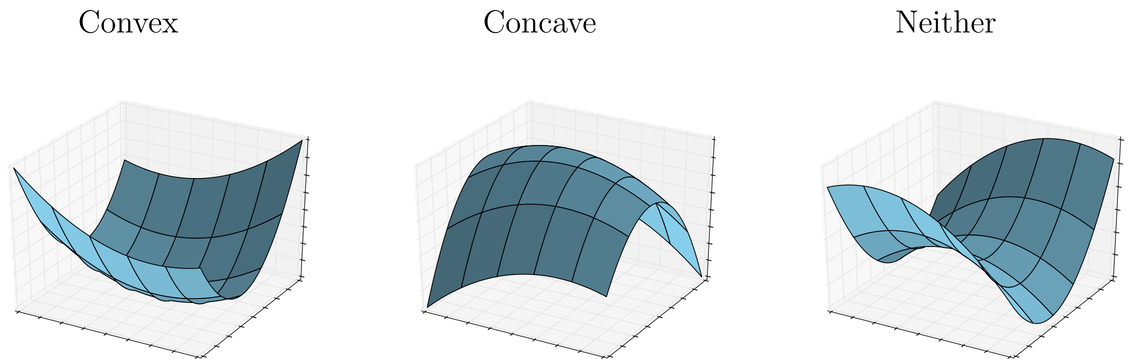

Figure 7: Quadratic forms for which the Hessian is positive definite (left), negative definite (center) and neither positive nor negative definite (right).

Remark 2.6 (Strict convexity)**.** *If the Hessian is positive definite, then the function is strictly convex (the proof is essentially the same). However, there are functions that are strictly convex for which the Hessian may equal zero at some points. An example is the univariate function* $f(x) = x^4$*, for which* $f''(0) = 0$*.*

We can interpret Corollary 2.5 in terms of the second-order Taylor expansion of $f : \mathbb{R}^n \rightarrow \mathbb{R}$ at x.

Definition 2.7 (Second-order approximation)**.** *The second-order or quadratic approximation of* f *at* x *is*

$$f_x^2(y) := f(x) + \nabla f(x)(y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(x)(y - x). \tag{40}$$

f_x^2 is a quadratic form that shares the same value at x. By the corollary, f is convex if and only if this quadratic approximation is always convex. This is illustrated in Figure 6. Figure 7 shows some examples of quadratic forms in two dimensions.

3 Nondifferentiable functions

If a function is not differentiable we cannot use its gradient to check whether it is convex or not. However, we can extend the first-order characterization derived in Section 2.1 by checking whether the function has a supporting hyperplane at every point. If a supporting hyperplane exists at x, the gradient of the hyperplane is called a subgradient of f at x.

Definition 3.1 (Subgradient)**.** *The subgradient of a function $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$ is a vector $q \in \mathbb{R}^n$ such that*

$$f(y) \geq f(x) + q^T(y - x), \quad \text{for all } y \in \mathbb{R}^n. \tag{41}$$

The set of all subgradients is called the subdifferential of the function at x.

The following theorem, proved in Section B.9 of the appendix, establishes that if a subgradient exists at every point, the function is convex.

Theorem 3.2. *If a function $f : \mathbb{R}^n \to \mathbb{R}$ has a non-empty subdifferential at any $x \in \mathbb{R}^n$ then f is convex.*

The subdifferential allows to obtain an optimality condition for nondifferentiable convex functions.

Proposition 3.3 (Optimality condition)**.** *A convex function attains its minimum value at a vector x if the zero vector is a subgradient of f at x.*

Proof. By the definition of subgradient, if $q := 0$ is a subgradient at x for any $y \in \mathbb{R}^n$

$$f(y) \geq f(x) + q^T(y - x) = f(x). \tag{42}$$

□

If a function is differentiable at a point, then the gradient is the only subgradient at that point.

Proposition 3.4 (Subdifferential of differentiable functions)**.** *If a convex function $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at $x \in \mathbb{R}^n$, then its subdifferential at x only contains $\nabla f(x)$.*

Proof. By Proposition 2.1 ∇f is a subgradient at x. Now, let q be an arbitrary subgradient at x. By the definition of subgradient,

$$\begin{aligned} f(x + \alpha\, e_i) &\geq f(x) + q^T \alpha\, e_i && (43)\\ &= f(x) + q_i\, \alpha, && (44)\\ f(x) &\geq f(x - \alpha\, e_i) + q^T \alpha\, e_i && (45)\\ &= f(x - \alpha\, e_i) + q_i\, \alpha. && (46) \end{aligned}$$

Combining both inequalities

$$\frac{f(x) - f(x - \alpha\, e_i)}{\alpha} \leq q_i \leq \frac{f(x + \alpha\, e_i) - f(x)}{\alpha}. \tag{47}$$

If we let $\alpha \to 0$, this implies $q_i = \frac{\partial f(x)}{\partial x_i}$. Consequently, $q = \nabla f$. □

An important nondifferentiable convex function in optimization-based data analysis is the ℓ_1 norm. The following proposition characterizes its subdifferential.

Proposition 3.5 (Subdifferential of ℓ_1 norm)**.** *The subdifferential of the ℓ_1 norm at $x \in \mathbb{R}^n$ is the set of vectors $q \in \mathbb{R}^n$ that satisfy*

$$q_i = \operatorname{sign}(x_i) \quad \text{if } x_i \neq 0, \tag{48}$$

$$|q_i| \leq 1 \qquad \text{if } x_i = 0. \tag{49}$$

Proof. The proof relies on the following simple lemma, proved in Section B.10.

Lemma 3.6. *q is a subgradient of $||\cdot||_1 : \mathbb{R}^n \to \mathbb{R}$ at x if and only if q_i is a subgradient of $|\cdot| : \mathbb{R} \to \mathbb{R}$ at x_i for all $1 \leq i \leq n$.*

If $x_i \neq 0$ the absolute-value function is differentiable at x_i, so by Proposition (3.4), q_i is equal to the derivative $q_i = \operatorname{sign}(x_i)$.

If $x_i = 0$, q_i is a subgradient of the absolute-value function if and only if $|\alpha| \geq q_i \alpha$ for any $\alpha \in \mathbb{R}$, which holds if and only if $|q_i| \leq 1$. □

4 Optimization problems

4.1 Definition

We start by defining a canonical optimization problem. The vast majority of optimization problems (certainly all of the ones that we will study in the course) can be cast in this form.

Definition 4.1 (Optimization problem)**.**

$$\text{minimize} \quad f_0(x) \tag{50}$$

$$\text{subject to} \quad f_i(x) \leq 0, \quad 1 \leq i \leq m, \tag{51}$$

$$h_i(x) = 0, \quad 1 \leq i \leq p, \tag{52}$$

where $f_0, f_1, \ldots, f_m, h_1, \ldots, h_p : \mathbb{R}^n \to \mathbb{R}$.

The problem consists of a cost function f_0, inequality constraints and equality constraints. Any vector that satisfies all the constraints in the problem is said to be feasible. A solution to the problem is any vector x^* such that for all feasible vectors x

$$f_0(x) \geq f_0(x^*). \tag{53}$$

If a solution exists $f(x^*)$ is the optimal value or optimum of the optimization problem.

An optimization problem is convex if it satisfies the following conditions:

- The cost function f_0 is convex.
- The functions that determine the inequality constraints $f_1, \ldots, f_m$ are convex.
- The functions that determine the equality constraints $h_1, \ldots, h_p$ are affine, i.e. $h_i(x) = a_i^T x + b_i$ for some $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$.

Note that under these assumptions the feasibility set is convex. Indeed, it corresponds to the intersection of several convex sets: the 0-sublevel sets of $f_1, \ldots, f_m$, which are convex by Lemma 1.14, and the hyperplanes $h_i(x) = a_i^T x + b_i$. The intersection is convex by Lemma 1.2.

If both the cost function and the constraint functions are all affine, the problem is a linear program (LP).

Definition 4.2 (Linear program)**.**

$$\begin{aligned} \text{minimize} \quad & a^T x && (54)\\ \text{subject to} \quad & c_i^T x \leq d_i, \quad 1 \leq i \leq m, && (55)\\ & Ax = b. && (56) \end{aligned}$$

It turns out that ℓ_1-norm minimization can be cast as an LP. The theorem is proved in Section B.11 of the appendix.

Theorem 4.3 (ℓ_1-norm minimization as an LP)**.** *The optimization problem*

$$\begin{aligned} \text{minimize} \quad & \|x\|_1 && (57)\\ \text{subject to} \quad & Ax = b && (58) \end{aligned}$$

can be recast as the linear program

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{n} t_i && (59)\\ \text{subject to} \quad & t_i \geq x_i, && (60)\\ & t_i \geq -x_i, && (61)\\ & Ax = b. && (62) \end{aligned}$$

If the cost function is a positive semidefinite quadratic form and the constraints are affine the problem is a quadratic program (QP).

Definition 4.4 (Quadratic program)**.**

$$\begin{aligned} \text{minimize} \quad & x^T Q x + a^T x && (63)\\ \text{subject to} \quad & c_i^T x \leq d_i, \quad 1 \leq i \leq m, && (64)\\ & Ax = b, && (65) \end{aligned}$$

where $Q \in \mathbb{R}^{n \times n}$ is positive semidefinite.

A corollary of Theorem 4.3 is that ℓ_1-norm regularized least squares can be cast as a QP.

Corollary 4.5 (ℓ_1-norm regularized least squares as a QP)**.** *The optimization problem*

$$\text{minimize} \quad ||Ax - y||_2^2 + \lambda \, ||x||_1 \tag{66}$$

can be recast as the quadratic program

$$\begin{aligned} \text{minimize} \quad & x^T A^T A x - 2y^T x + \lambda \sum_{i=1}^{n} t_i && (67) \\ \text{subject to} \quad & t_i \geq x_i, && (68) \\ & t_i \geq -x_i. && (69) \end{aligned}$$

We will discuss other types of convex optimization problems, such as semidefinite programs later on in the course.

4.2 Duality

The Lagrangian of the optimization problem in Definition 4.1 is defined as the cost function augmented by a weighted linear combination of the constraint functions,

$$L(x, \lambda, \nu) := f_0(x) + \sum_{i=1}^{m} \lambda_i \, f_i(x) + \sum_{j=1}^{p} \nu_j \, h_j(x), \tag{70}$$

where the vectors $\lambda \in \mathbb{R}^m, \nu \in \mathbb{R}^p$ are called Lagrange multipliers or dual variables. In contrast, x is the primal variable.

Note that as long as $\lambda_i \geq 0$ for $1 \leq i \leq m$, the Lagrangian is a lower bound for the value of the cost function at any feasible point. Indeed, if x is feasible and $\lambda_i \geq 0$ for $1 \leq i \leq m$ then

$$\begin{aligned} \lambda_i \, f_i(x) &\leq 0, && (71) \\ \nu_j \, h_j(x) &= 0, && (72) \end{aligned}$$

for all $1 \leq i \leq m, 1 \leq j \leq p$. This immediately implies,

$$L(x, \lambda, \nu) \leq f_0(x). \tag{73}$$

The Lagrange dual function is the infimum of the Lagrangian over the primal variable x

$$l(\lambda, \nu) := \inf_{x \in \mathbb{R}^n} f_0(x) + \sum_{i=1}^{m} \lambda_i \, f_i(x) + \sum_{j=1}^{p} \nu_j \, h_j(x). \tag{74}$$

Proposition 4.6 (Lagrange dual function as a lower bound of the primal optimum)**.** *Let p^* denote an optimal value of the optimization problem in Definition 4.1,*

$$l(\lambda, \nu) \leq p^*, \tag{75}$$

as long as $\lambda_i \geq 0$ for $1 \leq i \leq n$.

Proof. The result follows directly from (73),

$$\begin{aligned} p^* &= f_0(x^*) && (76)\\ &\geq L(x^*, \lambda, \nu) && (77)\\ &\geq l(\lambda, \nu). && (78) \end{aligned}$$

□

Optimizing this lower bound on the primal optimum over the Lagrange multipliers yields the dual problem of the original optimization problem, which is called the primal problem in this context.

Definition 4.7 (Dual problem)**.** *The dual problem of the optimization problem from Definition 4.1 is*

$$\begin{aligned} \text{maximize} \quad & l(\lambda, \nu) && (79)\\ \text{subject to} \quad & \lambda_i \geq 0, \quad 1 \leq i \leq m. && (80) \end{aligned}$$

Note that the cost function is a pointwise supremum of linear (and hence convex) functions, so by Proposition 1.21 the dual problem is a convex optimization problem even if the primal is nonconvex! The following result, which is an immediate corollary to Proposition 4.6, states that the optimum of the dual problem is a lower bound for the primal optimum. This is known as weak duality.

Corollary 4.8 (Weak duality)**.** *Let p^* denote an optimum of the optimization problem in Definition 4.1 and d^* an optimum of the corresponding dual problem,*

$$d^* \leq p^*. \tag{81}$$

In the case of convex functions, the optima of the primal and dual problems are often equal, i.e.

$$d^* = p^*. \tag{82}$$

This is known as strong duality. A simple condition that guarantees strong duality for convex optimization problems is Slater's condition.

Definition 4.9 (Slater's condition)**.** *A vector $x \in \mathbb{R}^n$ satisfies Slater's condition for a convex optimization problem if*

$$f_i(x) < 0, \quad 1 \leq i \leq m, \tag{83}$$

$$Ax = b. \tag{84}$$

A proof of strong duality under Slater's condition can be found in Section 5.3.2 of [1].

References

A very readable and exhaustive reference is Boyd and Vandenberghe's seminal book on convex optimization [1], which unfortunately does not cover subgradients.[1] Nesterov's book [2] and Rockafellar's book [3] do cover subgradients.

[1] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[2] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*.

[3] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*. Springer Science & Business Media, 2009.

A Norms

The norm of a vector is a generalization of the concept of *length*.

Definition A.1 (Norm)**.** *Let $\mathcal{V}$ be a vector space, a norm is a function $||\cdot||$ from $\mathcal{V}$ to $\mathbb{R}$ that satisfies the following conditions.*

- *It is homogeneous. For all $\alpha \in \mathbb{R}$ and $x \in \mathcal{V}$*

$$||\alpha\, x|| = |\alpha|\, ||x|| \,. \tag{85}$$

- *It satisfies the triangle inequality*

$$||x + y|| \leq ||x|| + ||y|| \,. \tag{86}$$

  In particular, it is nonnegative (set $y = -x$).

- $||x|| = 0$ *implies that x is the zero vector 0.*

[1] However, see http://see.stanford.edu/materials/lsocoee364b/01-subgradients_notes.pdf

The ℓ_2 norm is induced by the inner product $\langle x, y\rangle = x^T y$

$$||x||_2 := \sqrt{x^T x}. \tag{87}$$

Definition A.2 (ℓ_2 norm)**.** *The ℓ_2 norm of a vector $x \in \mathbb{R}^n$ is defined as*

$$||x||_2 := \sqrt{\sum_{i=1}^{n} x_i^2}. \tag{88}$$

Definition A.3 (ℓ_1 norm)**.** *The ℓ_1 norm of a vector $x \in \mathbb{R}^n$ is defined as*

$$||x||_1 := \sum_{i=1}^{n} |x_i| \,. \tag{89}$$

Definition A.4 (ℓ_∞ norm)**.** *The ℓ_∞ norm of a vector $x \in \mathbb{R}^n$ is defined as*

$$||x||_\infty := \max_{1\leq i\leq n} |x_i| \,. \tag{90}$$

Remark A.5. *The ℓ_0 "norm" is not a norm, as it is not homogeneous. For example, if x is not the zero vector,*

$$||2x||_0 = ||x||_0 \neq 2\, ||x||_0 \,. \tag{91}$$

B Proofs

B.1 Proof of Theorem 1.3

Existence

Since $\mathcal{S}$ is non-empty we can choose an arbitrary point $\tilde{s} \in \mathcal{S}$. Minimizing $||x - s||_2$ over $\mathcal{S}$ is equivalent to minimizing $||x - s||_2$ over $\mathcal{S} \cap \{y \mid \; ||x - y||_2 \leq ||x - \tilde{s}||_2\}$. Indeed, the solution cannot be a point that is farther away from x than $\tilde{s}$. By Weierstrass's extreme-value theorem, the optimization problem

$$\begin{aligned} \text{minimize} \quad & ||x - s||_2^2 && (92)\\ \text{subject to} \quad & s \in \mathcal{S} \cap \{y \mid \; ||x - y||_2 \leq ||x - \tilde{s}||_2\} && (93) \end{aligned}$$

has a solution because $||x - s||_2^2$ is a continuous function and the feasibility set is bounded and closed, and hence compact. Note that this also holds if $\mathcal{S}$ is not convex.

Uniqueness

Assume that there are two distinct projections $s_1 \neq s_2$. Consider the point

$$s := \frac{s_1 + s_2}{2}, \tag{94}$$

which belongs to $\mathcal{S}$ because $\mathcal{S}$ is convex. The difference between x and s and the difference between s_1 and s are orthogonal vectors,

$$\begin{aligned}
\langle x - s, s_1 - s \rangle &= \left\langle x - \frac{s_1 + s_2}{2}, s_1 - \frac{s_1 + s_2}{2} \right\rangle && (95)\\
&= \left\langle \frac{x - s_1}{2} + \frac{x - s_2}{2}, \frac{x - s_1}{2} - \frac{x - s_2}{2} \right\rangle && (96)\\
&= \frac{1}{4} \left(||x - s_1||^2 + ||x - s_2||^2 \right) && (97)\\
&= 0, && (98)
\end{aligned}$$

because $||x - s_1|| = ||x - s_2||$ by assumption. By Pythagoras's theorem this implies

$$\begin{aligned}
||x - s_1||_2^2 &= ||x - s||_2^2 + ||s_1 - s||_2^2 && (99)\\
&= ||x - s||_2^2 + \left|\left| \frac{s_1 - s_2}{2} \right|\right|_2^2 && (100)\\
&> ||x - s||_2^2 && (101)
\end{aligned}$$

because $s_1 \neq s_2$ by assumption. We have reached a contradiction, so the projection is unique.

B.2 Proof of Lemma 1.9

The proof for strict convexity is exactly the same, replacing the inequalities by strict inequalities.

<u>f being convex implies that $g_{x,y}$ is convex for any $x, y \in \mathbb{R}^n$</u>

For any $\alpha, \beta, \theta \in (0, 1)$

$$\begin{aligned}
g_{x,y}\left(\theta\alpha + (1-\theta)\beta\right) &= f\left(\left(\theta\alpha + (1-\theta)\beta\right)x + \left(1 - \theta\alpha - (1-\theta)\beta\right)y\right) && (102)\\
&= f\left(\theta\left(\alpha x + (1-\alpha)y\right) + (1-\theta)\left(\beta x + (1-\beta)y\right)\right) && (103)\\
&\leq \theta f\left(\alpha x + (1-\alpha)y\right) + (1-\theta) f\left(\beta x + (1-\beta)y\right) && \text{by convexity of } f\\
&= \theta g_{x,y}(\alpha) + (1-\theta)\, g_{x,y}(\beta). && (104)
\end{aligned}$$

<u>$g_{x,y}$ being convex for any $x, y \in \mathbb{R}^n$ implies that f is convex</u>

For any $\alpha, \beta, \theta \in (0,1)$

$$f(\theta x + (1-\theta) y) = g_{x,y}(\theta) \tag{105}$$
$$\leq \theta g_{x,y}(1) + (1-\theta) g_{x,y}(0) \quad \text{by convexity of } g_{x,y} \tag{106}$$
$$= \theta f(x) + (1-\theta) f(y). \tag{107}$$

B.3 Proof of Lemma 1.6

We prove that the ℓ_1-norm ball $\mathcal{B}_{\ell_1}$ is equal to the convex hull of the intersection between the ℓ_0 "norm" ball $\mathcal{B}_{\ell_0}$ and the ℓ_∞-norm ball $\mathcal{B}_{\ell_\infty}$ by showing that the sets contain each other.

$\underline{\mathcal{B}_{\ell_1} \subseteq \mathcal{C}(\mathcal{B}_{\ell_0} \cap \mathcal{B}_{\ell_\infty})}$

Let x be an n-dimensional vector in $\mathcal{B}_{\ell_1}$. If we set $\theta_i := |x(i)|$, where $x(i)$ is the ith entry of x by $x(i)$, and $\theta_0 = 1 - \sum_{i=1}^n \theta_i$ we have

$$\sum_{i=1}^{n+1} \theta_i = 1, \tag{108}$$
$$\theta_i \geq 0 \quad \text{for } 1 \leq i \leq n \text{ by definition,} \tag{109}$$
$$\theta_0 = 1 - \sum_{i=1}^{n+1} \theta_i \tag{110}$$
$$= 1 - ||x||_1 \tag{111}$$
$$\geq 0 \quad \text{because } x \in \mathcal{B}_{\ell_1}. \tag{112}$$

We can express x as a convex combination of the standard basis vectors $e_1, e_2, \ldots, e_n$, which belong to $\mathcal{B}_{\ell_0} \cap \mathcal{B}_{\ell_\infty}$ since they have a single nonzero entry equal to one, and the zero vector e_0, which also belongs to $\mathcal{B}_{\ell_0} \cap \mathcal{B}_{\ell_\infty}$,

$$x = \sum_{i=1}^{n} \theta_i e_i + \theta_0 e_0. \tag{113}$$

$\underline{\mathcal{C}(\mathcal{B}_{\ell_0} \cap \mathcal{B}_{\ell_\infty}) \subseteq \mathcal{B}_{\ell_1}}$

Let x be an n-dimensional vector in $\mathcal{C}(\mathcal{B}_{\ell_0} \cap \mathcal{B}_{\ell_\infty})$. By the definition of convex hull, we can write

$$x = \sum_{i=1}^{m} \theta_i y_i, \tag{114}$$

where $m > 0$, $y_1, \ldots, y_m \in \mathbb{R}^n$ have a single entry bounded by one, $\theta_i \geq 0$ for all $1 \leq i \leq m$ and $\sum_{i=1}^{m} \theta_i = 1$. This immediately implies $x \in \mathcal{B}_{\ell_1}$, since

$$\begin{aligned}
||x||_1 &\leq \sum_{i=1}^{m} \theta_i \, ||y_i||_1 \quad \text{by the Triangle inequality} && (115)\\
&\leq \sum_{i=1}^{n} \theta_i \, ||y_i||_\infty \quad \text{because each } y_i \text{ only has one nonzero entry} && (116)\\
&\leq \sum_{i=1}^{n} \theta_i && (117)\\
&\leq 1. && (118)
\end{aligned}$$

B.4 Proof of Theorem 1.12

We prove the result by contradiction. Let x_{loc} be a local minimum and x_{glob} a global minimum such that $f(x_{\text{glob}}) < f(x_{\text{loc}})$. Since x_{loc} is a local minimum, there exists $\gamma > 0$ for which $f(x_{\text{loc}}) \leq f(x)$ for all $x \in \mathbb{R}^n$ such that $||x - x_{\text{loc}}||_2 \leq \gamma$. If we choose $\theta \in (0,1)$ small enough, $x_\theta := \theta x_{\text{loc}} + (1-\theta) x_{\text{glob}}$ satisfies $||x - x_{\text{loc}}||_2 \leq \gamma$ and therefore

$$\begin{aligned}
f(x_{\text{loc}}) &\leq f(x_\theta) && (119)\\
&\leq \theta f(x_{\text{loc}}) + (1-\theta) f(x_{\text{glob}}) \quad \text{by convexity of } f && (120)\\
&< f(x_{\text{loc}}) \quad \text{because } f(x_{\text{glob}}) < f(x_{\text{loc}}). && (121)
\end{aligned}$$

B.5 Proof of Lemma 1.16

f being convex implies that $\operatorname{epi}(f)$ is convex

Let $x, y \in \mathbb{R}^{n+1} \in \operatorname{epi}(f)$, then for any $\theta \in (0,1)$

$$\begin{aligned}
f(\theta x_{1:n} + (1-\theta) y_{1:n}) &\leq \theta f(x_{1:n}) + (1-\theta) f(y_{1:n}) \quad \text{by convexity of } f && (122)\\
&\leq \theta x_{n+1} + (1-\theta) y_{n+1} && (123)
\end{aligned}$$

because $x, y \in \mathbb{R}^{n+1} \in \operatorname{epi}(f)$ so $f(x_{1:n}) \leq x_{n+1}$ and $f(y_{1:n}) \leq y_{n+1}$. This implies that $\theta x + (1-\theta) y \in \operatorname{epi}(f)$.

$\operatorname{epi}(f)$ being convex implies that f is convex

For any $x, y \in \mathbb{R}^n$, let us define $\tilde{x}, \tilde{y} \in \mathbb{R}^{n+1}$ such that

$$\begin{aligned}
\tilde{x}_{1:n} &:= x, \quad \tilde{x}_{n+1} := f(x), && (124)\\
\tilde{y}_{1:n} &:= y, \quad \tilde{y}_{n+1} := f(y). && (125)
\end{aligned}$$

By definition of $\mathrm{epi}(f)$, $\tilde{x}, \tilde{y} \in \mathrm{epi}(f)$. For any $\theta \in (0,1)$, $\theta\tilde{x} + (1-\theta)\tilde{y}$ belongs to $\mathrm{epi}(f)$ because it is convex. As a result,

$$\begin{aligned} f(\theta x + (1-\theta) y) &= f(\theta \tilde{x}_{1:n} + (1-\theta)\tilde{y}_{1:n}) && (126)\\ &\leq \theta \tilde{x}_{n+1} + (1-\theta)\tilde{y}_{n+1} && (127)\\ &= \theta f(x) + (1-\theta) f(y). && (128)\end{aligned}$$

B.6 Proof of Proposition 2.1

The proof for strict convexity is almost exactly the same; we omit the details.

The following lemma, proved in Section B.7 below establishes that the result holds for univariate functions

Lemma B.1. *A univariate differentiable function $g : \mathbb{R} \to \mathbb{R}$ is convex if and only if for all $\alpha, \beta \in \mathbb{R}$*

$$g(\beta) \geq g'(\alpha)(\beta - \alpha) \tag{129}$$

and strictly convex if and only if for all $\alpha, \beta \in \mathbb{R}$

$$g(\beta) > g'(\alpha)(\beta - \alpha). \tag{130}$$

To complete the proof we extend the result to the multivariable case using Lemma 1.9.

<u>If $f(y) \geq f(x) + \nabla f(x)^T (y-x)$ for any $x, y \in \mathbb{R}^n$ then f is convex</u>

By Lemma 1.9 we just need to show that the univariate function $g_{a,b}$ defined by (9) is convex for all $a, b \in \mathbb{R}^n$. Applying some basic multivariate calculus yields

$$g'_{a,b}(\alpha) = \nabla f(\alpha\, a + (1-\alpha) b)^T (a-b). \tag{131}$$

Let $\alpha, \beta \in \mathbb{R}$. Setting $x := \alpha\, a + (1-\alpha) b$ and $y := \beta\, a + (1-\beta) b$ we have

$$\begin{aligned} g_{a,b}(\beta) &= f(y) && (132)\\ &\geq f(x) + \nabla f(x)^T (y-x) && (133)\\ &= f(\alpha\, a + (1-\alpha) b) + \nabla f(\alpha\, a + (1-\alpha) b)^T (a-b)(\beta-\alpha) && (134)\\ &= g_{a,b}(\alpha) + g'_{a,b}(\alpha)(\beta - \alpha) \quad \text{by (131)}, && (135)\end{aligned}$$

which establishes that $g_{a,b}$ is convex by Lemma B.1 above.

<u>If f is convex then $f(y) \geq f(x) + \nabla f(x)^T (y-x)$ for any $x, y \in \mathbb{R}^n$</u>

By Lemma 1.9, $g_{x,y}$ is convex for any $x, y \in \mathbb{R}^n$.

$$\begin{aligned}
f(y) &= g_{x,y}(1) && (136)\\
&\geq g_{x,y}(0) + g'_{x,y}(0) \quad \text{by convexity of } g_{x,y} \text{ and Lemma B.1} && (137)\\
&= f(x) + \nabla f(x)^T (y - x) \quad \text{by (131).} && (138)
\end{aligned}$$

B.7 Proof of Lemma B.1

g being convex implies $g(\beta) \geq g'(\alpha)(\beta - \alpha)$ for all $\alpha, \beta \in \mathbb{R}$

If g is convex then for any $\alpha, \beta \in \mathbb{R}$ and any $0 \leq \theta \leq 1$

$$\theta (g(\beta) - g(\alpha)) + g(\alpha) \geq g(\alpha + \theta(\beta - \alpha)). \tag{139}$$

Rearranging the terms we have

$$g(\beta) \geq \frac{g(\alpha + \theta(\beta - \alpha)) - g(\alpha)}{\theta} + g(\alpha). \tag{140}$$

Setting $h = \theta(\beta - \alpha)$, this implies

$$g(\beta) \geq \frac{g(\alpha + h) - g(\alpha)}{h}(\beta - \alpha) + g(\alpha). \tag{141}$$

Taking the limit when $h \to 0$ yields

$$g(\beta) \geq g'(\alpha)(\beta - \alpha). \tag{142}$$

If $g(\beta) \geq g'(\alpha)(\beta - \alpha)$ for all $\alpha, \beta \in \mathbb{R}$ then g is convex

Let $z = \theta\alpha + (1 - \theta)\beta$, then by if $g(\beta) \geq g'(\alpha)(\beta - \alpha)$

$$\begin{aligned}
g(\alpha) &\geq g'(z)(\alpha - z) + g(z) && (143)\\
&= g'(z)(1 - \theta)(\alpha - \beta) + g(z) && (144)\\
g(\beta) &\geq g'(z)(\beta - z) + g(z) && (145)\\
&= g'(z)\,\theta(\beta - \alpha) + g(z) && (146)
\end{aligned}$$

Multiplying (144) by θ, then (146) by $1 - \theta$ and summing the inequalities, we obtain

$$\theta g(\alpha) + (1 - \theta) g(\beta) \geq g(\theta\alpha + (1 - \theta)\beta). \tag{147}$$

B.8 Proof of Lemma 2.4

The second derivative of g is nonnegative anywhere if and only if the first derivative is nondecreasing, because g'' is the derivative of g'.

If g is convex g' is nondecreasing

By Lemma B.1, if the function is convex then for any $\alpha, \beta \in \mathbb{R}$ such that $\beta > \alpha$

$$g(\alpha) \geq g'(\beta)(\alpha - \beta) + g(\beta), \tag{148}$$

$$g(\beta) \geq g'(\alpha)(\beta - \alpha) + g(\alpha). \tag{149}$$

Rearranging, we obtain

$$g'(\beta)(\beta - \alpha) \geq g(\beta) - g(\alpha) \geq g'(\alpha)(\beta - \alpha). \tag{150}$$

Since $\beta - \alpha > 0$, we have $g'(\beta) \geq g'(\alpha)$.

If g' is nondecreasing, g is convex

For arbitrary $\alpha, \beta, \theta \in \mathbb{R}$, such that $\beta > \alpha$ and $0 < \theta < 1$, let $\eta = \theta\beta + (1 - \theta)\alpha$. Since $\beta > \eta > \alpha$, by the mean-value theorem there exist $\gamma_1 \in [\alpha, \eta]$ and $\gamma_2 \in [\eta, \beta]$ such that

$$g'(\gamma_1) = \frac{g(\eta) - g(\alpha)}{\eta - \alpha}, \tag{151}$$

$$g'(\gamma_2) = \frac{g(\beta) - g(\eta)}{\beta - \eta}. \tag{152}$$

Since $\gamma_1 < \gamma_2$, if g' is nondecreasing

$$\frac{g(\beta) - g(\eta)}{\beta - \eta} \geq \frac{g(\eta) - g(\alpha)}{\eta - \alpha}, \tag{153}$$

which implies

$$\frac{\eta - \alpha}{\beta - \alpha} g(\beta) + \frac{\beta - \eta}{\beta - \alpha} g(\alpha) \geq g(\eta). \tag{154}$$

Recall that $\eta = \theta\beta + (1 - \theta)\alpha$, so that $\theta = (\eta - \alpha) / (\beta - \alpha)$ and $\theta = (\eta - \alpha) / (\beta - \alpha)$ and $1 - \theta = (\beta - \eta) / (\beta - \alpha)$. (154) is consequently equivalent to

$$\theta g(\beta) + (1 - \theta) g(\alpha) \geq g(\theta\beta + (1 - \theta)\alpha). \tag{155}$$

B.9 Proof of Theorem 3.2

For arbitrary $x, y \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ there exists a subgradient q of f at $\alpha x + (1-\alpha) y$. This implies

$$\begin{aligned} f(y) &\geq f(\alpha x + (1-\alpha) y) + q^T (y - \alpha x - (1-\alpha) y) && (156)\\ &= f(\alpha x + (1-\alpha) y) + \alpha\, q^T (y - x), && (157)\\ f(x) &\geq f(\alpha x + (1-\alpha) y) + q^T (x - \alpha x - (1-\alpha) y) && (158)\\ &= f(\alpha x + (1-\alpha) y) + (1-\alpha)\, q^T (y - x). && (159) \end{aligned}$$

Multiplying equation (157) by $1-\alpha$ and equation (159) by α and adding them together yields

$$\alpha f(x) + (1-\alpha) f(y) \geq f(\alpha x + (1-\alpha) y). \tag{160}$$

B.10 Proof of Lemma 3.6

<u>If q is a subgradient for $||\cdot||_1$ at x then q_i is a subgradient for $|\cdot|$ at $|x_i|$ for $1 \leq i \leq n$</u>

$$\begin{aligned} |y_i| &= ||x + (y_i - x_i) e_i||_1 - ||x||_1 && (161)\\ &\geq q^T (y_i - x_i) e_i && (162)\\ &= q_i (y_i - x_i). && (163) \end{aligned}$$

<u>If q_i is a subgradient for $|\cdot|$ at $|x_i|$ for $1 \leq i \leq n$ then q is a subgradient for $||\cdot||_1$ at x</u>

$$\begin{aligned} ||y||_1 &= \sum_{i=1}^{n} |y_i| && (164)\\ &\geq \sum_{i=1}^{n} |x_i| + q_i (y_i - x_i) && (165)\\ &= ||x||_1 + q^T (y - x). && (166) \end{aligned}$$

B.11 Proof of Theorem 4.3

To show that the linear problem and the ℓ_1-norm minimization problem are equivalent, we show that they have the same set of solutions.

Let us denote an arbitrary solution of the LP by $\left(x^{\mathrm{lp}}, t^{\mathrm{lp}}\right)$. For any solution x^{ℓ_1} of the ℓ_1-norm minimization problem, we define $t_i^{\ell_1} := \left|x_i^{\ell_1}\right|$. $\left(x^{\ell_1}, t^{\ell_1}\right)$ is feasible for the LP so

$$\begin{aligned}
\left|\left|x^{\ell_1}\right|\right|_1 &= \sum_{i=1}^{n} t_i^{\ell_1} && (167)\\
&\geq \sum_{i=1}^{n} t_i^{\mathrm{lp}} \quad \text{by optimality of } t^{\mathrm{lp}} && (168)\\
&\geq \left|\left|x^{\mathrm{lp}}\right|\right|_1 \quad \text{by constraints (60) and (61).} && (169)
\end{aligned}$$

This implies that any solution of the LP is also a solution of the ℓ_1-norm minimization problem.

To prove the converse, we fix a solution x^{ℓ_1} of the ℓ_1-norm minimization problem. Setting $t_i^{\ell_1} := \left|x_i^{\ell_1}\right|$, we show that $\left(x^{\ell_1}, t^{\ell_1}\right)$ is a solution of the LP. Indeed,

$$\begin{aligned}
\sum_{i=1}^{n} t_i^{\ell_1} &= \left|\left|x^{\ell_1}\right|\right|_1 && (170)\\
&\leq \left|\left|x^{\mathrm{lp}}\right|\right|_1 \quad \text{by optimality of } x^{\ell_1} && (171)\\
&\leq \sum_{i=1}^{n} t_i^{\mathrm{lp}} \quad \text{by constraints (60) and (61).} && (172)
\end{aligned}$$