Optimization

Abijith Jagannath Kamath

Contents

1 Convex Sets and Convex Functions **1**
1.1 Differentiable Functions 2
1.2 Optima of Convex Functions 2
1.3 First-order Conditions 2

2 Notions of Nonexpansiveness **5**

3 The Gradient-Descent Algorithm **7**
3.1 Convergence Analysis of Gradient Descent 8
3.2 Convergence Analysis of Gradient Descent Using Contraction Mapping Theory 9

4 Constrained Optimisation and Projected Gradient Descent **11**
4.1 Convergence Analysis of Projected Gradient Descent 12
4.2 Extended-Valued Functions 12
4.3 Lower Semicontinuity 14

5 Subgradient Methods **16**
5.1 Directional Derivatives and Support Functions 19

6 Proximal Methods **22**
6.1 Projected Subgradient Descent 22
6.2 Proximal Gradient Method 22

7 Fenchel Conjugate **25**
7.1 Properness of Fenchel Conjugates 25
7.2 Biconjugate Theorem 25

1 Convex Sets and Convex Functions

Definition 1.1 (Affine sets and Convex sets)
Let X be a linear space (vector space) equipped with an inner product denoted as $\langle \cdot, \cdot \rangle$.

1. $C \subseteq X$ is an affine set if

$$(\forall x_1, x_2 \in C, \forall \theta \in \mathbb{R}), \quad \theta x_1 + (1-\theta) x_2 \in C.$$

2. $C \subseteq X$ is a convex set if

$$(\forall x_1, x_2 \in C, \forall \theta \in [0,1]), \quad \theta x_1 + (1-\theta) x_2 \in C.$$

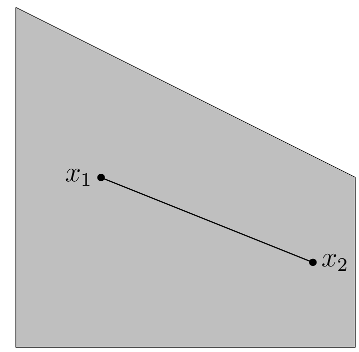

Figure 1: Every line segment of the form $[x_1, x_2]$ are completely inside the set.

The topology of convex sets may be open, for instance $]0,1[$ or closed, for instance $[0,1]$.

Definition 1.2 (Convex functions)
Let C be a convex set. $f : C \to \mathbb{R}$ is a convex function if

$$(\forall x_1, x_2 \in X, \forall \theta \in [0,1]), \;\; f(\theta x_1 + (1-\theta) x_2) \leq \theta f(x_1) + (1-\theta) f(x_2).$$

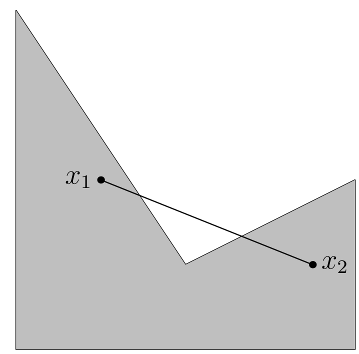

Figure 2: There is at least one line segment $[x_1, x_2]$ that is outside the set.

Definition 1.3 (Optimisation programme, minimum and minimiser)
Let $f : C \to \mathbb{R}$. We define an optimisation (minimisation) programme as

$$\underset{x \in \Omega}{\text{minimise}} \; f(x). \tag{P}$$

- f is called the objective function,
- Ω is called the feasible set.

Remark Well-posed and solvable optimisation programmes
Consider the canonical optimisation programme (P).

1. (P) is <u>well-posed</u> if there exists $f^* = \inf_{x \in \Omega} f(x)$. Then, f^* is called the <u>minimum value</u> of f.

2. (P) is <u>solvable</u> if there exists $x^* = \arg\min_{x \in \Omega} f(x)$, Then, x^* is called the (global) <u>minimiser</u> of f.

Definition 1.4 (Local minimiser)
Let $f : X \to \mathbb{R}$. x^* is called a local minimiser of f if $\exists \epsilon > 0$ such that

$$\forall x \in \mathcal{B}(x^*, \epsilon), \; f(x) \geq f(x^*),$$

where $\mathcal{B}(x^*, \epsilon) = \{y : \|y - x^*\| < \epsilon\}$ denotes the norm ball with centre x^* and radius ϵ.

1.1 Differentiable Functions

Definition 1.5 (Fréchet differentiable functions and gradient)
Let $f : X \to \mathbb{R}$, and $x \in X$. We say f is Fréchet differentiable at $x \in X$ if $\exists! g_x \in X$ such that

$$\forall h \in X,\ f(x+h) = f(x) + \langle g_x, h \rangle + o(\|h\|).$$

We write $g_x = \nabla f(x)$.

Exercise (Show g_x is unique)**.** Suppose there exists $g_x, \bar{g}_x \in X$ such that

$$\lim_{h \to 0} \frac{f(x+h) - f(x) - \langle g_x, h \rangle}{\|h\|}, \text{ and } \lim_{h \to 0} \frac{f(x+h) - f(x) - \langle \bar{g}_x, h \rangle}{\|h\|}.$$

Taking the difference, we have:

$$\lim_{h \to 0} \frac{\langle g_x - \bar{g}_x, h \rangle}{\|h\|} = 0,$$

i.e., a linear functional is identically zero, which implies $g_x = \bar{g}_x$. □

1.2 Optima of Convex Functions

Theorem 1.6 (Local minima of convex functions are global minima)
Let $x^* \in \Omega$ be a local minimiser of a convex function $f : \Omega \to \mathbb{R}$. Then, x^* is the global minimiser of f.

1.3 First-order Conditions

Theorem 1.7 (First-order conditions for convexity)
Let $f \in C^1(\mathbb{R}^n)$. Then, f is convex iff

$$(\forall x, y \in \mathbb{R}^n), \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$

Geometric interpretation For convex functions $f \in C^1(\mathbb{R})$, the tangent at a point x_0:

$$y(x) = f'(x_0)(x - x_0) + f(x_0) \leq f(x),$$

is a global underestimator to f.

Corollary 1.8 (Stationary points of convex functions)
Let $f \in C^1(\mathbb{R}^n)$ be convex. Then,

$$\nabla f(x^*) = 0 \iff x^* \in \arg\min_{x \in \mathbb{R}^n} f(x).$$

Corollary 1.9 (Convex functions and monotone gradients)
Let $f \in C^1(\mathbb{R}^n)$. Then, f is convex if and only if the mapping $x \mapsto \nabla f(x)$ is monotone, i.e.,

$$(\forall x, y \in \mathbb{R}^n), \quad \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0.$$

Definition 1.10 (Smooth functions)
A function $f \in C^1(X)$ is β-smooth, if ∇f is Lipschitz, i.e., $\exists \beta > 0$ such that,

$$\forall x, y \in X,\ \|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|.$$

Lemma 1.11 (Smooth functions have quadratic *majorisers*)
Let f be β-smooth. Then, $\forall x, y \in \mathbb{R}^n$,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2}\|y - x\|^2.$$

In particular, if f is convex, then, ∇f is strongly-monotonic with

$$(\forall x, y \in \mathbb{R}^n), \quad \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{\beta}\|\nabla f(x) - \nabla f(y)\|^2.$$

Definition 1.12 (Strongly convex function)
A function $f : \mathbb{R}^n \to \mathbb{R}$ is called σ-strongly convex, $\sigma > 0$, if the function $f - \frac{\sigma}{2}\|\cdot\|^2$ is convex.

Exercise. Let f be σ-strongly convex and β-smooth. Show that $\beta \geq \sigma$.
(Checkpoint) Suppose $f \in C^2(\mathbb{R}^n)$. Then, it is easy to verify that,

1. β-smoothness $\implies \nabla^2 f(x) \leq \beta I,\ \forall x \in \mathbb{R}^n$,
2. σ-strongly convex $\implies \nabla^2 g(x) = \nabla^2 f(x) - \sigma I \geq 0,\ \forall x \in \mathbb{R}^n$,

using the second-order test for convexity. Therefore, we have

$$\sigma I \leq \nabla^2 f(x) \leq \beta I,\ \forall x \in \mathbb{R}^n,$$

and in particular, $\sigma \leq \beta$. This gives a tractable method to compute the bounds σ, β using the eigenvalues of the Hessian matrix.

Can this result be shown in general? While smooth functions have a quadratic majoriser, strongly convex functions have a quadratic minoriser.

Lemma 1.13 (Strongly convex functions have quadratic minorisers)
Let $f \in C^1(\mathbb{R})$ and σ-strongly-convex. Then, $\forall x, y \in \mathbb{R}^n$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\sigma}{2}\|y - x\|^2.$$

Further, if f is β-smooth,

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{\sigma + \beta}\|\nabla f(x) - \nabla f(y)\|^2 + \frac{\beta\sigma}{\sigma + \beta}\|x - y\|^2.$$

Definition 1.14 (Strongly convex function)
A function $f : \mathbb{R}^n \to \mathbb{R}$ is called σ-strongly convex, $\sigma > 0$, if the function $f - \frac{\sigma}{2}\|\cdot\|^2$ is convex.

Exercise. Let f be σ-strongly convex and β-smooth. Show that $\beta \geq \sigma$.

(Checkpoint) Suppose $f \in C^2(\mathbb{R}^n)$. Then, it is easy to verify that,

1. β-smoothness $\implies \nabla^2 f(x) \leq \beta I, \; \forall x \in \mathbb{R}^n$,
2. σ-strongly convex $\implies \nabla^2 g(x) = \nabla^2 f(x) - \sigma I \geq 0, \; \forall x \in \mathbb{R}^n$,

using the second-order test for convexity. Therefore, we have

$$\sigma I \leq \nabla^2 f(x) \leq \beta I, \; \forall x \in \mathbb{R}^n,$$

and in particular, $\sigma \leq \beta$. This gives a tractable method to compute the bounds σ, β using the eigenvalues of the Hessian matrix.

Can this result be shown in general? While smooth functions have a quadratic majoriser, strongly convex functions have a quadratic minoriser.

Lemma 1.15 (Strongly convex functions have quadratic minorisers)
Let $f \in C^1(\mathbb{R})$ and σ-strongly-convex. Then, $\forall x, y \in \mathbb{R}^n$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\sigma}{2}\|y - x\|^2.$$

Further, if f is β-smooth,

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{\sigma + \beta}\|\nabla f(x) - \nabla f(y)\|^2 + \frac{\beta\sigma}{\sigma + \beta}\|x - y\|^2.$$

Lemma 1.16 (Polyak-Łojasiewicz inequality)
Let f be σ-strongly convex and $C^1(\mathbb{R}^n)$. Then, $\forall x \in \mathbb{R}^n$,

$$f(x) - f^* \leq \frac{1}{2\sigma}\|\nabla f(x)\|^2.$$

Remark The sub-optimality gap is controlled by the gradient.
(Checkpoint) Note that the optimality gap is zero when the gradient is zero, verifying Fermat's principle.

2 Notions of Nonexpansiveness

Definition 2.1 (Contraction mapping)
Let $T : X \to X$. We say T is a contraction if $\exists \kappa \in [0, 1[$ such that

$$(\forall x, y \in X), \quad \|Tx - Ty\| \leq \kappa \|x - y\|.$$

In other words, the operator norm $\|T\| = \sup_{\|x\|=1} \|Tx\| < 1$.

Definition 2.2 (Fixed points)
Let $T : X \to X$. We say that $x^* \in X$ is a fixed point of T is $Tx^* = x^*$. Define

$$\text{Fix}(T) = \{x \in X : Tx = x\}.$$

Theorem 2.3 (Banach Fixed-point Theorem)
Let $T : X \to X$ be a contraction with $\kappa = \|T\| < 1$. Then, T has a unique fixed point $x^* \in \text{Fix}(T)$ such that $x^* = Tx^*$.

Corollary 2.4 (Banach-Picard Theorem)
Let $T : X \to X$ be a contraction with $\kappa = \|T\| < 1$, and define $(x_k)_{k \in \mathbb{N}}$ as $x_{k+1} = Tx_k$. Then,

$$\|x_k - x^*\| \leq \kappa^k \|x_0 - x^*\|,$$

where x^* is a fixed point of T. Further, $\|x_k - x^*\| \leq \frac{\kappa^k}{1-\kappa} \|x_1 - x_0\|$.

Definition 2.5 (Nonexpansive and firmly nonexpansive operators)
Let $T : X \to X$.

1. T is nonexpansive if $\forall x, y \in X$,
$$\|Tx - Ty\| \leq \|x - y\|,$$
2. T is firmly nonexpansive if $\forall x, y \in X$,
$$\|Tx - Ty\|^2 + \|(I - T)x - (I - T)y\|^2 \leq \|x - y\|^2.$$

Definition 2.6 (Fejer monotone sequence)
A sequence (x_k) is said to be Fejer monotone with respect to $\Theta \subseteq X$ if

$$(\forall k \geq 1, \theta \in \Theta), \quad \|x_{k+1} - \theta\| \leq \|x_k - \theta\|.$$

Remark: A Fejer monotone sequence is necessarily bounded but need not be convergent.

Exercise (Averaging operator)**.** Show that

1. T is firmly nonexpansive if and only if $2T - I$ is nonexpansive.
   Remark: If T is firmly nonexpansive, then $T = \frac{1}{2}(I + G)$ can be written as an averaged operator, where $G = 2T - I$.

2. If T is a contraction, then T is firmly nonexpansive.

Exercise. Let T be firmly nonexpansive. Show that,

1. T is continuous.
2. Let x^* be a fixed point of T, and $x_{k+1} = T\{x_k\}$, $k > 0$. Then,
   (a) $\|x_{k+1} - x^*\| \leq \|x_k - x^*\|$, i.e., $(x_k)_{k\geq 0}$ is Fejer monotone with respect to X^*,
   (b) $\sum_{k=0}^{\infty} \|x_{k+1} - x_k\|^2 < \infty$.

3 The Gradient-Descent Algorithm

Proposition 3.1 (Gradient is an ascent direction)
Let $f \in C^1(\mathbb{R}^n)$. Then,

$$\forall x \in \mathbb{R}^n,\ \exists \delta > 0, \text{ such that } \forall t \in [0, \delta] : f(x - t\nabla f(x)) < f(x),$$

i.e., $-\nabla f(x)$ is a descent direction at x.

Proof. Let $f \in C^1(\mathbb{R}^n)$. Using Taylor's theorem, we have

$$\begin{aligned} f(x - t\nabla f(x)) &= f(x) + \langle \nabla f(x), -t\nabla f(x) \rangle + o(|t| \|\nabla f(x)\|), \\ &= f(x) - t\|\nabla f(x)\|^2 + o(t). \end{aligned}$$

It is possible to make $t > 0$ sufficiently small such that the second term is negative. More concretely, $\exists \delta > 0$ such that for $t \in [0, \delta]$,

$$\frac{o(t)}{t} < \|\nabla f(x)\|^2.$$

Then, we have $f(x - t\nabla f(x)) < f(x)$. □

This is the basis for the **gradient descent algorithm**. The parameter t is called the step-size, which is a small positive number, and may vary at every iteration as $\delta = \delta(x)$.

Algorithm 1: Gradient descent for unconstrained minimisation of differentiable functions

Input: Initialisation $x_0 \in \mathbb{R}^n$, gradient of the objective function ∇f,
Output: Minimiser $x^* = \arg\min_{x \in \mathbb{R}^n} f(x)$

for $k = 1, 2, \dots,$ *until convergence* **do**
  Choose a suitable step-size $t_k > 0$;
  $x_{k+1} = x_k - t_k \nabla f(x_k)$
return x_{k+1}

Example (Convergence of gradient descent for quadratic objective)
Consider $f(x) = \frac{1}{2}x^2$ the gradient descent iterates

$$\begin{aligned} x_{k+1} &= x_k - \alpha \nabla f(x_k), \\ &= (1 - \alpha) x_k. \end{aligned}$$

For convergence, $\alpha \in]0, 2[$, i.e., for convergence of gradient descent, the step-size must be sufficiently small.

Q: Can we ensure gradient descent is a descent method for a fixed step-size?
A: Yes, if the objective function f is "smooth".

Proposition 3.2 (Descent step for β-smooth functions)
Let f be β-smooth. Then, $\forall x \in \mathbb{R}^n$, $\forall 0 < t < \frac{1}{\beta}$,

$$f(x - t\nabla f(x)) \leq f(x) - \frac{t}{2}(2 - \beta t)\|\nabla f(x)\|^2 < f(x).$$

Proof. Using the quadratic majoriser Lemma 1.11, we have,

$$f(x - t\nabla f(x)) \leq f(x) + \langle \nabla f(x), x - t\nabla f(x) - x \rangle + \frac{\beta}{2}\|x - t\nabla f(x) - x\|^2,$$
$$= f(x) - t\|\nabla f(x)\|^2 + \frac{\beta t^2}{2}\|\nabla f(x)\|^2 = f(x) - \frac{t}{2}(2 - \beta t)\|f(x)\|^2.$$

Suppose the step-size $0 < t < \frac{1}{\beta}$, we get

$$f(x - t\nabla f(x)) \leq f(x) - \frac{t}{2}(2 - \beta t)\|\nabla f(x)\|^2 < f(x).$$

□

Corollary 3.3
Let f be β-smooth and bounded below. Consider the gradient descent iteration with any $x_0 \in \mathbb{R}^n$, $x_{k+1} = x_k - \frac{1}{\beta}\nabla f(x_k)$. Then,

$$\forall x_0 \in \mathbb{R}^n : \lim_{k\to\infty} \nabla f(x_k) = 0.$$

In particular, if x^* is a limit point of (x_k), then x^* is a stationary point, i.e., $\nabla f(x^*) = 0$.

Proof. We note that, since f is bounded below, $f^* > -\infty$ exists. Using the quadratic majoriser, we have

$$f(x_{k+1}) = f\left(x_k - \frac{1}{\beta}\nabla f(x_k)\right) \leq f(x_k) - \frac{1}{2\beta}\|\nabla f(x_k)\|^2,$$

By rearrangement, we have $\|\nabla f(x_k)\|^2 \leq 2\beta\left(f(x_k) - f(x_{k+1})\right)$. In particular, for some $N \in \mathbb{N}$, we have

$$\sum_{k=0}^{N-1}\|\nabla f(x_k)\|^2 \leq \sum_{k=0}^{N-1} 2\beta\left(f(x_k) - f(x_{k+1})\right) \leq 2\beta\left(f(x_0) - f(x_N)\right)$$
$$\leq 2\beta\left(f(x_0) - f^*\right).$$

Since the partial sum is bounded above, using the limit test for convergent series, we have $\lim_{k\to\infty} \nabla f(x_k) = 0$. □

Terminology

1. Objective convergence: $f(x_k) \to f^*$
2. Iterate convergence: $x_k \to x^*$
3. Minimising sequence: $(x_k)_{k\geq 0}$ is a minimising sequence if $x_k \to x^*$

Iterate convergence is stronger than objective convergence If f is β-smooth, then linear iterate convergence implies linear objective convergence, as

$$f(x_k) - f^* \leq \langle \nabla f(x^*), x_k - x^* \rangle + \frac{\beta}{2}\|x_k - x^*\|^2 = \frac{\beta}{2}\|x_k - x^*\|^2.$$

3.1 Convergence Analysis of Gradient Descent

Theorem 3.4 (Linear objective convergence of gradient descent for smooth and strongly convex functions)
Let f be β-smooth and σ-strongly convex. Fix $x_0 \in \mathbb{R}^n$, and let $x_{k+1} = x_k - \frac{1}{\beta}\nabla f(x_k)$, $k > 0$. Then, we have objective convergence with an exponential rate (also called linear convergence). More precisely,

$$f(x_k) - f^* \leq (f(x_0) - f^*)\left(1 - \frac{\sigma}{\beta}\right)^k.$$

Proof. Using Proposition 3.2, we have $f(x_{k+1}) \leq f(x_k) - \frac{1}{2\beta}\|\nabla f(x)\|^2$. Then, using Lemma 1.16, we have:

$$\begin{aligned} f(x_{k+1}) - f^* &\leq f(x_k) - f^* - \frac{1}{2\beta}\|\nabla f(x)\|^2, \\ &\leq f(x_k) - f^* - \frac{1}{2\beta} \cdot 2\sigma(f(x_k) - f^*) = \left(1 - \frac{\sigma}{\beta}\right)(f(x_k) - f^*), \\ &\leq \cdots, \\ &\leq \left(1 - \frac{\sigma}{\beta}\right)^{k+1}(f(x_0) - f^*). \end{aligned}$$

□

Theorem 3.5 (Linear iterate convergence of gradient descent for smooth and strongly convex functions)
Let f be β-smooth and σ-strongly convex. Fix $x_0 \in \mathbb{R}^n$, and let $x_{k+1} = x_k - \frac{1}{\beta}\nabla f(x_k)$ $k > 0$. Then, we have iterate convergence with an exponential rate. More precisely,

$$\|x_k - x^*\|^2 \leq \left(\frac{\beta - \sigma}{\beta + \sigma}\right)^k \|x_0 - x^*\|^2.$$

Proof.

□

Theorem 3.6 (Convergence of gradient descent for smooth and convex functions)
Let f be convex and β-smooth. Fix $x_0 \in \mathbb{R}^n$, and let $x_{k+1} = x_k - \frac{1}{\beta}\nabla f(x_k)$. Then,

$$f(x_k) \leq f^* + o\left(\frac{1}{k}\right) \text{ and } \|x_k - x^*\| \to 0.$$

3.2 Convergence Analysis of Gradient Descent Using Contraction Mapping Theory

Proposition 3.7
The sequence defined by $x_{k+1} = x_k - \frac{1}{\beta}\nabla f(x_k)$, for some $x_0 \in \mathbb{R}^n$, is Fejer monotone with respect to $X^* = \{x^* \in X : f(x^*) = \inf_{x \in X} f(x)\}$.

Theorem 3.8

$$(x_k) \to x^*$$

Define $T^f_{\text{grad}} : X \to X$ given by $T^f_{\text{grad}} = I - \dfrac{1}{\beta}\nabla f$. Each iteration of gradient descent is an application of this operator, i.e., it is a fixed-point iteration of T^f_{grad} — $x_{k+1} = T^f_{\text{grad}}\{x_k\}$.

Lemma 3.9
Let f be β-smooth and convex, and define $T^f_{\text{grad}} : X \to X$ given by $T^f_{\text{grad}} = I - \dfrac{1}{\beta}\nabla f$. Then,

1. T^f_{grad} is firmly nonexpansive,
2. x^* is a minimiser of $f \Leftrightarrow x^*$ is a fixed point of T^f_{grad}.

Proof.

□

Exercise. Let f be σ-strongly convex. Then, show that T^f_{grad} is a contraction.

Theorem 3.10
Let T be a firmly nonexpansive operator and let fix $T \neq \emptyset$. Then, for any $x_0 \in X$, the sequence $x_{k+1} = T\{x_k\}$ converges to a point in fix T.

4 Constrained Optimisation and Projected Gradient Descent

Theorem 4.1
Let $C \subseteq X$ be nonempty and convex. Let $f : X \to \mathbb{R}$ be convex and differentiable. Then,

$$x^* = \arg\min_{x \in C} f(x) \Leftrightarrow \forall x \in C : \langle x - x^*, \nabla f(x^*) \rangle \geq 0.$$

Proof.

□

Definition 4.2 (Projection operator)
Let $C \subseteq X$ be nonempty. The projection operator on C is defined as

$$\Pi_C : X \to X,\ \Pi_C(x) = \arg\min_{z \in C} \|z - x\|.$$

Remark: Suppose $C = (0, 1)$ and $x = 0$, $\Pi_C(x)$ is not defined. Closure of the set C is necessary for the projection operator to be well-defined for all points in X.

Theorem 4.3
Let C be nonempty, closed and convex. Then, $\Pi_C(x)$ is well-defined $\forall x \in X$, i.e., $\Pi_C(x)$ exists and is unique. In particular, the function $z \mapsto \|z - x\|$ has a unique minimiser over C.

Proof.

□

Exercise (Computing the projection operator)**.** Let $X \in \mathbb{R}^n$. Compute Π_C when C is the

1. closed unit ℓ_2-ball,
2. closed unit ℓ_∞-ball,
3. $\mathbb{R}^n_+$,
4. range of $A \in \mathbb{R}^{n \times n}$.

Lemma 4.4
Let C be nonempty, closed and convex. Let $\hat{x}, x \in X$. Then,

$$\hat{x} = \Pi_C(x) \Leftrightarrow \forall z \in C,\ \langle z - \hat{x}, x - \hat{x} \rangle \leq 0.$$

Geometric interpretation:

Algorithm 2: Projected gradient descent for constrained minimisation of differentiable functions

Input: Initialisation $x^{(0)} \in \mathbb{R}^n$, gradient of the objective function ∇f, projection operator Π_C onto the constrain set C

Output: Minimiser $x^* = \arg\min_{x \in C} f(x)$

for $k = 1, 2, \ldots,$ *until convergence* **do**
  Choose a suitable step-size $t_k > 0$;
  $z^{(k+1)} = x^{(k)} - t_k \nabla f(x^{(k)})$;
  $x^{(k+1)} = \Pi_C(z^{(k+1)})$
return $x^{(k+1)}$

4.1 Convergence Analysis of Projected Gradient Descent

Theorem 4.5
Let C be a nonempty, closed and convex; and let $f : X \to \mathbb{R}$ be β-smooth and convex. Then, for any $x^{(0)} \in X$, define

$$x^{(k+1)} = \Pi_C \left(x^{(k)} - \frac{1}{\beta} \nabla f(x^{(k)}) \right);\ k \geq 0,$$

Then, $(x^{(k)})$ converges to the minimiser x^* of f over C; and $(f(x^{(k)}))$ converges to f^*.

Proposition 4.6
Let C be nonempty, closed and convex. Then, the operator Π_C is (firmly) nonexpansive.

Define $T^f_{\text{proj}} : X \to X$ given by $T^f_{\text{proj}} = \Pi_C \circ \left(I - \frac{1}{\beta} \nabla f \right)$. Each iteration of projected gradient descent is an application of this operator, i.e., it is a fixed-point iteration of T^f_{proj} — $x^{(k+1)} = T^f_{\text{proj}}\{x^{(k)}\}$.

Lemma 4.7
Let f be β-smooth and convex, and define $T^f_{\text{proj}} : X \to X$ given by

$$T^f_{\text{proj}} = \Pi_C \circ \left(I - \frac{1}{\beta} \nabla f \right).$$

Then,

1. $x^* \in \arg\min_{x \in X} f(x) \iff x^* \in \operatorname{Fix}\left(T^f_{\text{proj}} \right)$,
2. T^f_{proj} is firmly nonexpansive.

4.2 Extended-Valued Functions

Definition 4.8 (Indicator function)
Let $C \subseteq X$ be nonempty. The (0-∞) indicator function of C, $\iota_C : X \to \mathbb{R} \bigcup \{-\infty, +\infty\} \stackrel{\text{def.}}{=} \bar{\mathbb{R}}$ is defined as

$$\iota_C(x) = \begin{cases} 0, & x \in C, \\ +\infty, & \text{otherwise.} \end{cases}$$

Remark: Constrained optimisation programmes can now be posed as an unconstrained optimisation programmes as

$$\inf_C f = \inf_X (f + \iota_C).$$

Definition 4.9 (Extended reals)
The set of extended reals $\bar{\mathbb{R}} \triangleq \mathbb{R} \bigcup \{-\infty, +\infty\}$ along with operations

1. $\forall a \in \mathbb{R},\ a + (\pm\infty) = \pm\infty,$
2. $\forall a > 0,\ a \cdot (\pm\infty) = \pm\infty,$
3. $\forall a < 0,\ a \cdot (\pm\infty) = \mp\infty,$
4. $0 \cdot (\pm\infty) = 0,$
5. $\forall a \in \mathbb{R} \bigcup \{-\infty\},\ a < +\infty,$
6. $\forall a \in \mathbb{R} \bigcup \{+\infty\},\ a > -\infty.$

Definition 4.10 (Extended real-valued functions, effective domain, epigraph and proper function)
An extended real-valued ($\bar{\mathbb{R}}$-valued) function on X is a function $f : X \to \bar{\mathbb{R}}$. The (effective) domain of f,

$$\operatorname{dom}\{f\} = \{x \in X : f(x) < +\infty\}.$$

The epigraph of f,

$$\operatorname{epi}\{f\} = \{(x, t) \in X \times \mathbb{R} : f(x) \leq t\}.$$

f is proper if,

$$\forall x \in X, f(x) \neq -\infty, \text{ or } f(x) > -\infty,$$
$$\exists x \in X, f(x) < +\infty, \text{ or } f(x) \in \mathbb{R}.$$

Definition 4.11 (Extended-valued convex functions)
A proper function $f : X \to \bar{\mathbb{R}}$ is convex if

$$\forall x_1, x_2 \in X, \forall \theta \in [0, 1],\ f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta) f(x_2).$$

Proposition 4.12
Let $f : X \to \bar{\mathbb{R}}$ be proper. Then, the following are equivalent.

1. f is convex,
2. $\operatorname{dom}\{f\}$ is convex, and $f|_{\operatorname{dom}\{f\}}$ is a real-valued convex function,
3. $\operatorname{epi}\{f\}$ is convex.

Exercise. Let $f : X \to \mathbb{R}$ be convex (and continuous). Then, show that $\operatorname{epi}\{f\}$ is closed in $X \times \mathbb{R}$.

Example
The epigraph of $f = \iota_{[0,1]}$ is not closed. What is the problem here?

Definition 4.13 (Closed functions)
A function $f : X \to \bar{\mathbb{R}}$ is closed of $\operatorname{epi}\{f\}$ is nonempty and closed.

Exercise. ι_C is closed $\Leftrightarrow$ C is closed in X.

Exercise. Show that if f, g is proper and closed, then, $\forall \alpha, \beta \in \mathbb{R} : \alpha f + \beta g$ is proper and closed.

4.3 Lower Semicontinuity

Definition 4.14 (Lower semicontinuous functions)
A function $f : X \to \bar{\mathbb{R}}$ is lower semicontinuous at $x \in X$ if

$$\forall (x_n) \subset X, (x_n) \to x : f(x) \leq \liminf f(x_n).$$

f is lower semicontinuous if f is lower semicontinuous $\forall x \in X$.

Exercise. Show that f is lower semicontinuous at $x \in X$ if and only if

$$\forall \epsilon > 0, \exists \delta > 0, \forall z \in \mathcal{B}(x, \delta) : f(z) > f(x) - \epsilon.$$

Exercise. Show that if (x_n) and λ are such that $\liminf f(x_n) < \lambda$, then $f(x_n) < \lambda$ infinitely often, i.e., $\exists (x_{n_k})_{k \geq 1}, f(x_{n_k}) < \lambda$, $\forall k \geq 1$.

Theorem 4.15
Let $f : X \to \bar{\mathbb{R}}$. Then, the following are equivalent.

1. f is lower semicontinuous.
2. f is closed, i.e., epi$\{f\}$ is closed in $X \times \mathbb{R}$.
3. $\forall \lambda \in \mathbb{R}$, $\{x : f(x) \leq \lambda\} = \text{lev}_\lambda\{f\}$ is closed.

Proof. $(1 \Rightarrow 2)$ Let $(x_n, t_n) \in$ epi$\{f\}$ such that $(x_n, t_n) \to (x, t)$. We need to show $f(x) \leq t$. However, we have $f(x_n) \leq t_n$. Therefore, $f(x) \leq \liminf f(x_n) \leq \liminf t_n = t$.

$(1 \Rightarrow 3)$ Let (x_n) be in $\{x : f(x) \leq \lambda\}$, i.e., $f(x_n) \leq \lambda$ such that $(x_n) \to x$. Show $f(x) \leq \lambda$. We have $(x_n, \lambda) \in$ epi$\{f\}$ and $(x_n, \lambda) \to (x, \lambda)$. Then, $f(x) \leq \liminf f(x_n) \leq \lambda$.

$(2 \Rightarrow 3)$

$(3 \Rightarrow 1)$ Suppose f is not lower semicontinuous at $x_0 \in X$, i.e., $\forall (x_n) \to x_0$ such that $f(x_0) > \liminf f(x_n)$. This means that $\exists \lambda \in \mathbb{R}$ such that $\liminf f(x_n) < \lambda < f(x_0)$. We need to show that $\{x : f(x) \leq \lambda\}$ is not closed. Since $\liminf f(x_n) < \lambda$, then $\exists (x_{n_k})_{k \geq 1}, f(x_{n_k}) < \lambda$, $\forall k \geq 1$, i.e., $x_{n_k} \in \{x : f(x) \leq \lambda\}$. However, $x_{n_k} \to x_0$ and $f(x_0) > \lambda$. So, $\{x : f(x) \leq \lambda\}$ is not closed. $\square$

Proposition 4.16
Let $f : X \to \bar{\mathbb{R}}$ be proper. Assume that dom$\{f\}$ is closed, and f is closed over dom$\{f\}$. Then, f is closed. In particular, if $f|_{\text{dom}\{f\}}$ is continuous, then f is closed.

Theorem 4.17 (Weierstrass Theorem)
Let $f : X \to \mathbb{R}$ be continuous, and let $C \subseteq X$ be compact. Then, f is bounded below on C, and $\exists x^* \in C$ such that $f(x^*) = \inf_C f$.

Proof. See text. $\square$

Example (Motivation)
Consider the function

$$f(x) = \begin{cases} \sqrt{x}, & x > 0, \\ 1, & x \leq 0. \end{cases}$$

The minimisation problem with f as the objective is not solvable.
Note: f is not lower semicontinuous at $x = 0$.

Theorem 4.18
Let $f : X \to \bar{\mathbb{R}}$ be closed, and let $C \subseteq X$ be compact. Then, $\exists x^* \in C$ such that $f(x^*) = \inf_C f$.

Proof. The case when $\inf_C f = \pm\infty$ is easy. Use compactness and lower semicontinuity. Then, consider the case when $C \bigcap \text{dom}\{f\} \neq \emptyset$. Note that $\inf\limits_C f = \inf\limits_{C \bigcap \text{dom}\{f\}} f$, and check, $\text{dom}\{f\}$ is closed. Therefore, $C \bigcap \text{dom}\{f\}$ is compact. Then, using Theorem 4.17, $\exists x^* \in C$ such that $f(x^*) = \inf\limits_{C \bigcap \text{dom}\{f\}} f = \inf\limits_C f$. $\square$

Definition 4.19 (Coervice functions)
We say $f : X \to \bar{\mathbb{R}}$ is coercive if

$$\lim_{\|x\| \to +\infty} f(x) = +\infty.$$

Exercise. Prove that any differentiable and strongly-convex function is coercive.

Example
Let $A \in \mathbb{R}^{n \times n}$ be symmetric and positive definite. Show that $f(x) = \frac{1}{2} x^\mathsf{T} A x$ is coercive. Show that this is NOT true if A is positive semi-definite and not positive definite.

Corollary 4.20
Let $f : X \to \bar{\mathbb{R}}$ be coercive and closed, and let $C \subseteq X$ be closed. Then, $\exists x^* \in C$ such that $f(x^*) = \inf_C f$.

Proof. Exercise. $\square$

Definition 4.21 (Extended-valued strictly convex function)
A proper function $f : X \to \bar{\mathbb{R}}$ is strictly convex if

$$\forall x_1, x_2 \in X, \forall \theta \in [0, 1],\ f(\theta x_1 + (1 - \theta) x_2) < \theta f(x_1) + (1 - \theta) f(x_2).$$

Example
Prove that any strongly convex function is strictly convex, but the converse is not true.

Exercise. Let $f : X \to \bar{\mathbb{R}}$ be proper and strictly convex. Then, for any $C \subseteq X$, f can have at most one (unique) minimiser over C.

5 Subgradient Methods

Definition 5.1 (Class of closed, proper and convex functions)

$$\Gamma_0(X) := \left\{ f : X \to \bar{\mathbb{R}} \;\middle|\; f \text{ is closed, proper and convex} \right\}.$$

__Recall:__ For convex, differentiable functions, we have the tangent property Theorem 1.7.

Definition 5.2 (Subdifferential of a function)
Let $f : X \to \bar{\mathbb{R}}$, and $x \in \text{dom}\{f\}$. Then, the subdifferential of f at x is the set

$$\partial f(x) = \{\zeta \in X : f(y) \geq f(x) + \langle \zeta, y - x\rangle, \forall y \in X\}.$$

The elements of $\partial f(x)$ are called subgradients, and, the (set-valued) operator $\partial f : X \to 2^X$ is called the subdifferential operator.

Remark
If f is differentiable on X, then $\nabla f(x) \in \partial f(x)$, and further, $\partial f(x) = \{\nabla f(x)\}$.

Lemma 5.3
Let $f : X \to]-\infty, +\infty]$ be proper, and $x \in \text{dom}\{f\}$. Then,

1. $\text{dom}\{\partial f\} \subseteq \text{dom}\{f\}$
2. $\partial f(x) = \cup_{y \in \text{dom}\{f\}} \{\zeta \in X : \langle y - x, \zeta\rangle \leq f(y) - f(x)\}$
3. $\partial f(x)$ is closed and convex

Remark
If $f \in \Gamma_0(X)$ then $\partial f(x) \neq \emptyset, \forall x \in X$ and $\text{dom}\{\partial f\} = \{x \in X : \partial f(x) \neq \emptyset\} = X$.

Theorem 5.4 (Existence of subgradients for convex functions)
Let $f : X \to \mathbb{R}$ be convex. Then,

$$\forall x \in X, \exists \zeta \in X : \forall y \in X,\ f(y) \geq f(x) + \langle \zeta, y - x\rangle$$

Remark
Nonconvex functions can have empty subdifferentials. Give an example.

Exercise. Consider $f(x) = |x|$. Show that

$$\partial f(x) = \begin{cases} \dfrac{x}{|x|}, & x \neq 0, \\ [-1, +1], & x = 0. \end{cases}$$

Then, by separable extension,

$$\partial \|x\|_1 = \begin{cases} \frac{x}{\|x\|_1}, & x \neq 0, \\ [-1, +1]^n, & x = 0. \end{cases}$$

Theorem 5.5 (Separating hyperplane theorem)
Exercise.

Theorem 5.6 (Supporting hyperplane theorem)
Let $C \subseteq X$ be closed, convex set, and let $x_0 \in \text{bdd}\{C\} = C \setminus \text{int}\{C\}$. Then,

$$\forall \zeta \in X \setminus \{0\}, \forall x \in C : \langle \zeta, x - x_0 \rangle \leq 0.$$

Proof. Since $x_0 \in \text{bdd}\{C\}$, there exists $(z_n) \subseteq C$ such that $z_n \to x_0$. Let $\xi_n = z_n - \Pi_C(z_n) \neq 0$. We have, $\langle x - \Pi_C(z_n), z_n - \Pi_C(z_n) \rangle \leq 0$. Using a renormalisation trick, i.e., using $\hat{\xi}_n \leftarrow \xi_n / \|\xi_n\|$, we have $\langle x - \Pi_C(z_n), \hat{\xi}_n \rangle \leq 0$. Setting $\zeta = \lim_{k \to \infty} \hat{\xi}_{n_k}$, and using continuity of inner product functionals, we have $\langle x - x_0, \zeta \rangle \leq 0$. We also have $\|\zeta\| = 1$. □

Proof of Theorem 5.4. Let $f : X \to \mathbb{R}$ be convex. We need to show $\partial f(x) \neq \emptyset, \forall x \in X$, i.e., $\text{dom}\{\partial f\} = X$. Let $C = \text{epi}\{f\} \in X \times \mathbb{R}$ and $x = x_0$. Since f is convex (and therefore continuous), C is closed and convex. Further, $(x, f(x)) \in \text{bdd}\{C\}$. Then, using Theorem 5.6, $\exists 0 \neq (\xi, \alpha) \in X \times \mathbb{R}, \forall (y, t) \in \text{epi}\{f\}$

$$\begin{aligned}
\langle (\xi, \alpha), (y, t) \rangle &\leq \langle (\xi, \alpha), (x, f(x)) \rangle, \\
\langle \xi, y \rangle + \alpha t &\leq \langle \xi, x \rangle + \alpha f(x).
\end{aligned}$$

Claim $\alpha \neq 0$: Suppose $\alpha = 0$, $\forall y \in X$, $\langle \xi, y - x \rangle \leq 0 \implies \xi = 0$, which is a contradiction.
Claim $\alpha < 0$: Suppose $\alpha > 0$, then, $\forall t \geq f(y), t \leq \gamma$, which is a contradiction.
Dividing by α and setting $\zeta = -\frac{1}{\alpha}\xi$, we have

$$\langle -\zeta, y \rangle + t \geq \langle -\zeta, x \rangle + f(x),$$

and in particular, $f(y) \geq \langle \zeta, y - x \rangle + f(x)$. □

Exercise. Let $(X, \|\cdot\|)$ be a normed space, and $f : X \to \mathbb{R}$, defined as $f(x) = \|x\|$. Show that $\partial f(0) = \{\zeta \in X : \|\zeta\|_* \leq 1\}$, where $\|x\|_* = \sup_{\|\xi\|=1} |\langle \xi, x \rangle|$.

Example
Construct a convex function $f : X \to \bar{\mathbb{R}}$ such that $\partial f(x) = \emptyset$ for some $x \in X$.

$$f(x) = \begin{cases} -\sqrt{x}, & x \geq 0, \\ +\infty, & x < 0. \end{cases}$$

(Claim) $\partial f(0) = \emptyset$. Supoose $\xi \in \partial f(0)$. Then,

$$\begin{aligned}
\forall x > 0 : -\sqrt{x} &\geq 0 + \langle \xi, x - 0 \rangle, \\
-\sqrt{x} &\geq \xi x, \\
\implies \sqrt{x} &\leq -1/\xi,
\end{aligned}$$

which is a contradiction.

Exercise. Let $f : X \to \bar{\mathbb{R}}$ be proper. Suppose, $\text{dom}\{f\}$ is convex and $\partial f(x) \neq \emptyset, \forall x \in \text{dom}\{f\}$. Then, show that f is convex.

Exercise. Let C be a closed, convex set, and let $x_0 \in C$. Show that

$$\partial \iota_C(x_0) = \mathrm{N}_C(x_0) \triangleq \{\zeta \in X : \langle \zeta, x - x_0 \rangle \leq 0\},$$

(the normal cone of C at x_0), $\forall x_0 \in C$.

Theorem 5.7 (Sub-differential inclusion)
Let $f : X \to \bar{\mathbb{R}}$ be proper. Then,

$$x^* = \arg\min_{x \in X} f(x) \iff 0 \in \partial f(x^*).$$

Proof. We have $0 \in \partial f(x^*)$ if and only if

$$\begin{aligned} \forall x \in X : f(x) &\geq f(x^*) + \langle 0, x - x^* \rangle, \\ f(x) &\geq f(x^*), \end{aligned}$$

$\iff x^* = \arg\min_{x \in X} f(x)$. □

Theorem 5.8
Let $f \in \Gamma_0(X)$. Then, $\text{int}(\text{dom}\{f\}) \subseteq \text{dom}\{\partial f\}$.

Proof. Exercise. □

Lemma 5.9
Let $f : X \to \bar{\mathbb{R}}$ be proper and convex. Then, f is locally Lipschitz on $\text{int}(\text{dom}\{f\})$, i.e.,

$$\exists \delta_0 > 0, L_x > 0, \forall u, v \in \mathcal{B}(x, \delta_0) : |f(u) - f(v)| \leq L_x \|u - v\|.$$

Theorem 5.10
Let $f : X \to \bar{\mathbb{R}}$ be proper and convex. Then, $\partial f(x)$ is closed and convex if $x \in \text{dom}\{f\}$. Moreover, if $x \in \text{int}(\text{dom}\{f\})$, then $\partial f(x)$ is bounded.

Proof. We have

$$\partial f(x) = \{\zeta \in X : \underbrace{f(y) - f(x)}_{\alpha_y} \geq \langle \zeta, \underbrace{y - x}_{d_y} \rangle\} = \bigcap_{y \in X} H_y,$$

where $H_y = \{\zeta \in X : \langle \zeta, d_y \rangle \leq \alpha_y\}$, which is closed and convex, therefore, $\partial f(x)$ is closed and convex. alternatively show using first principles.

Using Lemma 5.9, we have

$$\exists \delta > 0, L_x > 0, \forall z \in \mathcal{B}(x, \delta) : |f(z) - f(x)| \leq L_x \|z - x\|.$$

and from the definition of the subgradient,

$$\forall z \in X : f(z) \geq f(x) + \langle \zeta, z - x \rangle.$$

Choose $z = x + \frac{\delta}{2} \frac{\zeta}{\|\zeta\|}$ (w.l.o.g $\zeta \neq 0$) to show $\|\zeta\| \leq L_x$. □

Exercise. Show that any proper, convex function $f : X \to \bar{\mathbb{R}}$ is locally bounded on $\text{int}(\text{dom}\{f\})$, i.e.,

$$\forall x_0 \in \text{int}(\text{dom}\{f\}), \exists \delta > 0 : 0 \leq |f(x)| \leq M, \forall x \in \mathcal{B}(x_0, \delta).$$

Find such an M and a δ.

Proof of Lemma 5.9. Let $\delta_0 = \frac{1}{2}\delta$. **Claim:** $\forall u, v \in \mathcal{B}(x, \delta_0) : f(u) - f(v) \leq L_x \|u - v\|$. Then, using convexity, we have the desired (with the $|\cdot|$) result.

Let $w = v + \frac{\delta}{2}\frac{v-u}{\|v-u\|} \triangleq v + \gamma(v-u)$. **Claim:** $\|w - x\| \le \delta$ (Show).
Then, $v = \frac{\gamma}{1+\gamma}u + \frac{1}{1+\gamma}w$. By convexity,

$$f(v) \le \frac{\gamma}{1+\gamma}f(u) + \frac{1}{1+\gamma}f(w),$$
$$f(v) - f(u) \le \frac{1}{1+\gamma}(f(u) - f(w)) \le \frac{2M}{\delta} = \frac{4M}{\delta}\|u - v\|.$$

see the videos on YouTube by the Vietnamese dude. □

Corollary 5.11
Any real-valued convex function is continuous.

Proof.

□

Theorem 5.12
Let $f : X \to \bar{\mathbb{R}}$ be proper and $C \subseteq X$ be nonempty, closed and convex. Then,

$$x^* = \arg\min_C f \Leftrightarrow \exists \xi \in \partial f(x^*) : \langle \xi, x - x^* \rangle \ge 0, \forall x \in C.$$

5.1 Directional Derivatives and Support Functions

Theorem 5.13 (Directional derivative)
Let $f \in \Gamma_0(X)$ and let $x_0 \in \mathrm{int}(\mathrm{dom}\{f\})$. Then, for any $d \in X$, the following exists:

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

and is called the directional derivative of f at x along d, and is finite.

Proof.

$$\lim_{t \downarrow 0} \frac{f(x + td) - f(x)}{t} = \inf_{t > 0} \frac{f(x + td) - f(x)}{t}$$

Exercise. $\forall t_1 < 0 < t_2 < t_3, \theta(t_1) \le \theta(t_2) \le \theta(t_3)$ θ is what the limit is taken of

□

Proposition 5.14
Let $f \in \Gamma_0(X)$ and $x \in \mathrm{int}(\mathrm{dom}\{f\})$. Define $\psi : X \to \mathbb{R}$ as $\psi(d) = f'(x; d)$. Then,

1. ψ is sub-additive: $\psi(d_1 + d_2) \le \psi(d_1) + \psi(d_2)$
2. ψ is positive homogeneous: $\lambda > 0$, $\psi(\lambda d) = \lambda \psi(d)$

In particular, ψ is convex.

Proof.

□

Exercise. Compute the direction derivative of $f(x) = |x|$.

Lemma 5.15
Let $f \in \Gamma_0(X)$ and $x \in \text{int}(\text{dom}\{f\})$. Then,

$$\partial f(x) = \{\zeta \in X : \langle \zeta, d \rangle \leq f'(x; d), \forall d \in X\}.$$

Moreover, $f'(x; d) = \max_{\zeta \in \partial f(x)} \langle \zeta, d \rangle$. With this it is easy to show $\psi(d)$ is convex.

Proof.

□

Exercise. Let $A \in \mathbb{R}^{m \times n}$, $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$. Consider the linear programme:

$$p = \inf_{x \in \mathbb{R}^n} \left\{ c^\top x : Ax = b, x \geq 0 \right\}, \; d = \sup_{y \in \mathbb{R}^m} \left\{ b^\top y : c - A^\top y \geq 0 \right\}.$$

Suppose both are feasible. Then, show, using separating hyperplane theorem, that p, d exist and $p = d$ (strong duality).

Solution: Consider $K = \{(u, t) \in \mathbb{R}^m \times \mathbb{R} : \exists x \geq 0, s.t., Ax = u, c^\top x = t\}$. Show K is closed and use separating hyperplane theorem.

Theorem 5.16
Let $f \in \Gamma_0(X)$ and $x \in \text{int}(\text{dom}\{f\})$. Then, $\partial f(x)$ is a singleton if and only if f is differentiable at x.

Proof. ($\Leftarrow$) Suppose f is differentiable at x. Then, using Definition 1.5, we know $\nabla f(x) \in \partial f(x)$. We claim that if $\zeta \in \partial f(x)$, then $\zeta = \nabla f(x)$, which would imply $\partial f(x)$ is a singleton. We have

$$\forall h \in X : f(x + h) \geq f(x) + \langle \zeta, h \rangle.$$

Since f is differentiable,

$$\forall h \in X : f(x + h) = f(x) + \langle \nabla f(x), h \rangle + o(\|h\|).$$

Therefore, we have,

$$\begin{aligned} \forall h \in X : f(x) + \langle \nabla f(x), h \rangle + o(\|h\|) &\geq f(x) + \langle \zeta, h \rangle, \\ \langle \nabla f(x), h \rangle + o(\|h\|) &\geq \langle \zeta, h \rangle, \\ o(\|h\|) &\geq \langle \zeta - \nabla f(x), h \rangle. \end{aligned}$$

Put $h = t(\zeta - \nabla f(x)), t > 0$,

$$\frac{o(t)}{t} \geq \|\zeta - \nabla f(x)\|^2,$$

and as $t \to 0^+$, then $\zeta = \nabla f(x)$.

($\Rightarrow$) Suppose $\partial f(x)$ is a singleton, i.e., $\partial f(x) = \{\zeta_0\}$. Then, using Lemma 5.15, $f'(x; d) = \langle \zeta_0, d \rangle, \forall d \in X$, i.e.,

$$\langle \zeta_0, d \rangle = \lim_{t \to 0} \frac{f(x + td) - f(x)}{t}.$$

Taking d to be each of the orthonormal basis vectors $\{e_1, e_2, \cdots, e_n\}$, then, we have existence of all the partial derivatives. This implies f is differentiable at x. The implication requires convexity. □

Lemma 5.17
Let $f \in \Gamma_0(X)$. Consider an (ortho) basis $(e_i)_{i\in I}$ of X. Then,

$$f \text{ is differentiable at } x \in \text{int}(\text{dom}\{f\}) \Leftrightarrow \mathrm{D}f(x;e_i) \text{ exists } \forall i \in I.$$

Proof. Note that $\mathrm{D}f(x;e_i) = \lim_{t\to 0} \frac{f(x+te_i)-f(x)}{t}$. ($\Leftarrow$) Suppose $\mathrm{D}f(x;e_i)$ exists $\forall i \in I$, i.e., $\exists \zeta \in X$:

$$\forall h \in X : f(x+h) = f(x) + \langle \zeta, h\rangle + o(\|h\|).$$

Claim: $\zeta_i = \mathrm{D}f(x;e_i)$. Weaker form will suffice, i.e., it is enough to show ...complete the proof ☐

Exercise. Let $f \in \Gamma_0(X)$. Then, f is σ-strongly convex if and only if

$$\forall x, y \in X, \zeta \in \partial f(x) : f(y) \geq f(x) + \langle \zeta, y-x\rangle + \frac{\sigma}{2}\|y-x\|^2.$$

Proposition 5.18 (Sum rule)
Let $f : X \to \mathbb{R}$ be differentiable, and $g \in \Gamma_0(X)$. Then,
1. $\text{dom}\{f+g\} = \text{dom}\{g\}$
2. $\forall x \in \text{dom}\{\partial g\} : \partial(f+g)(x) = \nabla f(x) + \partial g(x)$

Note: In general, $f, g \in \Gamma_0(X)$, $\partial f + \partial g \subseteq \partial(f+g)$.

Example
Consider

$$f(x) = \begin{cases} -\sqrt{x}, & x \geq 0, \\ +\infty, & x < 0. \end{cases}, \; g(x) = \iota_{]0,\infty[}(x).$$

$\partial(f+g)(0) = \mathbb{R}$. But $\partial f(0) + \partial g(0) = \emptyset$.

Theorem 5.19
Let $f, g \in \Gamma_0(X)$ and let $x \in \text{int}(\text{dom}\{f\}) \bigcap \text{int}(\text{dom}\{g\})$. Then,

$$\partial(f+g)(x) = \partial f(x) + \partial g(x).$$

Exercise (Chain rule)**.** Let $f \in \Gamma_0(X)$, and let A be a linear map on X. Define $g : X \to \bar{\mathbb{R}}$, as $g(x) = f(Ax)$.

1. Show that g is convex and closed.
2. Assume $x, Ax \in \text{dom}\{f\}$. Then, show that $\partial g(x) = A^*\partial f(Ax)$.

Theorem 5.20
Let $f \in \Gamma_0(X)$. Then, the subdifferential operator $\partial f : X \to 2^X$ is maximal monotone. More precisely,
1. $\forall x, y \in X, \zeta \in \partial f(x), \xi \in \partial f(y) : \langle \zeta - \xi, x-y\rangle \geq 0$.
2. Suppose $z, v \in X$:
$$\forall x \in X, \zeta \in \partial f(x) : \langle z-x, v-\zeta\rangle \geq 0,$$
then, $v \in \partial f(z)$.

6 Proximal Methods

6.1 Projected Subgradient Descent

Algorithm 3: Projected subgradient descent for constrained minimisation

Input: Initialisation $x^{(0)} \in C$, projection operator Π_C onto the constrain set C
Output: Minimiser $x^* = \arg\min_{x \in C} f(x)$

for $k = 1, 2, \ldots,$ *until convergence* **do**
  Choose a suitable step-size $t_k > 0$;
  $z_{k+1} = x_k - t_k \zeta_k$, $\zeta_k \in \partial f(x_k)$;
  $x_{k+1} = \Pi_C(z_{k+1})$
return $x^{(k+1)}$

Assumptions:

1. $f : X \to \bar{\mathbb{R}}$ is proper, closed and σ-strongly convex,
2. $C \subseteq \text{int}(\text{dom}\{f\})$,
3. $f^* = \min_C f > -\infty, \arg\min_C f$ is nonempty,
4. $\exists L > 0 : \forall x \in C, \zeta \in \partial f(x), \|\zeta\| \leq L$.

Example

Theorem 6.1
Let (x_k) be the sequence generated by $x_{k+1} = \Pi_C(x_k - t_k \zeta_k)$, where $\zeta \in \partial f(x_k), t_k = \frac{2}{\sigma(k+1)}$. Then,

$$\underbrace{\min_{1 \leq k \leq n} f(x_k)}_{f_n^*} - f^* = o(1/n).$$

Moreover, if i_n such that $f(x_{i_n}) = f_n^*$, then $\|x_{i_n} - x^*\|^2 = o(1/n)$.

6.2 Proximal Gradient Method

Definition 6.2 (Proximal Operator)
Let $f \in \Gamma_0(X)$. The proximal operator of f is defined as

$$\text{prox}_f : X \to X, \ \text{prox}_f(x) = \arg\min_{\theta \in X} \frac{1}{2}\|\theta - x\|^2 + f(\theta).$$

Proposition 6.3
Let $f \in \Gamma_0(X)$. Then, for any $x \in X$, the function

$$\theta \mapsto \frac{1}{2}\|\theta - x\|^2 + f(\theta)$$

is strongly convex. In particular, prox_f is well-defined.

Proof. The proof is easy if $\mathrm{dom}\{f\} = X$. Else, we need the following observation.

Lemma 6.4 (Test 2 problem)
Let $f \in \Gamma_0(X)$. Then, $\exists \zeta \in X, \alpha \in \mathbb{R}$ such that

$$\forall x \in X : f(x) \geq \langle \zeta, x \rangle + \alpha.$$

finish this

□

Theorem 6.5
Let $f \in \Gamma_0(X)$. Then, prox_f is firmly-nonexpansive, i.e.,

$$\forall x, y \in X : \left\| \mathrm{prox}_f(x) - \mathrm{prox}_f(y) \right\|^2 + \left\| (x - \mathrm{prox}_f(x)) - (y - \mathrm{prox}_f(y)) \right\|^2 \leq \|x - y\|^2,$$
$$\Leftrightarrow \left\| \mathrm{prox}_f(x) - \mathrm{prox}_f(y) \right\|^2 \leq \langle x - y, \mathrm{prox}_f(x) - \mathrm{prox}_f(y) \rangle.$$

Proof.

□

Algorithm 4: Proximal gradient method for nonsmooth optimisation

Input: Initialisation $x_0 \in X$, proximal operator prox_g
Output: Minimiser $x^* = \arg\min_{x \in X} f(x) + g(x)$

```
for k = 1, 2, ..., until convergence do
    z_{k+1} = x_k − (1/β) ∇f(x_k) ;
    x_{k+1} = prox_{(1/β) g}(z_{k+1})
return x^{(k+1)}
```

This subsumes everything till date

1. Gradient descent $g = 0$
2. Projected gradient descent $g = \iota_C$
3. * Proximal point algorithm $f = 0$, $x_{k+1} = \mathrm{prox}_{\rho g}(x_k)$, $\rho > 0$.

Q: Is there a connection between proximal operators and convex optimisation?
A: . . .

Proposition 6.6
Let $\psi \in \Gamma_0(X)$ and $c > 0$. Then,

$$x^* = \arg\min_X \psi \Leftrightarrow x^* = \mathrm{prox}_{c\psi}(x^*).$$

Lemma 6.7
Let f be convex and β-smooth, and let $g \in \Gamma_0(X)$. Define $T : X \to X$ as follows:

$$T(x) = \mathrm{prox}_{\beta^{-1} g}(x - \beta^{-1} \nabla f(x)).$$

Then,

1. $x^* = \arg\min_X(f+g) \Leftrightarrow x^* \in \text{Fix}\{T\}$
2. $T = \text{prox}_\psi$ for some $\psi \in \Gamma_0(X)$. In particular, T is firmly-nonexpansive, and the fixed point iterations $x_{k+1} = T(x_k)$, which is also the proximal gradient method, is convergent.

7 Fenchel Conjugate

Definition 7.1 (Fenchel conjugate)
Let $f : X \to \bar{\mathbb{R}}$. Then, the conjugate function $f^* : X \to \bar{\mathbb{R}}$ is defined as

$$f^*(u) = \sup_{x \in X} \langle x, u \rangle - f(x)$$

Proposition 7.2
Let $f : X \to \bar{\mathbb{R}}$. Then, $f^* : X \to \bar{\mathbb{R}}$ is convex and lower semicontinuous.

Example
Let $C \subseteq X$ and let $f(x) = \iota_C(x)$. Then $f^*(u) = \sigma_C(u) \overset{\text{def.}}{=} \sup_{x \in C} \langle x, u \rangle$.

Example
Let $\alpha > 0$. Then,
1. $(\alpha|\cdot|)^* = \iota_{[-\alpha,+\alpha]}$,
2. $(\alpha\|\cdot\|_1)^* = \iota_{\mathcal{B}_{\infty,\alpha}}$.

Example
Let $y \in X$. Then, $\left(\frac{1}{2}\|\cdot - y\|^2\right)^* = \frac{1}{2}\|\cdot + y\|^2 - \frac{1}{2}\|y\|^2$

7.1 Properness of Fenchel Conjugates

Theorem 7.3
Let $f : X \to]-\infty, +\infty]$ be convex and proper. Then, f^* is convex, lower semicontinuous and proper.

Theorem 7.4 (Fenchel-Young inequality)
Let $f : X \to]-\infty, +\infty]$. Then,

$$\forall x, u \in X, \quad f(x) + f^*(u) \geq \langle x, u \rangle.$$

7.2 Biconjugate Theorem

Lemma 7.5
Let $f : X \to]-\infty, +\infty]$. Then, $f^{**} \leq f$.

Theorem 7.6
Let $f : X \to]-\infty, +\infty]$ be proper. Then,

$$f^{**} = f \iff f \text{ is convex and lower semicontinuous.}$$

Theorem 7.7
Let $f : X \to]-\infty, +\infty]$ be proper and convex, and let $x, u \in X$. Then, the following are equivalent.

1. $f(x) + f^*(u) = \langle x, u \rangle$,
2. $u \in \partial f(x)$,
   Further, if f is lower semicontinuous,
3. $x \in \partial f^*(u)$.

Corollary 7.8
Let $f : X \to]-\infty, +\infty] \in \Gamma_0(X)$. Then, $(\partial f)^{-1} = \partial f^*$.