Optimization Notes

Xiaopeng

September 7, 2025

1 Remainders from Multivariable Calculus

1.1 First-Order Conditions

The first-order necessary conditions for optimality can be expressed using the gradient of the objective function and the constraints. Specifically, if x^* is a local minimum of $f(x)$ Then

$$\nabla f(x^*) = 0$$

Proof: Writing the Taylor expansion of f around x^* gives

$$f(x) = f(x^*) + \nabla f(x^*)^T(x - x^*) + o(\|x - x^*\|)$$

Since x^* is a local minimum, we have $f(x) \geq f(x^*)$ for all x in a neighborhood of x^*. This implies that the first-order term must vanish, leading to the conclusion that $\nabla f(x^*) = 0$. More rigorously, we can consider the directional derivative of f at x^* in the direction of any vector d:

$$D_f(x^*; d) = \nabla f(x^*)^T d$$

Since x^* is a local minimum, the directional derivative must be non-negative for all feasible directions d (Univariate result: if f is differentiable and x^* is a local minimum, then $f'(x^*) = 0$). Otherwise this will be a decreasing direction, contradicting the local minimality of x^*. Therefore, we have:

$$D_f(x^*; d) \geq 0 \quad \forall d \in \mathcal{D}$$

In particular, if we take $d = -\nabla f(x^*)$, we find that

$$D_f(x^*; -\nabla f(x^*)) = -\|\nabla f(x^*)\|^2 \leq 0$$

This implies that $\nabla f(x^*) = 0$, completing the proof.

1.2 Second-Order Conditions

Aussume f is $\mathscr{C}^2$ and let x^* be a point such that $\nabla f(x^*) = 0$. Then

1. (Necessary condition) If x^* is a local minimum of f, then $\nabla^2 f(x^*) \in S_d^+(\mathbb{R})$ (the set of positive semi-definite matrices);

2. If $\nabla^2 \in S_d^{++}(\mathbb{R})$, then x^* is a strict local minimum of f;

3. If $\nabla^2 f(x^*)$ has at least one negative and one positive eigenvalue, then x^* is a saddle point of f: there exist two orthogonal directions e_1 and e_2 such that $t^* = 0$ is a local minimiser for $t \mapsto f(x^* + te_1)$ and a local maximiser for $t \mapsto f(x^* + te_2)$;

4. If $\nabla^2 f(x^*) \in S_d^{+}(\mathbb{R})$, but not in $S_d^{++}(\mathbb{R})$, then we cannot conclude and further analysis is required.

Proof: To prove the suffcient condition, we assume that

$$\nabla^2 f(x^*) \in S_d^{++}(\mathbb{R}) \tag{1}$$

Proposition 1.1 (Second-order mean value theorem)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be $\mathscr{C}^2$ on an open set containing the segment $[x, y] = \{x + t(y - x) : t \in [0, 1]\}$. Then there exists $\theta \in (0, 1)$ such that*

$$f(y) = f(x) + \nabla f(x)^\top (y - x) + \tfrac{1}{2} (y - x)^\top \nabla^2 f(x + \theta(y - x)) (y - x).$$

In particular, if $\nabla f(x) = 0$, then

$$f(y) = f(x) + \tfrac{1}{2} (y - x)^\top \nabla^2 f(x + \theta(y - x)) (y - x).$$

1. Let $d \in \mathbb{R}^d$ be any direction. By Taylor's theorem, we have

$$f(x^* + d) = f(x^*) + \nabla f(x^*)^T d + \frac{1}{2} d^T \nabla^2 f(x^*) d + o(\|d\|^2)$$

Since $\nabla f(x^*) = 0$, this simplifies to

$$f(x^* + d) = f(x^*) + \frac{1}{2} d^T \nabla^2 f(x^*) d + o(\|d\|^2)$$

By the minimality of x^*, we have $f(x^* + d) \geq f(x^*)$ for all sufficiently small d. This implies that

$$\frac{1}{2} d^T \nabla^2 f(x^*) d + o(\|d\|^2) \geq 0$$

for all sufficiently small d. Since the $o(\|d\|^2)$ term becomes negligible as d approaches zero, we conclude that

$$\frac{1}{2} d^T \nabla^2 f(x^*) d \geq 0$$

for all sufficiently small d. This is equivalent to saying that $\nabla^2 f(x^*)$ is positive semi-definite, which proves the necessary condition.

2. (Proof by contradiction, also a bit like contrapositive) Assume that there exists a sequence $\{x_k\}_{k \in \mathbb{N}}$ such that

$$\forall k \in \mathbb{N}, f(x_k) \leq f(x^*)$$

Mean value formula suggests that for x_k there exist $\xi_k \in [x_k, x^*]$ such that

$$f(x_k) = f(x^*) + \frac{1}{2} \langle \nabla^2 f(\xi_k)(x_k - x^*), (x_k - x^*) \rangle$$

The key point here is to use the exact mean value formula then which avoids any $o(\|d\|^2)$ terms. The tricky part is to carefully choose a sequence on the right hand side whose limit's Euclidean norm is strictly positive, and the sequence should also be a simple transformation from the sequence $x_k - x^*$. For this purpose, we define:

$$z_k := \frac{x_k - x^*}{\|x_k - x^*\|}$$

For any $k \in \mathbb{N}$ we have $\|z_k\| = 1$. By the compactness of the unit sphere, we can extract a convergent subsequence $z_{k_j} \to z^*$ for some $z^* \in S^{d-1}$. For the sake of clean notation, we will still denote the subsequence by z_k. We assume that z_k converges to $z_\infty \in \mathbb{S}^{d-1}$. Taking the result from mean value formula, we have (up to multiplying by a suitable factor) :

$$\langle \nabla^2 f(\xi_k) z_k, z_k \rangle \leq 0$$

Since $\xi_k \to x^*$ as $k \to \infty$, by the continuity of $\nabla^2 f$ we have $\nabla^2 f(\xi_k) \to \nabla^2 f(x^*)$. Taking the limit on both sides gives

$$\langle \nabla^2 f(x^*) z_\infty, z_\infty \rangle \leq 0$$

which contradicts the assumption that $\nabla^2 f(x^*)$ is positive definite ($\|z_\infty\|^2 = 1$).

3. Let e_1 and e_2 be the eigenvectors corresponding to the negative and positive eigenvalues of $\nabla^2 f(x^*)$, respectively. Consider the functions,

$$g_1(t) = f(x^* + te_1), \quad g_2(t) = f(x^* + te_2)$$

for $t \in \mathbb{R}$. By the chain rule (**use it or it's proving it myself in a paticular case**),

$$g'(t) = \frac{d}{dt} f(\gamma(t)) = \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(\gamma(t)) \frac{d}{dt}\gamma_k(t) = \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(x^* + tv)\, v_k.$$

$$g_1'(t) = \langle \nabla f(x^* + te_1), e_1 \rangle, \qquad g_2'(t) = \langle \nabla f(x^* + te_2), e_2 \rangle,$$

and

$$\frac{d}{dt}\Big(\frac{\partial f}{\partial x_k}(x^* + tv)\Big) = \sum_{\ell=1}^{n} \frac{\partial^2 f}{\partial x_k \partial x_\ell}(x^* + tv) \frac{d}{dt}(x_\ell^* + tv_\ell) = \sum_{\ell=1}^{n} \frac{\partial^2 f}{\partial x_k \partial x_\ell}(x^* + tv)\, v_\ell.$$

$$g_1''(t) = \langle \nabla^2 f(x^* + te_1)\, e_1, e_1 \rangle, \qquad g_2''(t) = \langle \nabla^2 f(x^* + te_2)\, e_2, e_2 \rangle.$$

In particular,

$$g_1'(0) = \langle \nabla f(x^*), e_1 \rangle = 0, \quad g_2'(0) = \langle \nabla f(x^*), e_2 \rangle = 0,$$

and, since e_i are eigenvectors of $\nabla^2 f(x^*)$ with eigenvalues λ_i,

$$g_1''(0) = \lambda_1 < 0, \qquad g_2''(0) = \lambda_2 > 0.$$

By continuity of $\nabla^2 f$, there exists $\delta > 0$ such that for $|t| < \delta$,

$$g_1''(t) < 0 \quad \text{and} \quad g_2''(t) > 0.$$

Hence $t = 0$ is a strict local maximizer of g_1 and a strict local minimizer of g_2, so x^* is a saddle point.

1.3 Coercive functions

Definition 1.2. We say that a continuous function $f : \mathbb{R}^d \to \mathbb{R}$ is coercive if

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

Remark

Another definition is for any $M \in \mathbb{R}$ the sub-level set $\{x \in \mathbb{R}^d : f(x) \leq M\}$ is compact (bounded).

Proposition 1.3. *If $f : \mathbb{R}^d \to \mathbb{R}$ is coercive, then it attains a global minimum.*

Proof: Let $(x_k)_{k \in \mathbb{N}} \subset \mathbb{R}^d$ be a minimizing sequence for f, i.e.,

$$\lim_{k \to \infty} f(x_k) = \inf_{x \in \mathbb{R}^d} f(x) = m.$$

Since f is coercive, we have $f(x_k) \to +\infty$ as $\|x_k\| \to \infty$. Thus, the sequence $(x_k)_{k \in \mathbb{N}}$ must be bounded. By the Bolzano-Weierstrass theorem, we can extract a convergent subsequence $(x_{k_j})_{j \in \mathbb{N}}$ such that

$$\lim_{j \to \infty} x_{k_j} = x^* \in \mathbb{R}^d.$$

By the continuity of f, we have

$$\lim_{j \to \infty} f(x_{k_j}) = f(x^*).$$

Combining these limits gives

$$f(x^*) = \lim_{j \to \infty} f(x_{k_j}) = m,$$

which shows that f attains its global minimum at x^*.

1.4 Exercises

Exercise 1.4. Let $A \in S_d(\mathbb{R})$.

1. Letting $\lambda_1(A) \leq \lambda_2(A) \leq \cdots \leq \lambda_d(A)$ be the eigenvalues of A, show that

$$\lambda_1(A) = \inf_{\|z\|_2 = 1} \langle Az, z \rangle.$$

2. Show that for any two $A, B \in S_d(\mathbb{R})$ there holds

$$|\lambda_1(A) - \lambda_1(B)| \leq \|A - B\|_{\text{op}}$$

where $\| \cdot \|_{\text{op}}$ stands for the standard operator norm on the set of matrices.

Réponse:

1. Let $(e_1 ... e_d)$ be a basis of unit eigenvectors of A such that $Ae_i = \lambda_i e_i$ for all i.

From the properties of eigenvectors and eigenvalues, we know that the eigenvectors corresponding to distinct eigenvalues are orthogonal. Therefore, we have:

$$\langle e_i, e_j \rangle = 0 \quad \text{for } i \neq j$$

and

$$\langle e_i, e_i \rangle = 1.$$

For any $z \in \mathbb{R}^d,\ \|z\|_2 = 1$, we can express z in terms of the basis $(e_1, \ldots, e_d)$:

$$z = \sum_{i=1}^{d} \langle z, e_i \rangle e_i,\ \|z\|_2^2 = \sum_{i=1}^{d} \langle z, e_i \rangle^2 = 1$$

Any vector z can be decomposed into its components along the eigenvectors of A. We can then compute $\langle Az, z \rangle$:

$$\langle Az, z \rangle = \langle A \left(\sum_{i=1}^{d} \langle z, e_i \rangle e_i \right), \sum_{j=1}^{d} \langle z, e_j \rangle e_j \rangle$$

Using the linearity of A and the fact that $Ae_i = \lambda_i e_i$, we get:

$$\langle Az, z \rangle = \langle \sum_{i=1}^{d} \langle z, e_i \rangle \lambda_i e_i, \sum_{j=1}^{d} \langle z, e_j \rangle e_j \rangle$$

By the orthogonality of the eigenvectors, this simplifies to:

$$\langle Az, z \rangle = \sum_{i=1}^{d} \langle z, e_i \rangle^2 \lambda_i$$

Since $\sum_{i=1}^{d} \langle z, e_i \rangle^2 = 1$ and each λ_i is bounded between $\lambda_{\min}$ and $\lambda_{\max}$, we have:

$$\inf_{\|z\|_2=1} \langle Az, z \rangle = \inf_{\|z\|_2=1} \sum_{i=1}^{d} \langle z, e_i \rangle^2 \lambda_i \leq \sum_{i=1}^{d} \langle z, e_i \rangle^2 \lambda_i$$

and

$$\sup_{\|z\|_2=1} \langle Az, z \rangle = \sup_{\|z\|_2=1} \sum_{i=1}^{d} \langle z, e_i \rangle^2 \lambda_i \geq \sum_{i=1}^{d} \langle z, e_i \rangle^2 \lambda_i$$

Assume that $c_1, c_2 \geq 0$ such that $c_1 + c_2 = 1$ and $a, b \in \mathbb{R}$ such that $a \leq b$.

$$a = c_1 a + c_2 a \leq c_1 a + c_2 b \leq c_1 b + c_2 b = b$$

Generalizing the above result to more variables, we have:

$$\lambda_1(A) = \inf_{\|z\|_2=1} \langle Az, z \rangle \leq \langle Az, z \rangle = \sum_{i=1}^{d} \langle z, e_i \rangle^2 \lambda_i \leq \sup_{\|z\|_2=1} \langle Az, z \rangle = \lambda_d(A).$$

2. Using the variational characterization from part (1), we have:

$$\lambda_1(A) = \inf_{\|z\|_2=1} \langle Az, z\rangle \quad \text{and} \quad \lambda_1(B) = \inf_{\|z\|_2=1} \langle Bz, z\rangle$$

For any unit vector z with $\|z\|_2 = 1$, we can write:

$$\langle Az, z\rangle - \langle Bz, z\rangle = \langle (A-B)z, z\rangle$$

By the Cauchy-Schwarz inequality and the definition of operator norm:

$$|\langle (A-B)z, z\rangle| \leq \|(A-B)z\|_2\|z\|_2 \leq \|A-B\|_{\text{op}}\|z\|_2^2 = \|A-B\|_{\text{op}}$$

Therefore:

$$\langle Az, z\rangle \leq \langle Bz, z\rangle + \|A-B\|_{\text{op}}$$

Taking the infimum over all unit vectors z on the left side:

$$\lambda_1(A) = \inf_{\|z\|_2=1} \langle Az, z\rangle \leq \inf_{\|z\|_2=1} (\langle Bz, z\rangle + \|A-B\|_{\text{op}}) = \lambda_1(B) + \|A-B\|_{\text{op}}$$

This gives us:

$$\lambda_1(A) - \lambda_1(B) \leq \|A-B\|_{\text{op}}$$

By symmetry (swapping the roles of A and B), we also have:

$$\lambda_1(B) - \lambda_1(A) \leq \|B-A\|_{\text{op}} = \|A-B\|_{\text{op}}$$

Combining both inequalities:

$$|\lambda_1(A) - \lambda_1(B)| \leq \|A-B\|_{\text{op}}$$

Remark

The operator norm (or spectral norm) of a matrix $A \in \mathbb{R}^{d\times d}$ is defined as:

$$\|A\|_{\text{op}} = \sup_{\|x\|_2=1} \|Ax\|_2 = \sup_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2}$$

This norm measures the maximum amplification factor of the matrix when applied to unit vectors. For symmetric matrices, the operator norm equals the largest absolute eigenvalue: $\|A\|_{\text{op}} = \max_i |\lambda_i(A)|$. The operator norm is induced by the Euclidean norm and satisfies the submultiplicative property: $\|AB\|_{\text{op}} \leq \|A\|_{\text{op}}\|B\|_{\text{op}}$. It provides a measure of how much a linear transformation can stretch vectors and is fundamental in analyzing the conditioning and stability of linear systems.

Exercise 1.5. Let $A \in S_d(\mathbb{R})$ and $b \in \mathbb{R}^d$. We consider

$$f : x \mapsto \tfrac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle.$$

1. Show that f is coercive if, and only if $A \in S_d^{++}(\mathbb{R})$.
2. Show that f is convex if, and only if $A \in S_d^{+}(\mathbb{R})$.

3. Show that f is strictly convex if, and only if $A \in S_d^{++}(\mathbb{R})$.

Réponse:

1. Let's show the contrapositive of " $\implies$ " : if $A \notin S_d^{++}(\mathbb{R})$, then f is not coercive. We can safely omit $\langle b, x\rangle$ in the definition since it is not a part of the dominant term when $\|x\|_2 \to \infty$. If $A \notin S_d^{++}(\mathbb{R})$, then an eigenvalue $\lambda_d \leq 0$ (adopting the notation from last exercise). This implies that there exists a sequence $(x_n) \subset \mathbb{R}^d$ (up to choosing from the eigenspace E_{λ_d} associated with λ_d) such that $\|x_n\|_2 \to \infty$ and $\langle Ax_n, x_n\rangle \to -\infty$ or 0 which shows that f is not coercive.

As for the converse " $\impliedby$ " : **We find a lower bound for** $\langle Ax, x\rangle$. Let $A \in S_d^{++}(\mathbb{R})$. Then all eigenvalues are positive, and we can find a constant $\alpha = \lambda_{\min}(A) > 0$ such that

$$\langle Ax, x\rangle \geq \alpha\|x\|_2^2 \quad \forall x \in \mathbb{R}^d.$$

This implies that

$$f(x) = \tfrac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle \geq \tfrac{\alpha}{2}\|x\|_2^2 - \langle b, x\rangle.$$

Now, if $\|x\|_2 \to \infty$, the term $\frac{\alpha}{2}\|x\|_2^2$ dominates $-\langle b, x\rangle$, and we conclude that $f(x) \to \infty$. Thus, f is coercive.

2. We use the characterization of convexity through the Hessian matrix: f is convex if, and only if, $\nabla^2 f(x) \succeq 0$ for all $x \in \mathbb{R}^d$. The result is trivial since $\nabla^2 f(x) = A \succeq 0$ for all $x \in \mathbb{R}^d$.

3. We use the characterization of strict convexity through the Hessian matrix: f is strictly convex if, and only if, $\nabla^2 f(x) \succ 0$ for all $x \in \mathbb{R}^d$. The result is again trivial since $\nabla^2 f(x) = A \succ 0$ for all $x \in \mathbb{R}^d$.

Exercise 1.6. Classify the critical points (local minimisers, local maximisers, saddle points, indeterminate critical points) of the following functions:

1. $f_1 : (x, y) \mapsto (x-y)^2 + (x+y)^3$,
2. $f_2 : (x, y) \mapsto x^2 - 2y^2 + 3xy$,
3. $f_3 : (x, y) \mapsto x^4 + y^3 - 3y - 2$.

Réponse:

1. The critical points of f_1 can be found by computing the gradient and setting it to zero:

$$\nabla f_1(x, y) = \begin{pmatrix} 2(x-y) + 3(x+y)^2 \\ 2(y-x) + 3(x+y)^2 \end{pmatrix} = 0.$$

This gives us a system of equations to solve for the critical points. Simplifying the equations,

$$\begin{cases} 3(x+y)^2 = 2(x-y) \\ 3(x+y)^2 = -2(x-y) \end{cases}$$

Which leads to $x = y = 0$ as the only critical point. This critical point is a saddle point, and we only need to take the direction $\vec{d} = (1, 1)$.

2. For f_2, it is worth noticing that f_2 can be expressed with a quadratic form:

$$f_2(x, y) = \frac{1}{2} \begin{pmatrix} x \\ y \end{pmatrix}^T \begin{pmatrix} 2 & 3 \\ 3 & -4 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

we compute the gradient:

$$\nabla f_2(x, y) = \begin{pmatrix} 2 & 3 \\ 3 & -4 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

Again, we have a system of equations to solve. The unique solution is given by:

$$\begin{pmatrix} 2 & 3 \\ 3 & -4 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 0$$

which leads to $x = 0$ and $y = 0$. To classify this critical point, we compute the Hessian matrix:

$$\nabla^2 f_2(x, y) = \begin{pmatrix} 2 & 3 \\ 3 & -4 \end{pmatrix}$$

The eigenvalues of this Hessian matrix can be found by solving the characteristic polynomial:

$$\det\left(\nabla^2 f_2(x, y) - \lambda I\right) = 0$$

which simplifies to:

$$\det \begin{pmatrix} 2 - \lambda & 3 \\ 3 & -4 - \lambda \end{pmatrix} = 0$$

The characteristic polynomial is given by:

$$(2 - \lambda)(-4 - \lambda) - 9 = 0$$

which leads to:

$$\lambda^2 + 2\lambda - 17 = 0$$

The eigenvalues are:

$$\lambda_{1,2} = -1 \pm \sqrt{18} = -1 \pm 3\sqrt{2}$$

Since one eigenvalue is positive and the other is negative, the critical point is a saddle point.

3. Finally, for f_3:

$$\nabla f_3(x, y) = \begin{pmatrix} 4x^3 \\ 3y^2 - 3 \end{pmatrix} = 0.$$

We can solve these equations to find the critical points. The solutions are:

$$x = 0, \quad y = \pm 1$$

To classify these critical points, we compute the Hessian matrix:

$$\nabla^2 f_3(x,y) = \begin{pmatrix} 12x^2 & 0 \\ 0 & 6y \end{pmatrix}$$

The eigenvalues of this Hessian matrix are given by the diagonal elements:

$$\lambda_1 = 12x^2, \quad \lambda_2 = 6y$$

At both critical points the Hessian is degenerate in the x-direction (entry $12x^2 = 0$), so the usual second derivative test is inconclusive and we use higher-order expansion.

1. Point $(0,1)$. Write $x = x - 0,\ y = y - 1$ (move first critical point to O):

$$f_3(x,y) = x^4 + y^3 - 3y^2$$

We examine $f(x,y) - f(0,0) = x^4 + y^3 - 3y^2$ along various paths through the origin:

(a) **Along the x-axis** ($y = 0$):

$$f(x,0) = x^4 \geq 0$$

with equality only at $x = 0$.

(b) **Along the y-axis** ($x = 0$):

$$f(0,y) = y^3 - 3y^2 = y^2(y-3)$$

For small $|y| > 0$: since $y - 3 < 0$, we have $f(0,y) < 0$.

(c) **Along the parabola** $y = x^2$:

$$f(x,x^2) = x^4 + x^6 - 3x^4 = x^6 - 2x^4 = x^4(x^2 - 2)$$

For small $|x| > 0$: since $x^2 - 2 < 0$, we have $f(x,x^2) < 0$.

Since the function takes both positive values (along the x-axis) and negative values (along the y-axis and the parabola $y = x^2$) in every neighborhood of $(0,0)$, we conclude that:

$$\boxed{(0,0) \text{ is a } \textbf{saddle point} \text{ of } f_3(x,y) = x^4 + y^3 - 3y^2}$$

2. Point $(0,-1)$. Write $y = -1 + s$:

$$f_3(x,-1+s) = x^4 + (-1+s)^3 - 3(-1+s) - 2 = x^4 - 3s^2 + s^3.$$

Thus

$$f_3(x,-1+s) - f_3(0,-1) = x^4 - 3s^2 + s^3.$$

Along $s = 0$, $x \neq 0$: difference $= x^4 > 0$. Along $x = 0$, $0 < s < 3$: difference $= -3s^2 + s^3 = -3s^2(1 - s/3) < 0$. Hence values of both signs occur arbitrarily close to $(0,-1)$: $(0,-1)$ is a saddle point.

Conclusion:

$$\boxed{\text{Both } (0,1) \text{ and } (0,-1) \text{ are } \textbf{saddle points} \text{ of } f_3(x,y) = x^4 + y^3 - 3y - 2}$$

Exercise 1.7 (Distance between two sets)**.** Let A and B be two closed, nonempty subsets of $\mathbb{R}^d$.

1. Show that if A is compact, then the problem

$$\min_{a\in A,\, b\in B} \|a-b\|$$

   has a solution (at least one).

2. Show with a counter-example that this problem need not have a solution if neither A nor B is assumed compact, even if A and B are convex.

Exercise 1.8. Give an example of a strictly convex function $\varphi : \mathbb{R}^d \to \mathbb{R}$ such that the equation

$$\nabla^2\varphi(x) = 0$$

has infinitely many solutions.

Exercise 1.9 (Carathéodory theorem)**.** Let $\Omega \subset \mathbb{R}^d$. We call the convex hull of Ω the smallest convex set containing Ω, denoted $C(\Omega)$.

1. Show that

$$C(\Omega) = \left\{ \sum_{i=0}^{N} t_i x_i \;\middle|\; N\in\mathbb{N},\ t_i\in[0,1],\ \sum_{i=0}^{N} t_i = 1,\ x_i\in\Omega \right\}.$$

2. Prove the Carathéodory theorem: for any $x \in C(\Omega)$, there exist $t_0, \dots, t_d \in [0,1]$ and $x_0, \dots, x_d \in \Omega$ such that

$$\sum_{i=0}^{d} t_i = 1, \quad x = \sum_{i=0}^{d} t_i x_i.$$

   (a) Using an example, show why one needs at least $(d+1)$ points.

   (b) Prove the theorem by descending induction, starting from a representation with $d+2$ points and eliminating one.

   (c) Deduce that if Ω is compact, then so is $C(\Omega)$.

Exercise 1.10 (Extreme points I: projection on closed convex sets)**.** Let $K \subset \mathbb{R}^d$ be a closed convex set. Show that there exists a unique $z \in K$, denoted by $\Pi_K(x)$ and called the orthogonal projection of x on K, such that

$$\|x - \Pi_K(x)\| = \min_{z\in K} \|x - z\|$$

and that

$$\forall y \in K,\ \langle x - \Pi_K(x),\, y - \Pi_K(x)\rangle \le 0.$$

Show that Π_K is 1-Lipschitz.

Exercise 1.11 (Extreme points II: The Krein-Milman theorem)**.** 1. Give an example of a convex set $K \subset \mathbb{R}^d$ that has no extreme points.

2. Assume K is closed. Prove that K has extreme points.

3. Prove the finite-dimensional Krein-Milman theorem: any $x \in K$ is a convex combination of extreme points of K.

   (a) Let $x \in \partial K$. Show that there exists a supporting hyperplane $H = \{\varphi = 0\}$ with $\varphi \in (\mathbb{R}^d)^*, \varphi \neq 0$, such that $x \in H$ and $\varphi(K) \subset (-\infty, 0]$.

   (b) Show that if $x \in H$ for some supporting hyperplane of K, then x is an extreme point of K iff it is an extreme point of $H \cap K$.

   (c) Conclude the theorem by induction on the dimension.

Exercise 1.12 (Polyak–Łojasiewicz Inequality)**.** Let $f : \mathbb{R}^d \to \mathbb{R}$ be an α-strongly convex function and let x^* be a minimiser of f.

1. Prove that
$$\forall x \in \mathbb{R}^d, \quad \|x - x^*\|^2 \leq \tfrac{2}{\alpha}\big(f(x) - f(x^*)\big).$$

2. Show that
$$\forall x \in \mathbb{R}^d, \quad f(x) - f(x^*) \leq \tfrac{1}{2\alpha}\|\nabla f(x)\|^2.$$

3. Deduce that
$$\forall x \in \mathbb{R}^d, \quad \|x - x^*\| \leq \tfrac{1}{\alpha}\|\nabla f(x)\|.$$