TC2: Optimization for Machine Learning

Master of Science in AI and Master of Science in Data Science
@ UPSaclay
2025/2026.

Faïcel Chamroukhi

chamroukhi.com

week 5 : December 04, 2025.

Constrained optimization

((in)equality constraints, Duality/Lagrangian, KKT optimality conditions)

Constrained Optimization Problem

- **Objective :** Minimize or maximize a function $f(x)$ subject to constraints.
- **General Form :**

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_j(x) = 0, \quad j = 1, \ldots, p \end{aligned}$$

  - $f(x)$: Objective function.
  - $g_i(x)$: Inequality constraints.
  - $h_j(x)$: Equality constraints.

- Budget limits in economics.
- Physical constraints in engineering.
- sparcity or regularity constraints in machine learning
- etc

Feasible Sets and Feasible Solutions I

1. **Feasible Set :**
   - The feasible set (or feasible region) is the set of all points that satisfy the constraints of an optimization problem.
   - Formally, for a problem with constraints $g_i(x) \leq 0$ and $h_j(x) = 0$, the feasible set S is :

$$S = \{x \in \mathbb{R}^n \mid g_i(x) \leq 0,\ h_j(x) = 0,\ \forall i, j\}$$

   - Only points within this set can be considered as potential solutions to the optimization problem.
   - Constraints restrict the feasible solutions domain, narrowing optimum search
2. **Feasible Solution :**
   - A feasible solution is any point $x \in S$ that satisfies all problem constraints.
   - An optimal solution, if it exists, is a feasible solution that minimizes (or maximizes) the objective function within the feasible set.

Example

Example of feasible region for a set of linear inequality constraints.

- Constraints for the feasible region :

$$
\begin{aligned}
&x + y \leq 4 \\
&x \geq 0 \\
&y \geq 0 \\
&y \leq 3
\end{aligned}
$$

- Plots of each constraint line :
  - $y = 4 - x$: Boundary for $x + y \leq 4$.
  - $x = 0$: Vertical line for $x \geq 0$.
  - $y = 3$: Horizontal line representing $y \leq 3$.

Example

- The feasible region is the intersection of the regions defined by each constraint.
- The feasible region, represented by the shaded area, satisfies all specified constraints. Only points within this shaded area are feasible solutions

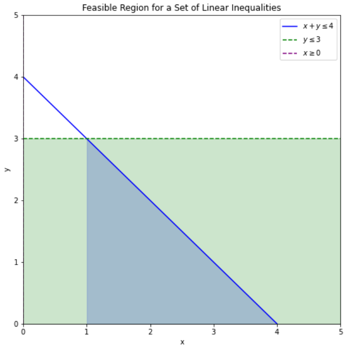

FIGURE – Feasible region for a set of linear inequalities : the constraints limit the solution space.

Mathematical tools help us handle constraints effectively.

Consider the problem (will be referred to as the **primal problem**)

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & h_j(x) = 0, \quad j = 1, \ldots, p \end{aligned}$$

Lagrange Multipliers Method :

- The **Lagrangian function** is defined as :

$$\mathcal{L}(x, \lambda) = f(x) + \sum_{j=1}^{p} \lambda_j h_j(x),$$

  where λ_j are the Lagrange multipliers.

- **Dual problem** : minimize w.r.t x and λ_i's the lagrangian $\mathcal{L}(x, \lambda)$
- Optimality conditions :

$$\nabla \mathcal{L}(x, \lambda) = 0, \quad h_j(x) = 0 \text{ for all } j.$$

Consider the problem (will be referred to as the **primal problem**)

$$\min_{x\in\mathbb{R}^n} \quad f(x)$$
$$\text{s.t.} \quad h_j(x) = 0, \quad j = 1, \dots, p$$

Lagrange Multipliers Method :

- The **Lagrangian function** is defined as :

$$\mathcal{L}(x, \lambda) = f(x) + \sum_{j=1}^{p} \lambda_j h_j(x),$$

  where λ_j are the Lagrange multipliers.
- **Dual problem** : minimize w.r.t x and λ_i's the lagrangian $\mathcal{L}(x, \lambda)$
- Optimality conditions :

$$\nabla \mathcal{L}(x, \lambda) = 0, \quad h_j(x) = 0 \text{ for all } j.$$

Theorem : Let x^* be a local minimum of $f(x)$ subject to equality constraints $h_j(x) = 0$ for $j = 1, \ldots, p$. If x^* is a *regular point* (the gradients $\nabla h_1(x^*), \ldots, \nabla h_p(x^*)$ are linearly independent), there exist Lagrange multipliers $\lambda_1, \lambda_2, \ldots, \lambda_p$ such that :

$$\nabla f(x^*) + \sum_{j=1}^{p} \lambda_j \nabla h_j(x^*) = 0, \quad h_j(x^*) = 0, \quad j = 1, \ldots, p.$$

- The condition $\nabla f(x^*) + \sum_{j=1}^{p} \lambda_j \nabla h_j(x^*) = 0$ ensures that the gradients of $f(x)$ and the constraints $h_j(x)$ are aligned, defining a stationary point of the Lagrangian function.
- The equality constraints $h_j(x^*) = 0$ ensure feasibility of the solution x^*.
- A *regular point* ensures that the constraint gradients are linearly independent, avoiding degenerate or redundant constraints.

Example : [Practical work]

$$\min_{x\in\mathbb{R}^2} \quad f(x_1, x_2) = x_1^2 + x_2^2$$
$$\text{s.t.} \quad h(x_1, x_2) = x_1 + x_2 - 1 = 0.$$

Using Lagrange Multipliers :

- The **Lagrangian function** is :

$$\mathcal{L}(x_1, x_2, \lambda) = x_1^2 + x_2^2 + \lambda(x_1 + x_2 - 1),$$

where λ is the Lagrange multiplier.

- Optimality conditions : $\nabla \mathcal{L}(x_1, x_2, \lambda) = \mathbf{0}$. Compute partial derivatives :

$$\frac{\partial \mathcal{L}}{\partial x_1} = 2x_1 + \lambda = 0,$$
$$\frac{\partial \mathcal{L}}{\partial x_2} = 2x_2 + \lambda = 0,$$

1. From $\frac{\partial \mathcal{L}}{\partial x_1} = 0$ and $\frac{\partial \mathcal{L}}{\partial x_2} = 0$, we have :

$$2x_1 + \lambda = 0 \implies \lambda = -2x_1,$$
$$2x_2 + \lambda = 0 \implies \lambda = -2x_2.$$

Equating the two expressions for λ : $-2x_1 = -2x_2 \implies x_1 = x_2$.

2. From the constraint : $h(x_1, x_2) = x_1 + x_2 - 1 = 0$: Substitute $x_1 = x_2$ into the constraint $x_1 + x_2 - 1 = 0$:

$$x_1 + x_1 = 1 \implies x_1 = x_2 = \frac{1}{2}.$$

3. The solution is : $x_1^* = \frac{1}{2}, \quad x_2^* = \frac{1}{2}, \quad \lambda^* = -1.$

Using Lagrange Multipliers :

- The **Lagrangian function** is :

$$\mathcal{L}(x_1, x_2, \lambda) = x_1^2 + x_2^2 + \lambda(x_1 + x_2 - 1),$$

  where λ is the Lagrange multiplier.

- Optimality conditions : $\nabla\mathcal{L}(x_1, x_2, \lambda) = \mathbf{0}$. Compute partial derivatives :

$$\frac{\partial \mathcal{L}}{\partial x_1} = 2x_1 + \lambda = 0,$$
$$\frac{\partial \mathcal{L}}{\partial x_2} = 2x_2 + \lambda = 0,$$

1. From $\frac{\partial \mathcal{L}}{\partial x_1} = 0$ and $\frac{\partial \mathcal{L}}{\partial x_2} = 0$, we have :

$$2x_1 + \lambda = 0 \quad \Longrightarrow \quad \lambda = -2x_1,$$
$$2x_2 + \lambda = 0 \quad \Longrightarrow \quad \lambda = -2x_2.$$

   Equating the two expressions for λ : $-2x_1 = -2x_2 \quad \Longrightarrow \quad x_1 = x_2.$

2. From the constraint : $h(x_1, x_2) = x_1 + x_2 - 1 = 0$: Substitute $x_1 = x_2$ into the constraint $x_1 + x_2 - 1 = 0$:

$$x_1 + x_1 = 1 \quad \Longrightarrow \quad x_1 = x_2 = \frac{1}{2}.$$

3. The solution is : $x_1^* = \frac{1}{2}, \quad x_2^* = \frac{1}{2}, \quad \lambda^* = -1.$

Optimization with Inequality Constraints

Remarks :

- If the regularity condition (linear independence of $\nabla h_j(x^*)$) is not satisfied, additional tools such as the Karush-Kuhn-Tucker (KKT) conditions are required to analyze the problem.
- Karush-Kuhn-Tucker (KKT) extend the method of Lagrange multipliers to handle inequality constraints.

Consider the optimization problem (primal form) :

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_j(x) = 0, \quad j = 1, \ldots, p \end{aligned}$$

The **Karush-Kuhn-Tucker (KKT) Conditions** are necessary conditions to check optimality in problems involving both equality and inequality constraints. They extend the method of Lagrange multipliers to handle inequality constraints.

$$\text{the Lagrangian : } \mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \mu_j h_j(x).$$

- **Stationarity :** The gradient of the Lagrangian w.r.t solution x must be zero :

$$\nabla \mathcal{L}(x, \lambda, \mu) = \nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x) + \sum_{j=1}^{\ell} \mu_j \nabla h_j(x) = 0.$$

- **Primal feasibility :** The solution x must satisfy all the constraints :

$$g_i(x) \leq 0, \quad h_j(x) = 0.$$

- **Dual feasibility :** The Lagrange multipliers $\lambda_i \geq 0$ for inequality constraints.
- **Complementary slackness :** For each i, either $\lambda_i = 0$ or $g_i(x) = 0$:

$$\lambda_i \cdot g_i(x) = 0, \quad \forall i = 1, \ldots, m.$$

The **Karush-Kuhn-Tucker (KKT) Conditions** are necessary conditions to check optimality in problems involving both equality and inequality constraints. They extend the method of Lagrange multipliers to handle inequality constraints.

$$\text{the Lagrangian : } \mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \mu_j h_j(x).$$

- **Stationarity :** The gradient of the Lagrangian w.r.t solution x must be zero :

$$\nabla \mathcal{L}(x, \lambda, \mu) = \nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x) + \sum_{j=1}^{\ell} \mu_j \nabla h_j(x) = 0.$$

- **Primal feasibility :** The solution x must satisfy all the constraints :

$$g_i(x) \leq 0, \quad h_j(x) = 0.$$

- **Dual feasibility :** The Lagrange multipliers $\lambda_i \geq 0$ for inequality constraints.
- **Complementary slackness :** For each i, either $\lambda_i = 0$ or $g_i(x) = 0$:

$$\lambda_i \cdot g_i(x) = 0, \quad \forall i = 1, \ldots, m.$$

Stationarity Condition

Stationarity :

$$\nabla \mathcal{L}(x, \lambda, \mu) = \nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x) + \sum_{j=1}^{p} \mu_j \nabla h_j(x) = 0.$$

Interpretation :

- At the optimal solution x^*, the gradient of the objective function $f(x)$ is balanced by the gradients of the active constraints $g_i(x)$ and $h_j(x)$.
- This condition ensures no further improvement in $f(x)$ is possible while satisfying the constraints.

Primal Feasibility

Primal Feasibility :

$$g_i(x) \leq 0, \quad h_j(x) = 0.$$

Interpretation :

- The solution x^* must satisfy :
  - All inequality constraints ($g_i(x) \leq 0$),
  - All equality constraints ($h_j(x) = 0$).
- Primal feasibility ensures the solution lies in the feasible region of the optimization problem.

Dual Feasibility

Dual Feasibility :

$$\lambda_i \geq 0, \quad \forall i = 1, \ldots, m.$$

Interpretation :

- The Lagrange multipliers λ_i associated with the inequality constraints must be non-negative.
- If $\lambda_i > 0$ this indicates the corresponding constraint $g_i(x)$ is active ($g_i(x) = 0$).
- If $\lambda_i = 0$, the corresponding inequality constraint $g_i(x)$ is inactive ($g_i(x) < 0$).

Complementary Slackness

Complementary Slackness :

$$\lambda_i \cdot g_i(x) = 0, \quad \forall i = 1, \ldots, m.$$

This implies one of the two things must be true for each constraint :

Interpretation :

- If $\lambda_i > 0$, then $g_i(x) = 0$, meaning the constraint is **active** and **binding** at the solution.
- If $g_i(x) < 0$, then $\lambda_i = 0$, meaning the constraint is **inactive** and does not affect the optimality condition.
- Complementary slackness ensures that inactive constraints do not influence the solution.

Hence the name *complementary slackness* :

- Either the constraint is **tight** (no slack, i.e., $g_i(x) = 0$) and it affects the optimum (the multiplier $\lambda_i \neq 0$) ;
- Or it is **slack** (strictly satisfied, i.e., $g_i(x) < 0$) and it does not affect the solution (the multiplier $\lambda_i = 0$).

Complementary Slackness

Complementary Slackness :

$$\lambda_i \cdot g_i(x) = 0, \quad \forall i = 1, \ldots, m.$$

This implies one of the two things must be true for each constraint :

Interpretation :

- If $\lambda_i > 0$, then $g_i(x) = 0$, meaning the constraint is **active** and **binding** at the solution.
- If $g_i(x) < 0$, then $\lambda_i = 0$, meaning the constraint is **inactive** and does not affect the optimality condition.
- Complementary slackness ensures that inactive constraints do not influence the solution.

Hence the name *complementary slackness* :

- Either the constraint is **tight** (no slack, i.e., $g_i(x) = 0$) and it affects the optimum (the multiplier $\lambda_i \neq 0$) ;
- Or it is **slack** (strictly satisfied, i.e., $g_i(x) < 0$) and it does not affect the solution (the multiplier $\lambda_i = 0$).

Summary

Summary of KKT Conditions :

- Stationarity : Ensures that the gradient of the objective function is aligned with the gradients of the active constraints.
- Primal Feasibility : Guarantees the solution lies within the feasible region.
- Dual Feasibility : Ensures the Lagrange multipliers λ_i are meaningful (non-negative).
- Complementary Slackness : Eliminates the influence of inactive constraints on the solution.

Optimality Check :

- Together, these conditions provide a framework to verify whether a candidate solution x^* is optimal in constrained optimization problems.

Summary

- Inequality constraints become **active** when $g_i(x^*) = 0$, contributing to the optimality conditions through $\lambda_i > 0$.
- Inactive constraints $(g_i(x^*) < 0)$ have $\lambda_i = 0$, meaning they do not influence the solution.
- Complementary slackness ensures that inactive constraints (those with $g_i(x^*) < 0$) do not contribute to the optimality condition.
- Equality constraints $(h_j(x^*) = 0)$ are always active and satisfied exactly.

- The gradient of the resulting objective function is a linear combination of the gradients of the active constraints : The gradients of $f(x)$, $g_i(x)$, and $h_j(x)$ at x^* reflecting a balance between optimizing the objective function and respecting the constraints.

- Second-order conditions extend the KKT (first optimality) conditions to ensure a critical point satisfying KKT (a local minimum, maximum, (or saddle point)) is a local optimum.
- **Hessian of the Lagrangian :**

$$H = \nabla^2 f(x) + \sum_i \lambda_i \nabla^2 g_i(x) + \sum_j \mu_j \nabla^2 h_j(x)$$

1. **Second-Order Sufficiency :** For a minimization problem, $d^T H d > 0$ for all $d \neq 0$ satisfying :

$$\nabla g_i(x)^T d = 0 \quad \text{(active inequality constraints)}$$

$$\nabla h_j(x)^T d = 0 \quad \text{(equality constraints).}$$

2. **Second-Order Necessity :** $d^T H d \geq 0$ for all $d \neq 0$ satisfying the same constraint conditions as SOSC.

Theorem : KKT Conditions

Theorem : Let $f(x)$, $g_i(x)$, and $h_j(x)$ be continuously differentiable. If x^* is a local minimum and satisfies certain regularity conditions, then there exist $\lambda_i \geq 0$ and μ_j such that the KKT conditions hold.

Definition :

- The **dual function**, $\varphi(\lambda, \mu)$, is obtained by minimizing the Lagrangian with respect to the primal variable x :

$$\varphi(\lambda, \mu) = \inf_x \mathcal{L}(x, \lambda, \mu).$$

- The dual function $\varphi(\lambda, \mu)$ provides a lower bound to the primal problem for any $\lambda \geq 0$ and any μ.
- The dual function is always concave (the inf of an affine transformation is a concave function, and $\mathcal{L}$ is a linear combination of λ and μ, so produces a function that is concave in λ and μ, regardless of whether $\mathcal{L}$ is convex or not in x.

Dual function importance

- **Duality Gap :** The difference between the primal optimal value $f(x^*)$ and the dual optimal value $\varphi(\lambda^*, \mu^*)$, known as the **duality gap**, quantifies how close the solution of the dual problem is to the solution of the primal problem.
- If the duality gap is zero, the dual solution exactly matches the primal solution, indicating perfect alignment between the two.

Dual Problem :

- The dual problem is derived by minimizing the Lagrangian over x :

$$\varphi(\lambda^*, \mu^*) = \inf_x \mathcal{L}(x, \lambda, \mu).$$

- The dual problem is (recall the dual function is concave in λ and μ) :

$$\max_{\lambda \geq 0, \mu} \varphi(\lambda^*, \mu^*).$$

- Weak Duality :

$$f(x^*) \geq \varphi(\lambda^*, \mu^*).$$

- **Strong Duality :** If strong duality holds :

$$\varphi(\lambda^*, \mu^*) = f(x^*),$$

  where x^* is the optimal solution of the primal problem, and (λ^*, μ^*) are the optimal dual variables.

Strong Duality :

- If strong duality holds, $f(x^*) = \varphi(\lambda^*, \mu^*)$.

Theorem : (Slater's Condition) For a convex optimization problem, if there exists a strictly feasible point x (one that satisfies $g_i(x) < 0, h_j(x) = 0$), then strong duality holds.

- Strong duality ensures that solving the dual problem gives the exact same result as solving the primal problem :
  - primal problem (minimizing the original objective function, i.e. f s.t. the constraints),
  - dual problem (maximizing the dual function, ie. the Lagrangian $\mathcal{L}$).

- **Duality** consists of associating to every constrained minimization problem (the **primal problem**) another problem called the **dual problem**.
- Duality connects a constrained optimization problem (primal) to an associated maximization problem (dual), built from the **Lagrangian function**.
  - The **dual function** always provides a **lower bound** to the primal optimal value ; $\hookrightarrow$ this is known as **weak duality**.
  - Under suitable conditions such as **convexity** and **Slater's condition**, the primal and dual problems have the same optimal value ; $\hookrightarrow$ this is known as **strong duality**. In this case, the **duality gap** is zero, and solving the dual problem yields the same optimal value as solving the primal one.