# Modern Optimization Techniques

## 3. Equality Constrained Optimization / 3.1. Duality

Lars Schmidt-Thieme

Information Systems and Machine Learning Lab (ISMLL)
Institute for Computer Science
University of Hildesheim, Germany

# Syllabus

| | | |
|---|---|---|
| Mon. 30.10. | (1) | 0. Overview |
| | | **1. Theory** |
| Mon. 6.11. | (2) | 1. Convex Sets and Functions |
| | | **2. Unconstrained Optimization** |
| Mon. 13.11. | (3) | 2.1 Gradient Descent |
| Mon. 20.11. | (4) | 2.2 Stochastic Gradient Descent |
| Mon. 27.11. | (5) | 2.3 Newton's Method |
| Mon. 4.12. | (6) | 2.4 Quasi-Newton Methods |
| Mon. 11.12. | (7) | 2.5 Subgradient Methods |
| Mon. 18.12. | — | — *canceled* — |
| | — | — *Christmas Break* — |
| Mon. 8.01. | (8) | 2.6 Coordinate Descent |
| | | **3. Equality Constrained Optimization** |
| Mon. 15.01. | (9) | 3.1 Duality |
| Mon. 22.01. | (10) | 3.2 Methods |
| | | **4. Inequality Constrained Optimization** |
| Mon. 29.01. | (11) | 4.1 Primal Methods |
| Mon. 5.02. | (12) | 4.2 Barrier and Penalty Methods |
| Mon. 12.02. | (13) | 4.3 Cutting Plane Methods |
| | | Q & A |

# Outline

1. Constrained Optimization

2. Duality

3. Karush-Kuhn-Tucker Conditions

# Outline

1. Constrained Optimization

2. Duality

3. Karush-Kuhn-Tucker Conditions

# Constrained Optimization Problems

A **constrained optimization problem** has the form:

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_p(\mathbf{x}) = 0, \quad p = 1, \ldots, P \\ & h_q(\mathbf{x}) \leq 0, \quad q = 1, \ldots, Q \end{aligned}$$

where:

- $f : \mathbb{R}^N \to \mathbb{R}$ is called the **objective** or **cost function**,
- $g_1, \ldots, g_P : \mathbb{R}^N \to \mathbb{R}$ are called **equality constraints**,
- $h_1, \ldots, h_Q : \mathbb{R}^N \to \mathbb{R}$ are called **inequality constraints**,
- a feasible, optimal $\mathbf{x}^*$ exists

# Constrained Optimization Problems

A **convex constrained optimization problem**:

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_p(\mathbf{x}) = 0, \quad p = 1, \ldots, P \\ & h_q(\mathbf{x}) \leq 0, \quad q = 1, \ldots, Q \end{aligned}$$

is **convex** iff:

- $f$, the objective function is **convex**,
- $g_1, \ldots, g_P$ the equality constraint functions are **affine**: $g_p(x) = \mathbf{a}_p^T \mathbf{x} - b_p$, and
- $h_1, \ldots, h_Q$ the inequality constraint functions are **convex**.

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & \mathbf{a}_p^T \mathbf{x} - b_p = 0, \quad p = 1, \ldots, P \\ & h_q(\mathbf{x}) \leq 0, \qquad\quad\; q = 1, \ldots, Q \end{aligned}$$

# Linear Programming

A convex problem with an

- **affine** objective and
- **affine** constraints

is called **Linear Program (LP)**.

**Standard form LP:**

$$\begin{aligned} \text{minimize} \quad & \mathbf{c}^T \mathbf{x} \\ \text{subject to} \quad & \mathbf{a}_p^T \mathbf{x} = b_p, \quad p = 1, \ldots, P \\ & \mathbf{x} \geq 0 \end{aligned}$$

**Inequality form LP:**

$$\begin{aligned} \text{minimize} \quad & \mathbf{c}^T \mathbf{x} \\ \text{subject to} \quad & \mathbf{a}_q^T \mathbf{x} \leq b_q, \quad q = 1, \ldots, Q \end{aligned}$$

# Quadratic Programming

A convex problem with

- a **quadratic** objective and
- **affine** constraints

is called **Quadratic Program (QP)**.

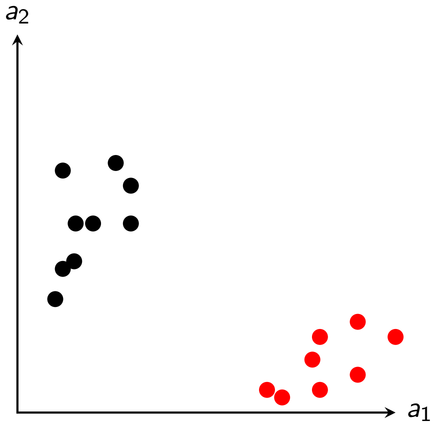
**Inequality form QP:**

$$
\begin{aligned}
\text{minimize} \quad & \frac{1}{2}\mathbf{x}^T C\mathbf{x} + \mathbf{c}^T\mathbf{x} \\
\text{subject to} \quad & \mathbf{a}_q^T\mathbf{x} \leq b_q, \quad q = 1, \ldots, Q
\end{aligned}
$$

where:

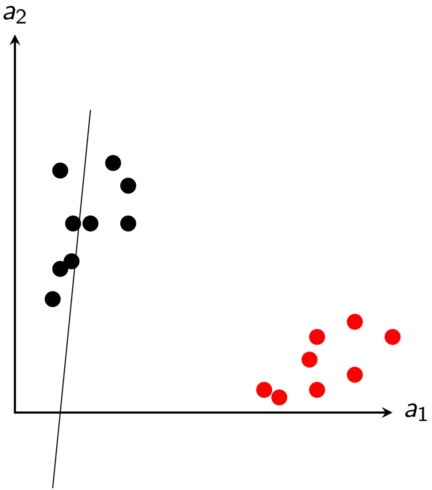
- $C \succ 0$ pos.def. or
- $C = 0$, a special case: linear programs.

# Example: Maximum Margin Separating Hyperplanes



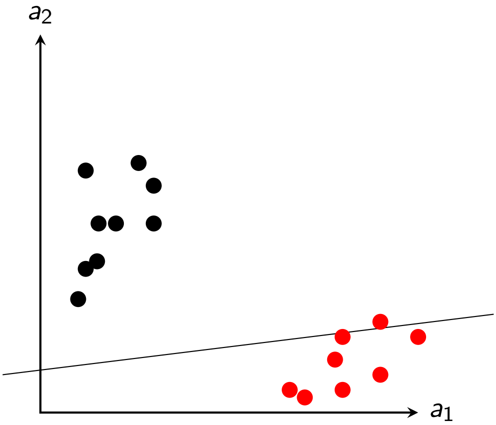

Note: $N$ points $a_n \in \mathbb{R}^M$ ($M = 2$) with labels $b_n \in \{+1, -1\}$ (black/red).
Hyperplane defined by $\mathbf{x}, x_0$: $\mathbf{x}^T \mathbf{a} + x_0 = 0$.

# Example: Maximum Margin Separating Hyperplanes



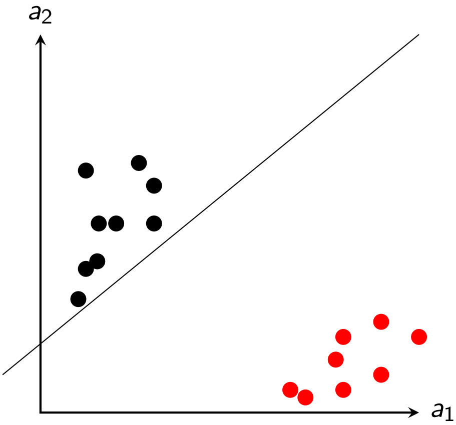

Note: $N$ points $a_n \in \mathbb{R}^M$ ($M = 2$) with labels $b_n \in \{+1, -1\}$ (black/red).
Hyperplane defined by $\mathbf{x}, x_0$: $\mathbf{x}^T \mathbf{a} + x_0 = 0$.

# Example: Maximum Margin Separating Hyperplanes



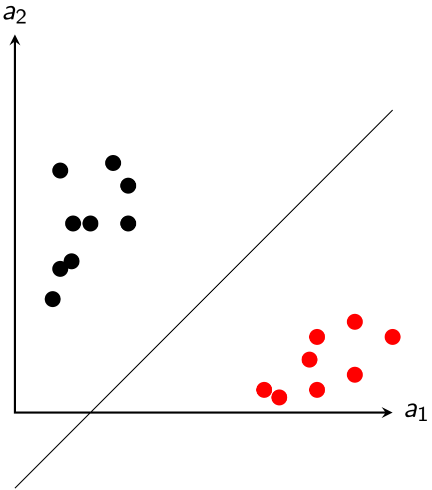

Note: $N$ points $a_n \in \mathbb{R}^M$ ($M = 2$) with labels $b_n \in \{+1, -1\}$ (black/red).
Hyperplane defined by $\mathbf{x}, x_0$: $\mathbf{x}^T \mathbf{a} + x_0 = 0$.

# Example: Maximum Margin Separating Hyperplanes



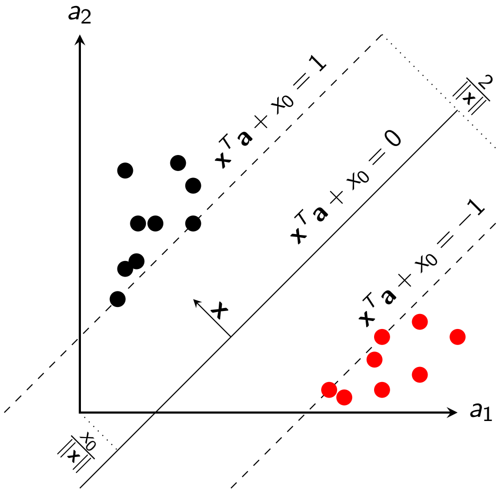

Note: $N$ points $a_n \in \mathbb{R}^M$ ($M = 2$) with labels $b_n \in \{+1, -1\}$ (black/red).
Hyperplane defined by $\mathbf{x}, x_0$: $\mathbf{x}^T \mathbf{a} + x_0 = 0$.

# Example: Maximum Margin Separating Hyperplanes




Note: $N$ points $a_n \in \mathbb{R}^M$ ($M = 2$) with labels $b_n \in \{+1, -1\}$ (black/red).
Hyperplane defined by $\mathbf{x}, x_0$: $\mathbf{x}^T \mathbf{a} + x_0 = 0$.

# Example: Maximum Margin Separating Hyperplanes




Note: $N$ points $a_n \in \mathbb{R}^M$ ($M = 2$) with labels $b_n \in \{+1, -1\}$ (black/red).
Hyperplane defined by $\mathbf{x}, x_0$: $\mathbf{x}^T\mathbf{a} + x_0 = 0$.

# Example: Support Vector Machines

For linear separable problems $(a_n, b_n) \in \mathbb{R}^M \times \{0, 1\}$ $(n = 1, \ldots, N)$:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}||\mathbf{x}||^2 \\ \text{subject to} \quad & b_n(x_0 + \mathbf{x}^T \mathbf{a_n}) \geq 1, \quad n = 1, \ldots, N \\ \text{over} \quad & x \in \mathbb{R}^M, \quad x_0 \in \mathbb{R} \end{aligned}$$

Note: For linear inseparable problems, optimization variables $x = (\beta, \beta_0, \xi)$ are the hyperplane and the slack variables.

# Outline

# Lagrangian

Given a constrained optimization problem in the standard form:

$$\begin{array}{rll} \text{minimize} & f(\mathbf{x}) & \\ \text{subject to} & g_p(\mathbf{x}) = 0, & p = 1, \ldots, P \\ & h_q(\mathbf{x}) \leq 0, & q = 1, \ldots, Q \end{array}$$

We can put

- the objective function $f$ and
- the constraints $g_p$ and $h_q$

in a joint function called **primal Lagrangian**:

$$f(\mathbf{x}) + \sum_{p=1}^{P} \nu_p \, g_p(\mathbf{x}) + \sum_{q=1}^{Q} \lambda_q \, h_q(\mathbf{x})$$

# Primal Lagrangian

The **primal Lagrangian** of a constrained optimization problem is a function

$$L\colon \mathbb{R}^N \times \mathbb{R}^P \times \mathbb{R}^Q \to \mathbb{R}$$

$$L(\mathbf{x}, \nu, \lambda) := f(\mathbf{x}) + \sum_{p=1}^{P} \nu_p \, g_p(\mathbf{x}) + \sum_{q=1}^{Q} \lambda_q \, h_q(\mathbf{x})$$

where:

- $\nu_p$ and $\lambda_q$ are called **Lagrange multipliers**.
  - $\nu_p$ is the Lagrange multiplier associated with the constraint $g_p(\mathbf{x}) = 0$
  - $\lambda_q$ is the Lagrange multiplier associated with the constraint $h_q(\mathbf{x}) \leq 0$.

# Dual Lagrangian

Let $\mathcal{X}$ be the domain of the problem.

**Dual Lagrangian** of a constrained optimization problem:

$$g : \mathbb{R}^P \times \mathbb{R}^Q \rightarrow \mathbb{R}$$

$$\begin{aligned} g(\nu, \lambda) &:= \inf_{\mathbf{x} \in \mathcal{X}} L(\mathbf{x}, \nu, \lambda) \\ &= \inf_{\mathbf{x} \in \mathcal{X}} \left( f(\mathbf{x}) + \sum_{p=1}^{P} \nu_p \, g_p(\mathbf{x}) + \sum_{q=1}^{Q} \lambda_q \, h_q(\mathbf{x}) \right) \end{aligned}$$

- Q: What type of function is $g$?

Note: From here onwards, $g$ denotes the dual Lagrangian, not the equality constraints anymore.

# Dual Lagrangian

Let $\mathcal{X}$ be the domain of the problem.
**Dual Lagrangian** of a constrained optimization problem:

$$\begin{aligned} g : \mathbb{R}^P \times \mathbb{R}^Q &\to \mathbb{R} \\ g(\nu, \lambda) &:= \inf_{\mathbf{x} \in \mathcal{X}} L(\mathbf{x}, \nu, \lambda) \\ &= \inf_{\mathbf{x} \in \mathcal{X}} \left( f(\mathbf{x}) + \sum_{p=1}^{P} \nu_p \, g_p(\mathbf{x}) + \sum_{q=1}^{Q} \lambda_q \, h_q(\mathbf{x}) \right) \end{aligned}$$

- $g$ is concave.
  - as infimum over concave (affine) functions
- for non-negative $\lambda_q$, $g$ is a **lower bound** on $f(\mathbf{x}^*)$:

$$g(\nu, \lambda) \leq f(\mathbf{x}^*) \quad \text{for } \lambda \geq 0$$

Note: From here onwards, $g$ denotes the dual Lagrangian, not the equality constraints anymore.

# Dual Lagrangian / Proof

Proof of the lower bound property of:

$$
\begin{aligned}
g(\nu, \lambda) &:= \inf_{\mathbf{x} \in \mathcal{X}} L(\mathbf{x}, \nu, \lambda) \\
&= \inf_{\mathbf{x} \in \mathcal{X}} \left( f(\mathbf{x}) + \sum_{p=1}^{P} \nu_p \, g_p(\mathbf{x}) + \sum_{q=1}^{Q} \lambda_q \, h_q(\mathbf{x}) \right) \\
&\underset{\text{inf}}{\leq} f(\mathbf{x}^*) + \sum_{p=1}^{P} \nu_p \, \underbrace{g_p(\mathbf{x}^*)}_{=0} + \sum_{q=1}^{Q} \underbrace{\lambda_q}_{\geq 0} \, \underbrace{h_q(\mathbf{x}^*)}_{\leq 0} \\
&\leq f(\mathbf{x}^*)
\end{aligned}
$$

# Example / Least-norm solution of linear equations

$$\begin{array}{ll} \text{minimize} & \mathbf{x}^T\mathbf{x} \\ \text{subject to} & A\mathbf{x} = \mathbf{b} \end{array}$$

- **Lagrangian:** $L(\mathbf{x}, \nu) = \mathbf{x}^T\mathbf{x} + \nu^T(A\mathbf{x} - \mathbf{b})$
- **Dual Lagrangian:**
  - minimize $L$ over $\mathbf{x}$:

$$\nabla_{\mathbf{x}} L(\mathbf{x}, \nu) = 2\mathbf{x} + A^T\nu = 0$$
$$\mathbf{x} = -\frac{1}{2}A^T\nu$$

  - Substituting $\mathbf{x}$ in $L$ we get $g$:

$$\begin{aligned} g(\nu) &= \frac{1}{4}\nu^T AA^T\nu - \frac{1}{2}\nu^T AA^T\nu - \nu^T b \\ &= -\frac{1}{4}\nu^T AA^T\nu - b^T\nu \end{aligned}$$

# The Dual Problem

Compute the **best** lower bound on $f(\mathbf{x}^*)$:

$$\begin{aligned} \text{maximize} \quad & g(\nu, \lambda) \\ \text{subject to} \quad & \lambda \geq 0 \end{aligned}$$

- this is a convex optimization problem (g is concave).

# Weak and Strong Duality

Say $p^*$ is the optimal value of $f$
and $d^*$ is the optimal value of $g$.

**Weak duality**: $d^* \leq p^*$

- always holds.
- can be useful to find informative lower bounds for difficult problems.

**Strong duality**: $d^* = p^*$

- does not always hold.
- but holds for a range of convex problems.
- properties that guarantee strong duality are called **constraint qualifications**.

# Weak Duality / Example

- convex optimization problem:

$$\begin{aligned} \text{min. } & f(x_1, x_2) := e^{-x_1} \\ \text{s.t. } & h(x_1, x_2) := \frac{x_1^2}{x_2} \leq 0 \\ \text{over } & (x_1, x_2) \in \mathcal{X} := \mathbb{R} \times \mathbb{R}^+ \end{aligned}$$

- $h$ is a complicated way to say: $x_1 = 0$,
  and thus $x^* = (0, x_2)$ for any $x_2 > 0$ and $p^* = 1$.

- dual Lagrange function:

$$\begin{aligned} g(\lambda) &:= \inf_{(x_1,x_2)\in\mathcal{X}} e^{-x_1} + \lambda \frac{x_1^2}{x_2} \\ &= 0, \quad \text{e.g., as limit of } x^{(n)} := (n, n^3), \quad n \in \mathbb{N} \\ \rightsquigarrow \quad d^* &:= \sup_{\lambda \in \mathbb{R}_0^+} g(\lambda) = 0 < p^* = 1 \end{aligned}$$

# Strong Duality of Equality Constraints Only

Convex problems with only equality constraints

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & A\mathbf{x} = \mathbf{b} \end{aligned}$$

are always strongly dual.

Proof:

$$\begin{aligned} g(\nu^*) = \sup_\nu \inf_x f(x) + \nu^T(Ax - b) &\underset{(*)}{=} \sup_\nu \inf_{x:Ax-b=0} f(x) + \nu^T 0 \\ &= \inf_{x:Ax-b=0} f(x) \\ &= f(x^*) \end{aligned}$$

where $(*)$ holds as:

- assume $g(\nu^*) = f(x') + \nu^*(Ax' - b)$ with $Ax' - b \neq 0$.
- then $\sup_\nu \nu^T(Ax' - b) = \infty$. Contradiction to $g(\nu^*)$ being finite.

# Slater's Condition / Strict Feasibility

If a convex problem

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & A\mathbf{x} = \mathbf{b} \\ & h_q(\mathbf{x}) \leq 0, \quad q = 1, \ldots, Q \end{aligned}$$

is **strictly feasible**, i.e.

$$\exists \mathbf{x} : \quad A\mathbf{x} = \mathbf{b} \quad \text{and} \quad h_q(\mathbf{x}) < 0, \quad \forall q = 1, \ldots, Q$$

then strong duality holds for this problem.

# Duality Gap

How close is the value of the dual lagrangian to the primal objective?

For a primal feasible $\mathbf{x}$ (i.e., $Ax = b$ and $h(x) \leq 0$) and
dual feasible $\nu, \lambda$ (i.e., $\lambda \geq 0$),
the **duality gap** is defined as:

$$f(\mathbf{x}) - g(\nu, \lambda)$$

Since $g(\nu, \lambda)$ is a lower bound on $f$:

$$f(\mathbf{x}) - f(\mathbf{x}^*) \leq f(\mathbf{x}) - g(\nu, \lambda)$$

If the duality gap is zero, then $\mathbf{x}$ is primal optimal.

- This is a useful stopping criterion:
  if $f(\mathbf{x}) - g(\nu, \lambda) \leq \epsilon$, then we are sure that $f(\mathbf{x}) - f(\mathbf{x}^*) \leq \epsilon$

# Outline

# Consequences of Optimality under Strong Duality

Assume strong duality:

- let $\mathbf{x}^*$ be primal optimal and
- $(\nu^*, \lambda^*)$ be dual optimal.

$$
\begin{aligned}
f(\mathbf{x}^*) \underset{\text{s.d.}}{=} g(\nu^*, \lambda^*) = \inf_{\mathbf{x} \in \mathcal{X}} & L(\mathbf{x}, \nu^*, \lambda^*) \\
\leq & L(\mathbf{x}^*, \nu^*, \lambda^*) \\
\underset{\text{lower bound}}{\leq} & f(\mathbf{x}^*)
\end{aligned}
$$

hence

$$
L(\mathbf{x}^*, \nu^*, \lambda^*) = \inf_{\mathbf{x} \in \mathcal{X}} L(\mathbf{x}, \nu^*, \lambda^*) = f(\mathbf{x}^*)
$$

# Consequences of Optimality under Strong Duality I: Stationarity

Assume strong duality:

- let $\mathbf{x}^*$ be primal optimal and
- $(\nu^*, \lambda^*)$ be dual optimal.

$$L(\mathbf{x}^*, \nu^*, \lambda^*) = \inf_{\mathbf{x} \in \mathcal{X}} L(\mathbf{x}, \nu^*, \lambda^*)$$

i.e., $\mathbf{x}^*$ minimizes $L(\mathbf{x}, \nu^*, \lambda^*)$ and thus

$$\nabla_{\mathbf{x}} L(\mathbf{x}^*, \nu^*, \lambda^*) = \nabla f(\mathbf{x}^*) + \sum_{p=1}^{P} \nu_p^* \nabla g_p(\mathbf{x}^*) + \sum_{q=1}^{Q} \lambda_q^* \nabla h_q(\mathbf{x}^*) \stackrel{!}{=} 0$$

- condition called **stationarity**.

Note: $g_p$ denote again the equality constraints, not the dual Lagrangian.

# Consequences of Optimality under Strong Duality II: Complementary Slackness

Assume strong duality:

- let $\mathbf{x}^*$ be primal optimal and
- $(\nu^*, \lambda^*)$ be dual optimal.

$$L(\mathbf{x}^*, \nu^*, \lambda^*) = f(\mathbf{x}^*) + \sum_{p=1}^{P} \nu_p^* \, g_p(\mathbf{x}^*) + \sum_{q=1}^{Q} \lambda_q^* \, h_q(\mathbf{x}^*) = f(\mathbf{x}^*)$$

$\rightsquigarrow$ **complementary slackness**:

$$\lambda_q^* \, h_q(\mathbf{x}^*) = 0, \quad q = 1, \ldots, Q$$

which means that

- If $\lambda_q^* > 0$, then $h_q(\mathbf{x}^*) = 0$
- If $h_q(\mathbf{x}^*) < 0$, then $\lambda_q = 0$

# Karush-Kuhn-Tucker (KKT) Conditions

The following conditions on $\mathbf{x}, \nu, \lambda$ are called the KKT conditions:

1. **primal feasibility:** $g_p(\mathbf{x}) = 0$ and $h_q(\mathbf{x}) \leq 0, \quad \forall p, q$
2. **dual feasibility:** $\lambda \geq 0$
3. **complementary slackness:** $\lambda_q \, h_q(\mathbf{x}) = 0, \quad \forall q$
4. **stationarity:** $\nabla f(\mathbf{x}) + \sum_{p=1}^{P} \nu_p \nabla g_p(\mathbf{x}) + \sum_{q=1}^{Q} \lambda_q \nabla h_q(\mathbf{x}) = 0$

If strong duality holds and $\mathbf{x}, \nu, \lambda$ are optimal,
then they **must** satisfy the KKT conditions.

**If $\mathbf{x}, \lambda, \nu$ satisfy the KKT conditions,**
**then $\mathbf{x}$ is primal optimal and $(\nu, \lambda)$ is dual optimal.**

# Karush-Kuhn-Tucker (KKT) Conditions

Theorem (Karush-Kuhn-Tucker)

*For a strongly dual problem, if* $\mathbf{x}, \lambda, \nu$ *satisfy the KKT conditions,*

1. **primal feasibility:** $g_p(\mathbf{x}) = 0$ *and* $h_q(\mathbf{x}) \leq 0, \quad \forall p, q$
2. **dual feasibility:** $\lambda \geq 0$
3. **complementary slackness:** $\lambda_q\, h_q(\mathbf{x}) = 0, \quad \forall q$
4. **stationarity:** $\nabla f(\mathbf{x}) + \sum_{p=1}^{P} \nu_p \nabla g_p(\mathbf{x}) + \sum_{q=1}^{Q} \lambda_q \nabla h_q(\mathbf{x}) = 0$

*then* $\mathbf{x}$ *is the primal solution and* $(\nu, \lambda)$ *is the dual solution.*

# Karush-Kuhn-Tucker (KKT) Conditions / Proof

Proof:

$$
\begin{aligned}
g(\lambda, \nu) &= \inf_{x' \in \mathcal{X}} f(x') + \sum_{p=1}^{P} \nu_p g_p(x') + \sum_{q=1}^{Q} \lambda_q h_q(x') \\
&\underset{\text{4. stat.}}{=} f(x) + \sum_{p=1}^{P} \nu_p g_p(x) + \sum_{q=1}^{Q} \lambda_q h_q(x) \\
&\underset{1,3}{=} f(x)
\end{aligned}
$$

i.e. duality gap is 0, and thus $x$ and $\lambda, \nu$ optimal.

# Summary

- The **primal Lagrangian** combines objective and constraints linearly
  - constraint weights called **multipliers**
  - multipliers viewed as additional variables
  - inequality multipliers $\geq 0$
- The **dual Lagrangian** $g$ is the pointwise infimum of the primal Lagrangian over the primal variables $\mathbf{x}$.
  - a lower-bound for $f(\mathbf{x}^*)$
  - difference $f(x) - g(\nu, \lambda)$ called **duality gap**
- **Dual problem**: Maximizing the dual Lagrangian
  - $=$ finding the best lower bound
  - a convex problem
  - solves the primal problem under **strong duality** (duality gap $= 0$)
- **Constraint qualifications** guarantee strong duality for a problem
  - e.g., **Slater's condition**: existence of a **strictly feasible** point.

# Summary (2/2)

- **Karush-Kuhn-Tucker (KKT) conditions** for $(x, \nu, \lambda)$:
  1. **primal feasibility**
  2. **dual feasibility**
  3. **complementary slackness**
  4. **stationarity**

- KKT is a necessary condition for primal/dual optimality under strong duality.

- If a problem is strongly dual,
  KKT are also a sufficient condition for a primal/dual solution.

# Further Readings

- Boyd and Vandenberghe, 2004, ch. 5
- The proof that Slater's condition is sufficient for strong duality can be found in Boyd and Vandenberghe, 2004, ch. 5.3.2.

# References

Boyd, Stephen and Lieven Vandenberghe (2004). *Convex Optimization*. Cambridge University Press.