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. 28.10.	(0)	0. Overview
		1. Theory
Mon. 4.11.	(1)	1. Convex Sets and Functions
		2. Unconstrained Optimization
Mon. 11.11.	(2)	2.1 Gradient Descent
Mon. 18.11.	(3)	2.2 Stochastic Gradient Descent
Mon. 25.11.	(4)	2.3 Newton's Method
Mon. 2.12.	(5)	2.4 Quasi-Newton Methods
Mon. 19.12.	(6)	2.5 Subgradient Methods
Mon. 16.12.	(7)	2.6 Coordinate Descent
	—	— *Christmas Break* —
		3. Equality Constrained Optimization
Mon. 6.1.	(8)	3.1 Duality
Mon. 13.1.	(9)	3.2 Methods
		4. Inequality Constrained Optimization
Mon. 20.1.	(10)	4.1 Primal Methods
Mon. 27.1.	(11)	4.2 Barrier and Penalty Methods
Mon. 3.2.	(12)	4.3 Cutting Plane Methods

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:

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

Constrained Optimization Problems

A **convex constrained optimization problem**:

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

is **convex** iff:

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

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

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:

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

- No analytical solution
- There are specialized algorithms available

Quadratic Programming

A convex problem with

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

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

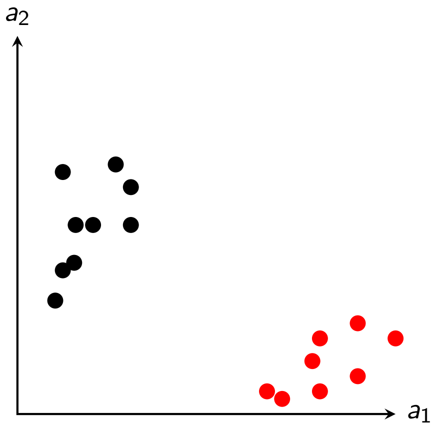
Inequality form QP:

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

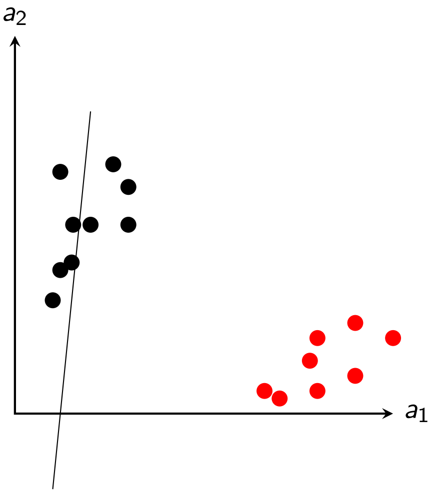
where:

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

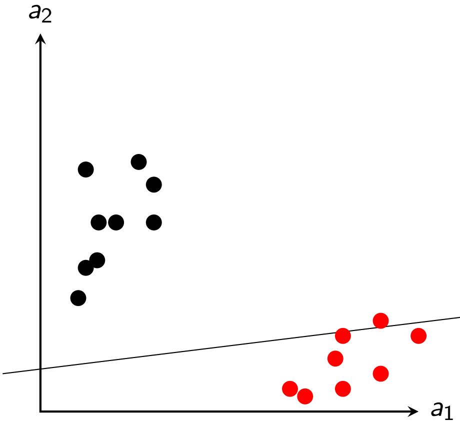
Example: Maximum Margin Separating Hyperplanes



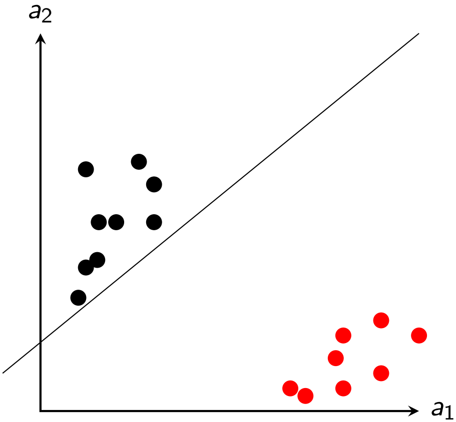
Example: Maximum Margin Separating Hyperplanes



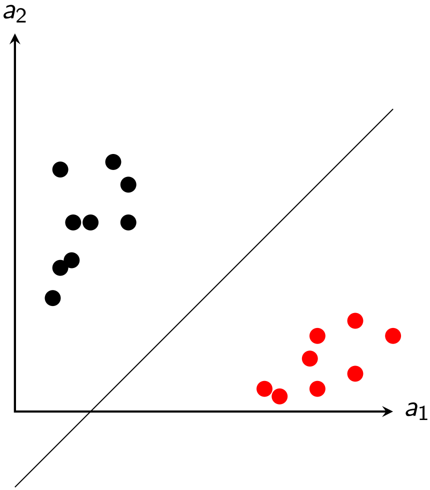
Example: Maximum Margin Separating Hyperplanes



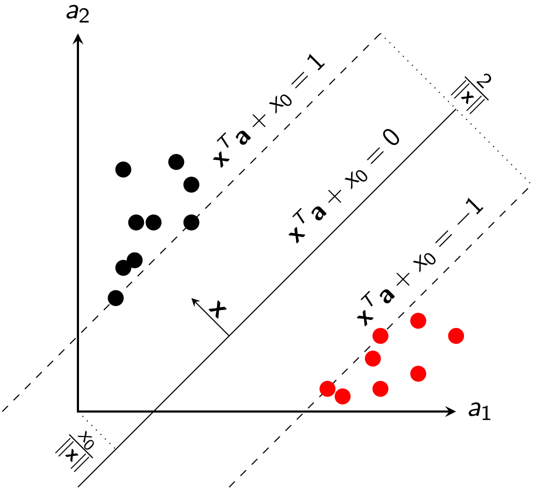
Example: Maximum Margin Separating Hyperplanes



Example: Maximum Margin Separating Hyperplanes



Example: Maximum Margin Separating Hyperplanes



Example: Support Vector Machines

If the instances are not completely separable,
we can allow some of them to be on the wrong side of the decision boundary.

The closer the “wrong” points are to the boundary,
the better (modeled by slack variables ξ_n).

$$
\begin{aligned}
\text{minimize} \quad & \frac{1}{2}||\mathbf{x}||^2 + \gamma \sum_{n=1}^{N} \xi_n \\
\text{subject to} \quad & y_n(a_0 + \mathbf{x}^T \mathbf{a_n}) \geq 1 - \xi_n, \quad n = 1, \ldots, N \\
& \xi_n \geq 0, , \quad n = 1, \ldots, N
\end{aligned}
$$

Outline

1. Constrained Optimization

2. Duality

3. Karush-Kuhn-Tucker Conditions

Lagrangian

Given a constrained optimization problem in the standard form:

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

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

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

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

where:

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

Dual Lagrangian

Be $\mathcal{D}$ the domain of the problem, the **dual Lagrangian** of a constrained optimization problem is a function $g : \mathbb{R}^P \times \mathbb{R}^Q \rightarrow \mathbb{R}$:

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

- g is concave.
  - as infimum over concave (affine) functions
- for non-negative λ_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:

$$g(\nu, \lambda) := \inf_{\mathbf{x} \in \mathcal{D}} L(\mathbf{x}, \nu, \lambda)$$

$$= \inf_{\mathbf{x} \in \mathcal{D}} \left(f(\mathbf{x}) + \sum_{p=1}^{P} \nu_p \, g_p(\mathbf{x}) + \sum_{q=1}^{Q} \lambda_q \, h_q(\mathbf{x}) \right)$$

for any feasible $\mathbf{x}$ we have:

- $g_p(\mathbf{x}) = 0$
- $h_q(\mathbf{x}) \leq 0$

thus, with $\lambda \geq 0$:

$$f(\mathbf{x}) \geq L(\mathbf{x}, \nu, \lambda) \geq \inf_{\mathbf{x}' \in \mathcal{D}} L(\mathbf{x}', \nu, \lambda) = g(\nu, \lambda)$$

minimizing over all feasible $\mathbf{x}$, we have $f(\mathbf{x}^*) \geq g(\nu, \lambda)$

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:

$$g(\nu) = -\frac{1}{4}\nu^T AA^T\nu - b^T\nu$$

The Dual Problem

Once we know how to compute the dual, we are interested in computing the **best** lower bound on $f(\mathbf{x}^*)$:

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

where:

- this is a convex optimization problem (g is concave)
- let d^* be the maximal value of g

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**

Slater's Condition / Strict Feasibility

If the following primal 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:

- convex and
- **strictly feasible**, i.e.

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

then strong duality holds for this problem.

Duality Gap

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

Given a primal feasible $\mathbf{x}$ and a dual feasible ν, λ,
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 Strong Duality

Assume strong duality:

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

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

hence

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

Consequences of Strong Duality I: Stationarity

Assume strong duality:

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

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

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

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

- condition called **stationarity**.

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

Consequences of Strong Duality II: Complementary Slackness

Assume strong duality:

- let $\mathbf{x}^*$ be primal optimal and
- (ν^*, λ^*) 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 the primal solution and (ν, λ) is the dual solution.

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* (ν, λ) *is the dual solution.*

Karush-Kuhn-Tucker (KKT) Conditions / Proof

Proof:

$$g(\lambda, \nu) = \sup_{x' \in \mathcal{D}} f(x') + \sum_{p=1}^{P} \nu_p g_p(x') + \sum_{q=1}^{Q} \lambda_q h_q(x')$$

$$\underset{4.stat.}{=} f(x) + \sum_{p=1}^{P} \nu_p g_p(x) + \sum_{q=1}^{Q} \lambda_q h_q(x)$$

$$= f(x)$$

i.e. duality gap is 0, and thus x and λ, ν optimal.

Summary

- The **primal Lagrangian** combines objective and constraints linearly
  - constraint weights called **multipliers**
  - multipliers viewed as additional variables
  - inequality multipliers ≥ 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 $g(\nu, \lambda) - f(x^*)$ 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, ν, λ):
  1. **primal feasibility**
  2. **dual feasibility**
  3. **complementary slackness**
  4. **stationarity**

- KKT is a necessary condition for a primal/dual solution.

- 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

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