# Optimization (2MMD10,2DME20), lecture 3

Gerhard Woeginger

Technische Universiteit Eindhoven

Fall 2015, Q1

# Program for this week

- Recall: unconstrained optimization in $\mathbb{R}^n$
- The Lagrangian approach
- Karush-Kuhn-Tucker conditions
- Lagrangian duality
- Lagrangian duality for convex programs
- A catalogue of duality theorems
- Solving convex programs

# Recall, recall, recall (1)

**Definition**

In a continuous optimization problem, we want to solve

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in U \end{array}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is the objective function and
where $U \subseteq \mathbb{R}^n$ is the feasible region.

- "minimize $-f(x)$" equivalent to "maximize $f(x)$"
- $U = \mathbb{R}^n$: unconstrained optimization problem

# Recall, recall, recall (2)

- Compact subset of $\mathbb{R}^n$: closed and bounded

**Theorem (Weierstrass)**

If function $f$ is continuous and if the feasible region $U$ is compact, then $f$ attains its minimum (and its maximum) in $U$.

**Theorem (Weierstrass)**

If continuous function $f : \mathbb{R}^n \to \mathbb{R}$ goes to +infinity whenever $\|x\| \to \infty$, then $f$ attains its minimum.

# Unconstrained optimization (1)

Recall from the univariate case ($f : \mathbb{R} \to \mathbb{R}$):

- $\bar{x}$ minimizes $f \implies f'(\bar{x}) = 0$
- $f''(\bar{x}) > 0$ or $f''(\bar{x}) < 0$ or $f''(\bar{x}) = 0$

**Definition**

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable.
A point $\bar{x}$ with $\nabla f(\bar{x}) = 0$ is called stationary point.

**Necessary optimality condition (first order)**

If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable and if $\bar{x} \in \mathbb{R}$ is a local minimum, then $\bar{x}$ is a stationary point.

- Often called "Fermat's theorem"
- Condition is necessary, but not sufficient ($f(x) = x^3$)

# Unconstrained optimization (2)

**Necessary optimality condition (second order)**

If $f$ is twice continuously differentiable and if $\bar{x}$ is a local minimum, then

(i) $\bar{x}$ is a stationary point, and

(ii) the Hessian $\nabla^2 f(\bar{x})$ is positive semi-definite.

- Condition is necessary, but not sufficient ($f(x) = x^3$)

**Sufficient optimality condition (second order)**

If $f$ is twice continuously differentiable and if $\bar{x}$ is a local minimum, then

(i) $\bar{x}$ is a stationary point, and

(ii) the Hessian $\nabla^2 f(\bar{x})$ is positive definite.

- Condition is sufficient, but not necessary ($f(x) = x^4$)

**Example**

$f(x, y) = x^2 + xy + y^2 - 2x - y + 3$

# Unconstrained optimization (3)

**Iterative algorithms**

An iterative algorithm for the optimization problem $p^* = \min_x f(x)$ generates a sequence $x^{(0)}, x^{(1)}, \ldots$ of points,
so that $f(x^{(k)}) \to p^*$ as $k \to \infty$.

In the $k$-th step, a typical iterative algorithm

- chooses a search direction $\Delta x^{(k)} \in \mathbb{R}^n$
- chooses a *step size* $t^{(k)} \in \mathbb{R}$
- puts $x^{(k+1)} = x^{(k)} + t^{(k)} \Delta x^{(k)}$

The search direction $\Delta x$ is a descent direction, if $(\nabla f(x))^T \Delta x < 0$.
Then there exists a $t$ so that $f(x + t\Delta x) < f(x)$.

- gradient descent; steepest descent; Newton descent; line search
- Michiel Hochstenbach will talk on this in weeks 6 and 7
- methods work well (and fast) for minimization of convex functions

# Constrained optimization (1)

## Joseph-Louis Lagrange (1788)

"One can state the following general principle:
If one is looking for the maximum or minimum of some function of many variables subject to the condition that these variables are related by a constraint given by one or more equations,

then one should add to the function whose extremum is sought the functions that yield the constraint equations each multiplied by undetermined multipliers and seek the maximum or minimum of the resulting sum as if the variables were independent.

The resulting equations, combined with the constraint equations, will serve to determine all unknowns."

# Constrained optimization (2)

## Continuous optimization problem

$$
\begin{array}{lll}
\text{minimize} & f_0(x) & \\
\text{subject to} & f_i(x) \le 0 & i = 1, \ldots, r \\
& h_i(x) = 0 & i = 1, \ldots, s \\
& x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n &
\end{array}
$$

- an equality constraint can be rewritten into two inequality constraints
- we assume throughout that all $f_i$ and all $h_i$ are differentiable

## Definition (technical)

For a point $x \in \mathbb{R}^n$, let $J(x) = \{i \ge 1 \mid f_i(x) = 0\}$.
Point $x$ is called regular,
if the set of all vectors $\nabla f_i(x)$ with $i \in J(x)$ and
of all vectors $\nabla h_i(x)$ with $i = 1, \ldots, s$ is linearly independent.

# Constrained optimization (3)

**Definition**

The Lagrangian of this problem is:

$$L(x,\lambda,\mu) \;:=\; f_0(x) + \sum_{i=1}^{r} \lambda_i f_i(x) + \sum_{i=1}^{s} \mu_i h_i(x)$$

where $\lambda \geq 0$ and $\mu$ arbitrary.

**Lemma**

*If* $\lambda \geq 0$*, then*

$$f_0(x) \geq L(x,\lambda,\mu)$$

*for any* $x$ *such that all* $f_i(x) \leq 0$ *and* $h_i(x) = 0$*.*

# Constrained optimization (4)

**Karush-Kuhn-Tucker (KKT) conditions**

If a feasible point $\bar{x}$ is a local optimum,
then there exist real numbers $\lambda_0, \lambda_1, \ldots, \lambda_r$ and $\mu_1, \ldots, \mu_s$
(that not all are 0) such that

- $$\lambda_0 \nabla f(\bar{x}) + \sum_{i=1}^{r} \lambda_i \nabla f_i(\bar{x}) + \sum_{i=1}^{s} \mu_i \nabla h_i(\bar{x}) \;=\; 0$$
- $\lambda_i f_i(\bar{x}) = 0$ for $i = 1, \ldots, r$
- $\lambda_i \geq 0$ for $i = 1, \ldots, r$

- $\lambda_0, \ldots, \lambda_r$ and $\mu_1, \ldots, \mu_s$ are called Lagrange multipliers
- Often convenient: separate treatment of non-negativity constraints
- Often convenient: separate treatment of entire boundary

If point $\bar{x}$ is regular, then one may choose $\lambda_0 = 1$

# Constrained optimization (5)

**Example**

Let $a, c \in \mathbb{R}^n$ with $c \neq 0$.

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} c_i x_i \\ \text{subject to} & \sum_{i=1}^{n} (x_i - a_i)^2 \leq 1 \end{array}$$

- KKT yields $c_i + 2\lambda(x_i - a_i) = 0$ for $i = 1, \ldots, n$
  and $\lambda \left( \sum_{i=1}^{n} (x_i - a_i)^2 - 1 \right) = 0$
  and $\lambda \geq 0$
- This leads to $x_i = a_i - c_i / \|c\|$
  and $c^T x = c^T a - \|c\|$

# Constrained optimization (6)

Where does the name KKT come from?

- For a long time known as KT (Kuhn-Tucker) conditions
- First appeared in publication by Kuhn and Tucker in 1951
- Later people found out that Karush had the condition in his master thesis of 1939 (Univ. of Chicago)

- William Karush (1917–1997): American mathematician
- Harold Kuhn (1925–2014): famous American mathematician; inventor of the Hungarian method for the assignment problem
- Albert Tucker (1905–1995): famous American mathematician

# Lagrangian duality (1)

$$\begin{array}{lll} \text{minimize} & f_0(x) & \\ \text{subject to} & f_i(x) \leq 0 & i = 1, \ldots, r \\ & h_i(x) = 0 & i = 1, \ldots, s \end{array}$$

## Definition

The Lagrangian $L(x, \lambda, \mu) \; = \; f_0(x) + \sum_{i=1}^{r} \lambda_i f_i(x) + \sum_{i=1}^{s} \mu_i h_i(x)$

## Lemma

*If $\lambda \geq 0$, then*

$$f_0(x) \geq L(x, \lambda, \mu)$$

*for any $x$ such that all $f_i(x) \leq 0$ and $h_i(x) = 0$.*

# Lagrangian duality (2)

Let $p^* :=$

$$\begin{array}{lll} \text{minimize} & f_0(x) & \\ \text{subject to} & f_i(x) \leq 0 & i = 1, \ldots, r \\ & h_i(x) = 0 & i = 1, \ldots, s \end{array}$$

The Lagrange dual function of this problem is:

$$g(\lambda, \mu) := \min_x L(x, \lambda, \mu)$$

**Lemma**

*If $\lambda \geq 0$, then*

$$p^* \geq g(\lambda, \mu)$$

# Lagrangian duality (3)

Consider the (primal) problem

$$\begin{array}{lll} \text{minimize} & f_0(x) & \\ \text{subject to} & f_i(x) \leq 0 & i = 1, \ldots, r \\ & h_i(x) = 0 & i = 1, \ldots, s \end{array}$$

with value $p^*$.

Its Lagrange dual is the problem

$$\begin{array}{lll} \text{maximize} & g(\lambda, \mu) & \\ \text{subject to} & \lambda_i \geq 0 & i = 1, \ldots, r \end{array}$$

with value $d^*$.

**Lemma**

*We have $p^* \geq d^*$.*

# Lagrangian duality (4a)

**Example**

For $A \in S^n$, show that $\min\{x^T Ax \mid x^T x = 1\}$ yields the smallest Eigenvalue $\lambda_{\min}(A)$ of matrix $A$.

$$L(x,\mu) \;=\; x^T Ax + \mu(1 - x^T x) \;=\; \mu + x^T(A - \mu I)x$$

$$\nabla x^T Ax + \mu\nabla(1 - x^T x) \;=\; 0$$

$$2x^T A - 2\mu x^T \;=\; 0$$

- If $x^T A = \mu x^T$ then
  either: $x = 0$
  or: $\mu$ is an Eigenvalue and $x$ is the corresponding Eigenvector

## Example (continued)

For $A \in S^n$, give the Lagrangian, Lagrange dual function, and Lagrange dual of the problem $\min\{x^T Ax \mid x^T x = 1\}$.

$$L(x,\mu) = x^T Ax + \mu(1 - x^T x) = \mu + x^T(A - \mu I)x$$

$$g(\mu) = \min_x \ \mu + x^T(A - \mu I)x$$

- If $A - \mu I \in PSD$, then $\min_x \ x^T(A - \mu I)x = 0$
  If $A - \mu I \notin PSD$, then $\min_x \ x^T(A - \mu I)x = -\infty$
- $A - \mu I \in PSD$ if and only if $\mu \leq \lambda_{\min}(A)$
- Hence $g(\mu) = \mu$ for $\mu \leq \lambda_{\min}$
  and $g(\mu) = -\infty$ for $\mu > \lambda_{\min}$
- Lagrange dual: maximize $\mu$ subject to $\mu \leq \lambda_{\min}$
- Hence in this case $p^* = d^*$ (no duality gap in this case)

# Lagrangian duality (5)

### Example

Determine the optimal primal and the optimal dual objective value for the problem $\min\{x \mid x^3 - y \geq 0,\ y \geq 0\}$.

$$L(x, y, \alpha, \beta) \quad = \quad x + \alpha(y - x^3) - \beta y$$

### Primal problem (P)

$p^* = 0$ with $x^* = y^* = 0$.

### Dual problem (D)

$g(\alpha, \beta) \ = \ \min_{x,y}\ (x - \alpha x^3) + (\alpha - \beta)y$

- If $\alpha > 0$, then $x - \alpha x^3$ goes to $-\infty$ as $x$ goes to $\infty$
- If $\alpha = 0$, then $x - \alpha x^3 = x$ can be made arbitrarily small

As $p^* = 0$ and $d^* = -\infty$, the duality gap is infinite in this case.

# Convex Lagrangian duality (1)

Let $p^* :=$

$$\begin{array}{lll} \text{minimize} & f_0(x) & \\ \text{subject to} & f_i(x) \leq 0 & i = 1, \ldots, r \\ & h_i(x) = 0 & i = 1, \ldots, s \end{array}$$

and $d^* :=$

$$\begin{array}{lll} \text{maximize} & g(\lambda, \mu) & \\ \text{subject to} & \lambda_i \geq 0 & i = 1, \ldots, r \end{array}$$

**Theorem (strong duality for convex optimization)**

Suppose (convex program) and (Slater's condition is satisfied):

- $f_0, \ldots, f_r$ are convex and $h_1, \ldots, h_s$ are affine
- $\exists y : \; f_i(y) < 0$ for $i = 1, \ldots, r$, and
  $h_i(y) = 0$ for $i = 1, \ldots, s$

Then $p^* = d^*$.

# Convex Lagrangian duality (2)

Assume $s = 0$.

**Sketch of proof (strong duality for convex optimization)**

Consider $\mathcal{A} := \{ \begin{bmatrix} u \\ t \end{bmatrix} \mid \exists x : f_i(x) \leq u_i,\ i = 1, \ldots, r,\ f_0(x) \leq t \}$.

- $\mathcal{A}$ is a convex set
- $p^* = \min\{ t \mid \begin{bmatrix} 0 \\ t \end{bmatrix} \in \mathcal{A} \}$
- some hyperplane supports $\mathcal{A}$ at $(0, p^*)$; say

$$\begin{bmatrix} \lambda^* \\ \mu^* \end{bmatrix}^T \begin{bmatrix} u \\ t \end{bmatrix} \geq \alpha \text{ for } \begin{bmatrix} u \\ t \end{bmatrix} \in \mathcal{A}; \qquad \begin{bmatrix} \lambda^* \\ \mu^* \end{bmatrix}^T \begin{bmatrix} 0 \\ p^* \end{bmatrix} = \alpha$$

- $\mu^* \geq 0, \lambda^* \geq 0$
- If $\mu^* = 0$, then $0 > \sum_i \lambda_i^* f_i(y) \geq \alpha = 0$; contradiction.
- If $\mu^* > 0$, then $g(\lambda^* / \mu^*) = p^*$.

# Convex Lagrangian duality (3)

Equivalent formulation of strong duality for convex optimization:

**Theorem**

The KKT conditions characterize an optimal solution (and hence are both necessary and sufficient),

- if $f_0, \ldots, f_r$ are convex and $h_1, \ldots, h_s$ are affine; and
- if Slater's condition is satisfied: there exists a point that satisfies all inequality constraints strictly.

# Convex Lagrangian duality (4)

Another equivalent formulation of strong duality for convex optimization:

**Theorem**

Assume that a convex optimization problem satisfies Slater's condition (there exists a point that satisfies all inequality constraints strictly).

Then vector $x^*$ is an optimal solution,
if and only if there exists $(\lambda^*, \mu^*)$ with $\lambda^* \geq 0$ such that the inequality

$$L(x^*, \lambda, \mu) \;\leq\; L(x^*, \lambda^*, \mu^*) \;\leq\; L(x, \lambda^*, \mu^*)$$

is satisfied for all vectors $x$ and for all vectors $(\lambda, \mu)$ with $\lambda \geq 0$.

- Such a point $(x^*, \lambda^*, \mu^*)$ is called saddle point
- Under the assumptions of the theorem, saddle point = KKT point

# Duality theorems (1)

**Linear programming duality**

Consider the linear program

$$\min\{c^T x \mid Ax \geq b\}$$

The Lagrangian is

$$L(x,\lambda) := c^T x + \lambda^T (b - Ax)$$

The Lagrange dual function is

$$g(\lambda) = \min_x c^T x + \lambda^T (b - Ax) = \begin{cases} \lambda^T b & \text{if } c^T = \lambda^T A \\ -\infty & \text{otherwise} \end{cases}$$

So the dual is

$$\max\{g(\lambda) \mid \lambda \geq 0\} = \max\{\lambda^T b \mid c^T = \lambda^T A, \lambda \geq 0\}$$

Strong convex duality implies LP duality.

# Duality theorems (2)

- Dénes König (1884–1944):
  Hungarian mathematician; founder of graph theory

**Theorem (König, 1931)**

In a bipartite graph $G = (X \cup Y, E)$,
the number of edges in a maximum matching
is equal to
the number of vertices in a minimum vertex cover.

# Duality theorems (3)

- Philip Hall (1904–1982): English algebraist; Cambridge

**Theorem (Hall's marriage theorem, 1935)**

A bipartite graph $G = (X \cup Y, E)$ contains a matching that covers $X$,
if and only if
$|N(S)| \geq |S|$ for all $S \subset X$.

# Duality theorems (4)

- Robert Palmer Dilworth (1914–1993): American mathematician

**Theorem (Dilworth, 1950)**

For any (finite) partially ordered set,
the maximum size of an antichain
is equal to
the minimum number of chains in a partition into chains

# Duality theorems (5)

- Karl Menger (1902–1985): Austrian mathematician

**Theorem (Menger, 1927)**

For any undirected graph $G = (V, E)$ and $x, y \in V$,
the maximum number of pairwise vertex-disjoint paths from $x$ to $y$
is equal to
the minimum of vertices in a cut separating $x$ from $y$.

- Predecessor and close relative of the max-flow min-cut theorem

# Solving convex programs (1)

Reduction to equality constrained minimization:

**Convex optimization problem**

$$\min\{f_0(x) \mid f_i(x) \leq 0, i = 1\ldots, r,\ Ax = b\}$$

This problem is equivalent to the problem

$$\min\{f_0(x) + \sum_{i=1}^{r} I_-(f_i(x)) \mid Ax = b\}$$

where $I_- : \mathbb{R} \to \mathbb{R}$ is the indicator function of the set $\{u \mid u \leq 0\}$ :

$$I_-(u) = \begin{cases} 0 & \text{if } u \leq 0 \\ \infty & \text{otherwise} \end{cases}$$

**Trouble**

This indicator function $I_-$ is not differentiable!

# Solving convex programs (2): the logarithmic barrier

**Definition**

Consider the following approximate indicator function:

$$\widehat{I}_-(u) = \begin{cases} -\frac{1}{t}\log(-u) & \text{if } u < 0 \\ \infty & \text{otherwise} \end{cases}$$

- The function $\widehat{I}_-$ is convex and non-decreasing
- The function $\widehat{I}_-$ is differentiable
- All sublevel sets of $\widehat{I}_-$ are closed

- As $t \to \infty$, function $\widehat{I}_-$ approximates $I_-$ increasingly well

# Solving convex programs (3): the logarithmic barrier

We approximate $\min\{f_0(x) \mid f_i(x) \leq 0, i = 1 \ldots, r,\ Ax = b\}$ by

$$\min\{f_0(x) + \frac{1}{t}\phi(x) \mid Ax = b\}$$

where $t > 0$ and $\phi(x) := \sum_{i=1}^{r} -\log(-f_i(x))$.

Function $\phi(x)$ is called logarithmic barrier.

**Definition**

The set $\{x^*(t) \mid t > 0\}$ is the central path, where

$$x^*(t) := \arg\min\{f_0(x) + \frac{1}{t}\phi(x) \mid Ax = b\}$$

# Solving convex programs (4): the central path

**Theorem**

*The central path leads to the optimum.*

The Lagrange dual function of $p^* = \min\{f_0(x) \mid f_i(x) \le 0, Ax = b\}$ is

$$g(\lambda,\mu) \;:=\; \inf_x \; f_0(x) + \sum_i \lambda_i f_i(x) + \mu^T(Ax-b)$$

For $x^* = \arg\min\{f_0(x) + \frac{1}{t}\phi(x) \mid Ax = b\}$, we have by KKT that

$$\nabla f_0(x^*) + \sum_i \frac{1}{-tf_i(x^*)} \nabla f_i(x^*) + A^T\mu^* = 0$$

for some $\mu^*$. By setting $\lambda_i^* := 1/(-tf_i(x^*)) > 0$, we get

$$g(\lambda^*,\mu^*) = f_0(x^*) + \sum_{i=1}^r \lambda_i^* f_i(x^*) + (\mu^*)^T(Ax^*-b) = f_0(x^*) - \frac{r}{t}$$

Hence $p^* \;\le\; f_0(x^*) \;=\; g(\lambda^*,\mu^*) + r/t \;\le\; p^* + r/t$.

# Solving convex programs (5)

In order to approximate $p^* = \min\{f_0(x) \mid f_i(x) \leq 0, i = 1\ldots, r,\ Ax = b\}$ within some additive error $\varepsilon > 0$, it hence suffices to solve

$$x^*(t) \ = \ \arg\min\{f_0(x) + \frac{1}{t}\phi(x) \mid Ax = b\}$$

with $t = r/\varepsilon$ (so that $r/t = \varepsilon$).

# Solving convex programs (6)

Linear equality constraints are harmless:

**Equality constrained minimization problem**

Consider the optimization problem $\min\{f(x) \mid Ax = b\}$.

If $\{x \mid Ax = b\} = \{Wy + v \mid y \in \mathbb{R}^k\}$,
then this problem is equivalent to the

**Unconstrained minimization problem**

$\min\{f(Wy + v) \mid y \in \mathbb{R}^k\}$,
that is, the unconstrained minimization of $g : y \mapsto f(Wy + v)$.

In the same fashion, the equality constraints can be eliminated from

$$\min\{f_0(x) + \frac{1}{t}\phi(x) \mid Ax = b\}$$

# Solving convex programs (7): the barrier method

In order to approximate $p^* = \min\{f_0(x) \mid f_i(x) \leq 0, i = 1\ldots, r,\ Ax = b\}$ within error $\varepsilon > 0$, it suffices to solve the unconstrained problem

$$y^*(t) \;=\; \arg\min\{f_0(Wy+v) + \frac{1}{t}\phi(Wy+v)\}$$

with $t = r/\varepsilon$.

As directly solving this problem for high $t$ is usually inefficient, there is:

**The barrier method**

Given a strictly feasible $y = y^{(0)}$ and $t = t^{(0)} > 0$, do

1. compute $y^*(t)$, starting the solution algorithm with $y$ (centering)
2. put $y \leftarrow y^*(t)$ (update)
3. if $r/t < \varepsilon$ then quit; else put $t \leftarrow \mu t$ and repeat. (increase)

- Centering step: by using the Newton method

# A final example

**Problem**

Design a cylindrical tin can with volume at least $v$ units, such that the total surface area is minimal.

For height $h$ and radius $r$, this problem becomes

$$\begin{aligned} \text{minimize} \quad & f(r,h) := 2\pi(r^2 + rh) \\ \text{subject to} \quad & \pi r^2 h \geq v \\ & r > 0 \text{ and } h > 0 \end{aligned}$$

This is not a convex optimization problem!
The substitution $r = e^x$ and $h = e^y$ yields the convex problem

$$\begin{aligned} \text{minimize} \quad & g(x,y) := 2\pi(e^{2x} + e^{x+y}) \\ \text{subject to} \quad & \ln(v/\pi) - 2x - y \leq 0 \\ & x, y \in \mathbb{R} \end{aligned}$$

# A final example (continued)

$$\begin{array}{ll} \text{minimize} & g(x,y) := 2\pi(e^{2x} + e^{x+y}) \\ \text{subject to} & \ln(v/\pi) - 2x - y \le 0 \\ & x, y \in \mathbb{R} \end{array}$$

$$L(x,\lambda) = 2\pi(e^{2x} + e^{x+y}) + \lambda\left(\ln(v/\pi) - 2x - y\right)$$

$$2\pi \begin{bmatrix} 2e^{2x} + e^{x+y} \\ e^{x+y} \end{bmatrix} = \lambda \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

- Hence $2e^{2x} + e^{x+y} = 2e^{x+y}$, whence $2e^x = e^y$ and $2r = h$.
- As the constraint holds with equality (why?), we can express $r$ and $h$ in terms of $v$.

# Homework 3

- Read chapters 4, 5 and 11.1–11.3 in Boyd & Vandenberghe
- Recommended exercises:
  43, 45, 51, 55, 60, 62, 63, 66, 68

Collection of exercises can be downloaded from:
`http://www.win.tue.nl/~gwoegi/optimization/`

---

Attention!
Weeks 2-5 (Sep 8; Sep 15; Sep 22; Sep 29):
- Tuesday 1+2: instructions
- Tuesday 3+4: lecture