Discrete Mathematics for Bioinformatics (P1)

Lecture & Tutorials (4 + 2)

Winter 2009/10, FU Berlin

Prof. Dr. Alexander Bockmayr
bockmayr <AT> mi.fu-berlin dot de
Prof. Dr. Knut Reinert
reinert <AT> mi.fu-berlin dot de
Sandro Andreotti
andreott <AT> mi.fu-berlin dot de

Web page at http://www.inf.fu-berlin.de/

Organizational issues

Lecture and slides will be in English.

Credits. Contributed to the slides have Prof. Alexander Bockmayr, Prof. Knut Reinert, Dr. Gunnar Klau, and Prof. Daniel Huson (Tübingen)

Please note the following:

- You can download all slides (also in script form) typically the day after the lecture. It is not necessary to copy the slides.
- However, the slides are **not** a complete script. Please take notes and maybe copy things on the blackboard.
- Exercises will be held by Sandro Andreotti.
- **For all current information please check frequently `http://www.inf.fu-berlin.de/groups/ag-bio`.**

Lectures

Lectures take place

- Tuesdays from 10-12 am in SR 006 (Takustraße 9)
- Thursdays from 10-12 am in SR 006 (Takustraße 9)

Exercises

The exercises are mandatory and will be counted as *active participation* in the three column model of the FU Berlin.

Active participation = presence + reviews + some blackboard action.

There will also be some practical exercices (programming).

Reviews: 2 $\times$ 60 min (End November, Mid January)

Exercises (2)

Exercises will be held once weekly in room SR 032 (A6) (Fridays from 10-12 am).

First meeting: Friday, 24 Oct 2007

Office hours

See websites.

Content

1. Linear Programming
   (a) (L1, 13 Oct) Optimization problems (AB)
   (b) (L2, 15 Oct) Polyhedra (AB)
   (c) (L3, 20 Oct) Simplex Algorithm (AB)
   (d) (L4, 22 Oct) Duality & Applications (AB)
2. Combinatorial Optimization I
   (a) (L5, 27 Oct) Branch-and-Cut (Intro) (KR)
   (b) (L6, 29 Oct) B&C for sequence alignment (Modelling) (KR)
   (c) (L7, 3 Nov) B&C for sequence alignment (Algorithms) (KR)
   (d) (L8, 5 Nov) B&C for sequence alignment II (KR)
   (e) (L9, 10 Nov) B&C for RNA alignment (KR)
   (f) (L10, 12 Nov) Lagrangian Relaxation for RNA alignment (KR)
3. Data structures and analysis methods I
   (a) (L19, 17 Nov) Hashing (KR)
   (b) (L20, 19 Nov) Hashing II (KR)
   (c) (L21, 24 Nov) Randomized algorithms: search trees/skiplists (KR)
   (d) (L22, 26 Nov) Randomized analysis, Chernoff bounds I (KR)
   (e) (L23, 1 Dec) Randomized analysis, Chernoff bounds II (KR)
4. Combinatorial Optimization II
   (a) (L11, 3 Dec) Constraint programming I (AB)
   (b) (L12, 8 Dec) Constraint programming II (AB)
   (c) (L13, 10 Dec) Local search and metaheuristics I (AB)
   (d) (L14, 15 Dec) Local search and metaheuristics II (AB)
5. Graph algorithms
   (a) (L15, 17 Dec) Shortest path (AB)
   (b) (L16, 5 Jan) Network flow I (AB)
   (c) (L17, 7 Jan) Network flow II (AB)
   (d) (L18, 12 Jan) Matching (AB)
6. Data structures and analysis methods II
   (a) (L24, 14 Jan) $\approx$ Review 2
   (b) (L25, 19 Jan) Locality sensitive hashing and applications (KR)

   (c) (L26, 21 Jan) Tree decompositions and applications I (KR)
   (d) (L27, 26 Jan) Tree decompositions and applications II (KR)

7. Computability and complexity theory
   (a) (L28, 28 Jan) Computability I (AB)
   (b) (L29, 2 Feb) Computability II (AB)
   (c) (L30, 4 Feb) Complexity theory I (AB)
   (d) (L31, 9 Feb) Complexity theory II (AB)
8. (L32, 11 Feb) Written examination

Discrete mathematics

What is *discrete mathematics*? (discrete, from lat. *discernere*, to discern, to separate)

From `http://mathworld.wolfram.com`:

> Discrete mathematics is the branch of mathematics dealing with objects that can assume only distinct, separated values. The term "discrete mathematics" is therefore used in contrast with "continuous mathematics", which is the branch of mathematics dealing with objects that can vary smoothly (and which includes, for example, calculus). Whereas discrete objects can often be characterized by integers, continuous objects require real numbers.
>
> The study of how discrete objects combine with one another and the probabilities of various outcomes is known as combinatorics. Other fields of mathematics that are considered to be part of discrete mathematics include graph theory and the theory of computation. Topics in number theory such as congruences and recurrence relations are also considered part of discrete mathematics.
>
> The study of topics in discrete mathematics usually includes the study of algorithms, their implementations, and efficiencies. Discrete mathematics is the mathematical language of computer science, and as such, its importance has increased dramatically in recent decades.

Why is it important in bioinformatics?

Linear programming

Optimization Problems

- *General optimization* problem

$$\max\{z(x) \mid f_j(x) \leq 0, x \in D\} \text{ or } \min\{z(x) \mid f_j(x) \leq 0, x \in D\}$$

  where $D \subseteq \mathbb{R}^n, f_j : D \to \mathbb{R}$, for $j = 1, \ldots, m$, $z : D \to \mathbb{R}$.

- *Linear optimization* problem

$$\max\{c^T x \mid Ax \lesseqgtr b, x \in \mathbb{R}^n\}, \text{ with } c \in \mathbb{R}^n, A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$$

- *Integer* optimization problem: $x \in \mathbb{Z}^n$
- *0-1* optimization problem: $x \in \{0,1\}^n$

Feasible and optimal solutions

- Consider the optimization problem

$$\max\{z(x) \mid f_j(x) \leq 0, x \in D, j = 1, \ldots, m\}$$

- A *feasible solution* is a vector $x^* \in D \subseteq \mathbb{R}^n$ such that $f_j(x^*) \leq 0$, for all $j = 1, \ldots, m$.
- The set of all feasible solutions is called the *feasible region*.
- An *optimal solution* is a feasible solution such that

$$z(x^*) = \max\{z(x) \mid f_j(x) \leq 0, x \in D, j = 1, \ldots, m\}.$$

- Feasible or optimal solutions
  - need not exist,
  - need not be unique.

Transformations

- $\min\{z(x) \mid x \in S\} = \max\{-z(x) \mid x \in S\}$.
- $f(x) \geq a$ if and only if $-f(x) \leq -a$.
- $f(x) = a$ if and only if $f(x) \leq a \wedge -f(x) \leq -a$.

Lemma

Any linear programming problem can be brought to the form

$$\max\{c^T x \mid Ax \leq b\} \text{ or } \max\{c^T x \mid Ax = b, x \geq 0\}.$$

Proof: a) $a^T x \leq \beta \rightsquigarrow a^T x + x' = \beta, x' \geq 0$ *(slack variable)*
b) x free $\rightsquigarrow x = x^+ - x^-$, $x^+, x^- \geq 0$.

Practical problem solving

1. Model building
2. Model solving
3. Model analysis

Example: Production problem

A firm produces n different goods using m different raw materials.

- b_i: available amount of the i-th raw material
- a_{ij}: number of units of the i-th material needed to produce one unit of the j-th good
- c_j: revenue for one unit of the j-th good.

Decide how much of each good to produce in order to maximize the total revenue $\rightsquigarrow$ *decision variables* x_j.

Linear programming formulation

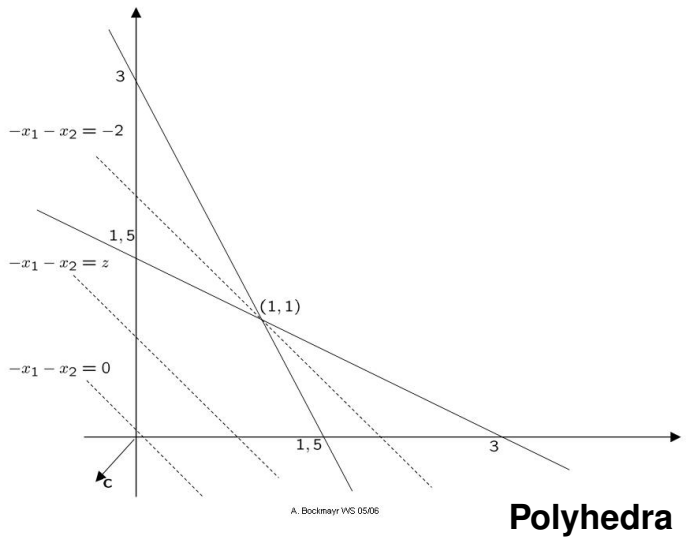
$$\begin{array}{lccccccc}
\text{max} & c_1x_1 & + & \cdots & + & c_nx_n & & \\
\text{w.r.t.} & a_{11}x_1 & + & \cdots & + & a_{1n}x_n & \leq & b_1, \\
& \vdots & & & & \vdots & & \\
& a_{m1}x_1 & + & \cdots & + & a_{mn}x_n & \leq & b_m, \\
& x_1, & & \ldots & & ,x_n & \geq & 0.
\end{array}$$

In matrix notation:

$$\max\{c^T x \mid Ax \leq b, x \geq 0\},$$

where $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m, c \in \mathbb{R}^n, x \in \mathbb{R}^n$.

Geometric illustration



$$\begin{array}{lccccc}
\text{max} & x_1 & + & x_2 & & \\
\text{w.r.t.} & x_1 & + & 2x_2 & \leq & 3 \\
& 2x_1 & + & x_2 & \leq & 3 \\
& x_1 & , & x_2 & \geq & 0
\end{array}$$

Polyhedra

- *Hyperplane* $H = \{x \in \mathbb{R}^n \mid a^T x = \beta\}, a \in \mathbb{R}^n \setminus \{0\}, \beta \in \mathbb{R}$
- *Closed halfspace* $\overline{H} = \{x \in \mathbb{R}^n \mid a^T x \leq \beta\}$

- *Polyhedron* $P = \{x \in \mathbb{R}^n \mid Ax \leq b\}, A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$
- *Polytope* $P = \{x \in \mathbb{R}^n \mid Ax \leq b,\ l \leq x \leq u\}, l, u \in \mathbb{R}^n$
- *Polyhedral cone* $P = \{x \in \mathbb{R}^n \mid Ax \leq 0\}$

The feasible set

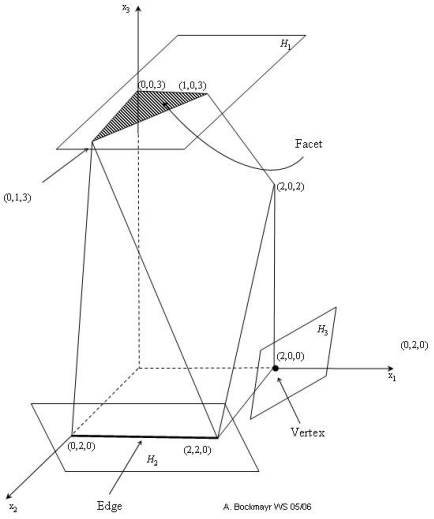
$$P = \{x \in \mathbb{R}^n \mid Ax \leq b\}$$

of a linear optimization problem is a polyhedron.

Vertices, Faces, Facets

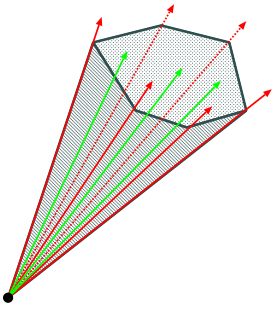
- $P \subseteq \overline{H}, H \cap P \neq \emptyset$ *(Supporting hyperplane)*
- $F = P \cap H$ *(Face)*
- $\dim(F) = 0$ *(Vertex)*
- $\dim(F) = 1$ *(Edge)*
- $\dim(F) = \dim(P) - 1$ *(Facet)*
- P *pointed:* P has at least one vertex.

Illustration



Rays and extreme rays

- $r \in \mathbb{R}^n$ is a *ray* of the polyhedron P if for each $x \in P$ the set $\{x + \lambda r \mid \lambda \geq 0\}$ is contained in P.
- A ray r of P is *extreme* if there do not exist two linearly independent rays r^1, r^2 of P such that $r = \frac{1}{2}(r^1 + r^2)$.



Hull operations

- $x \in \mathbb{R}^n$ is a *linear combination* of $x^1, \ldots, x^k \in \mathbb{R}^n$ if

$$x = \lambda_1 x^1 + \cdots + \lambda_k x^k, \text{ for some } \lambda_1, \ldots, \lambda_k \in \mathbb{R}.$$

- If, in addition

$$\left\{ \begin{array}{ll} \lambda_1, \ldots, \lambda_k \geq 0, & \\ & \lambda_1 + \cdots + \lambda_k = 1, \\ \lambda_1, \ldots, \lambda_k \geq 0, & \lambda_1 + \cdots + \lambda_k = 1, \end{array} \right\} x \text{ is a } \left\{ \begin{array}{c} \textit{conic} \\ \textit{affine} \\ \textit{convex} \end{array} \right\} \text{combination.}$$

- For $S \subseteq \mathbb{R}^n, S \neq \emptyset$, the set $\mathrm{lin}(S)$ (resp. $\mathrm{cone}(S), \mathrm{aff}(S), \mathrm{conv}(S)$) of all linear (resp. conic, affine, convex) combinations of finitely many vectors of S is called the *linear (resp. conic, affine, convex) hull of* S.

Outer and inner descriptions

- A subset $P \subseteq \mathbb{R}^n$ is a H*-polytope,* i.e., a *bounded* set of the form *Outer*

$$P = \{x \in \mathbb{R}^n \mid Ax \leq b\}, \text{ for some } A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m.$$

if and only if it is a V*-polytope,* i.e., *Inner*

$$P = \mathrm{conv}(V), \text{ for some finite } V \subset \mathbb{R}^n$$

- A subset $C \subseteq \mathbb{R}^n$ is a H*-cone,* i.e., *Outer*

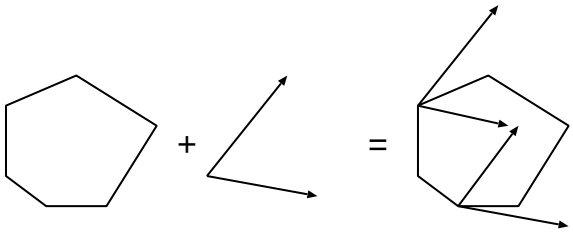
$$C = \{x \in \mathbb{R}^n \mid Ax \leq 0\}, \text{ for some } A \in \mathbb{R}^{m \times n}.$$

if and only if it is a V*-cone,* i.e., *Inner*

$$C = \mathrm{cone}(Y), \text{ for some finite } Y \subset \mathbb{R}^n$$

Minkowski sum

- $X, Y \subseteq \mathbb{R}^n$
- $X + Y = \{x + y \mid x \in X, y \in Y\}$ *(Minkowski sum)*



Main theorem for polyhedra

A subset $P \subseteq \mathbb{R}^n$ is a H*-polyhedron,* i.e., *Outer*

$$P = \{x \in \mathbb{R}^n \mid Ax \leq b\}, \text{ for some } A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m.$$

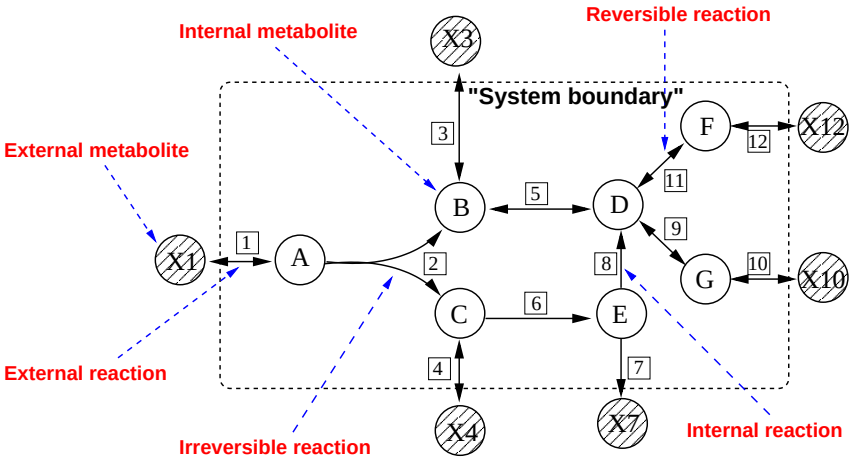
if and only if it is a V*-polyhedron,* i.e., *Inner*

$$P = \mathrm{conv}(V) + \mathrm{cone}(Y), \text{ for some finite } V, Y \subset \mathbb{R}^n$$

Theorem of Minkowski

- For each polyhedron $P \subseteq \mathbb{R}^n$ there exist finitely many points $p^1, \ldots, p^k$ in P and finitely many rays $r^1, \ldots, r^l$ of P such that
$$P = \text{conv}(p^1, \ldots, p^k) + \text{cone}(r^1, \ldots, r^l).$$
- If the polyhedron P is pointed, then $p^1, \ldots, p^k$ may be chosen as the uniquely determined vertices of P, and $r^1, \ldots, r^l$ as representatives of the up to scalar multiplication uniquely determined extreme rays of P.
- *Special cases*
  - A polytope is the convex hull of its vertices.
  - A pointed polyhedral cone is the conic hull of its extreme rays.

Application: Metabolic networks



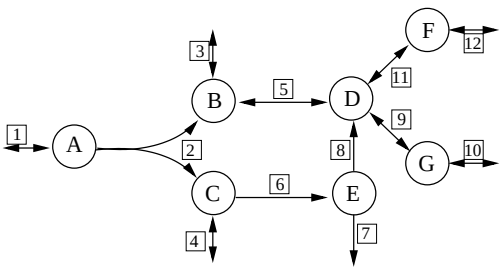
Stoichiometric matrix

- Metabolites (internal) $\rightsquigarrow$ rows
- Biochemical reactions $\rightsquigarrow$ columns

$$kA + lC \xrightarrow{j} mE + nH$$

$$\begin{array}{c} A \\ B \\ C \\ D \\ E \\ F \\ G \\ H \end{array} \begin{pmatrix} \ldots & -k & \ldots \\ & 0 & \\ & -l & \\ & 0 & \\ & m & \\ & 0 & \\ & 0 & \\ \ldots & n & \ldots \end{pmatrix}$$

Example network



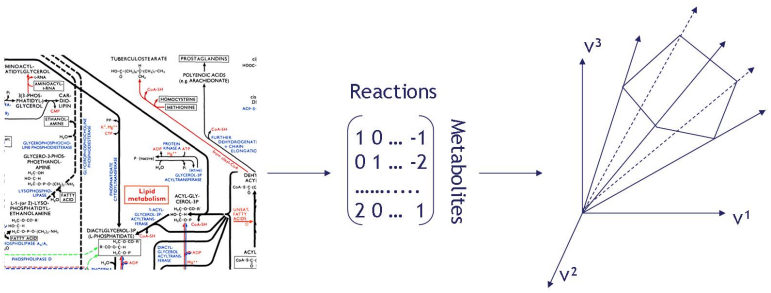
$$S = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & -1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \end{pmatrix}.$$

Flux cone

- Flux balance: $Sv = 0$
- Irreversibiliy of some reactions: $v_i \geq 0, i \in Irr$.
- *Steady-state flux cone*

$$C = \{v \in \mathbb{R}^n \mid Sv = 0, v_i \geq 0, \text{ for } i \in Irr\}$$

- *Metabolic network analysis* $\rightsquigarrow$ find $p^1, \ldots, p^k \in C$ with $C = \text{cone}\{p^1, \ldots, p^k\}$.



Simplex Algorithm: Geometric view

Linear optimization problem

$$\max\{c^T x \mid Ax \leq b, x \in \mathbb{R}^n\} \qquad \text{(LP)}$$

Simplex-Algorithm (Dantzig 1947)

1. Find a vertex of P.
2. Proceed from vertex to vertex along edges of P such that the objective function $z = c^T x$ increases.
3. Either a vertex will be reached that is optimal, or an edge will be chosen which goes off to infinity and along which z is unbounded.

Basic solutions

- $Ax \leq b$, $A \in \mathbb{R}^{m \times n}, rank(A) = n$.
- $M = \{1, \ldots, m\}$ row indices, $N = \{1, \ldots, n\}$ column indices
- For $I \subseteq M, J \subseteq N$ let A_{IJ} denote the submatrix of A defined by the rows in I and the columns in J.
- $I \subseteq M, |I| = n$ is called a *basis of* A iff $A_{I*} = A_{IN}$ is non-singular.
- In this case, $A_{I*}^{-1} b_I$, where b_I is the subvector of b defined by the indices in I, is called a *basic solution.*
- If $x = A_{I*}^{-1} b_I$ satisfies $Ax \leq b$, then x called a *basic feasible solution* and I is called a *feasible basis.*

Algebraic characterization of vertices

Theorem

Given the non-empty polyhedron $P = \{x \in \mathbb{R}^n \mid Ax \leq b\}$, where $\mathrm{rank}(A) = n$, a vector $v \in \mathbb{R}^n$ is a vertex of P if and only if it is a basic feasible solution of $Ax \leq b$, for some basis I of A.

For any $c \in \mathbb{R}^n$, either the maximum value of $z = c^T x$ for $x \in P$ is attained at a vertex of P or z is unbounded on P.

Corollary

P has at least one and at most finitely many vertices.

Remark

In general, a vertex may be defined by several bases.

Simplex Algorithm: Algebraic version

- Suppose $\mathrm{rank}(A) = n$ (otherwise apply Gaussian elimination).
- Suppose I is a feasible basis with corresponding vertex $v = A_{I*}^{-1} b_I$.
- Compute $u^T \stackrel{\text{def}}{=} c^T A_{I*}^{-1}$ (vector of n components indexed by I).
- If $u \geq 0$, then v is an optimal solution, because for each feasible solution x

$$c^T x = u^T A_{I*} x \leq u^T b_I = u^T A_{I*} v = c^T v.$$

- If $u \ngeq 0$, choose $i \in I$ such that $u_i < 0$ and define the direction $d \stackrel{\text{def}}{=} -A_{I*}^{-1} e_i$, where e_i is the i-th unit basis vector in $\mathbb{R}^I$.
- Next increase the objective function value by going from v in direction d, while maintaining feasibility.

Simplex Algorithm: Algebraic version (2)

1. If $Ad \nleq 0$, the largest $\lambda \geq 0$ for which $v + \lambda d$ is still feasible is

$$\lambda^* = \min\{\frac{b_l - A_{l*} v}{A_{l*} d} \mid l \in \{1, \ldots, m\}, A_{l*} d > 0\}. \qquad \text{(PIV)}$$

   Let this minimum be attained at index k. Then $k \notin I$ because $A_{I*} d = -e_i \leq 0$.

   Define $I' = (I \setminus \{i\}) \cup \{k\}$, which corresponds to the vertex $v + \lambda^* d$.

   Replace I by I' and repeat the iteration.

2. If $Ad \leq 0$, then $v + \lambda d$ is feasible, for all $\lambda \geq 0$. Moreover,

$$c^T d = -c^T A_{I*}^{-1} e_i = -u^T e_i = -u_i > 0.$$

   Thus the objective function can be increased along d to infinity and the problem is unbounded.

Termination and complexity

- The method terminates if the indices i and k are chosen in the right way (such choices are called *pivoting rules*).

- Following the rule of Bland, one can choose the smallest i such that $u_i < 0$ and the smallest k attaining the minimum in (PIV).
- For most known pivoting rules, sequences of examples have been constructed such that the number of iterations is exponential in $m+n$ (e.g. Klee-Minty cubes).
- Although no pivoting rule is known to yield a polynomial time algorithm, the Simplex method turns out to work very well in practice.

Simplex : Phase I

- In order to find an *initial feasible basis*, consider the auxiliary linear program

$$\max\{y \mid Ax - by \leq 0,\ -y \leq 0,\ y \leq 1\}, \tag{Aux}$$

  where y is a new variable.
- Given an arbitrary basis K of A, obtain a feasible basis I for (Aux) by choosing $I = K \cup \{m+1\}$. The corresponding basic feasible solution is 0.
- Apply the Simplex method to (Aux). If the optimum value is 0, then (LP) is infeasible. Otherwise, the optimum value has to be 1.
- If I' is the final feasible basis of (Aux), then $K' = I' \setminus \{m+2\}$ can be used as an initial feasible basis for (LP).

Duality

- *Primal problem:* $z_P = \max\{\mathbf{c^T}x \mid \quad Ax \leq b, \quad x \in \mathbb{R}^n\}$ (P)
- *Dual problem:* $w_D = \min\{b^T u \mid \quad A^T u = \mathbf{c}, \quad u \geq 0\}$ (D)

General form

(P)			(D)		
min	$c^T x$		max	$u^T b$	
w.r.t.	$A_{i*}x \geq b_i,$	$i \in M_1$	w.r.t	$u_i \geq 0,$	$i \in M_1$
	$A_{i*}x \leq b_i,$	$i \in M_2$		$u_i \leq 0,$	$i \in M_2$
	$A_{i*}x = b_i,$	$i \in M_3$		u_i free,	$i \in M_3$
	$x_j \geq 0,$	$j \in N_1$		$(A_{*j})^T u \leq c_j,$	$j \in N_1$
	$x_j \leq 0,$	$j \in N_2$		$(A_{*j})^T u \geq c_j,$	$j \in N_2$
	x_j free,	$j \in N_3$		$(A_{*j})^T u = c_j,$	$j \in N_3$

Duality theorems

Theorem

- If x^* is primal feasible and u^* is dual feasible, then

$$c^T x^* \leq z_P \leq w_D \leq b^T u^*.$$

- Only four possibilities:
  1. z_P and w_D are both finite and equal.
  2. $z_P = +\infty$ and (D) is infeasible.

3. $w_D = -\infty$ and (P) is infeasible.
4. (P) and (D) are both infeasible.

Complexity of linear programming

Theorem (Khachyian 79)
The following problems are solvable in polynomial time:

- Given a matrix $A \in \mathbb{Q}^{m \times n}$ and a vector $b \in \mathbb{Q}^m$, decide whether $Ax \leq b$ has a solution $x \in \mathbb{Q}^n$, and if so, find one.
- (Linear programming problem) Given a matrix $A \in \mathbb{Q}^{m \times n}$ and vectors $b \in \mathbb{Q}^m, c \in \mathbb{Q}^n$, decide whether $\max\{c^T x \mid Ax \leq b, x \in \mathbb{Q}^n\}$ is infeasible, finite, or unbounded. If it is finite, find an optimal solution. If it is unbounded, find a feasible solution x_0, and find a vector $d \in \mathbb{Q}^n$ with $Ad \leq 0$ and $c^T d > 0$.

Complexity of integer linear programming

Satisfiability	over $\mathbb{Q}$	over $\mathbb{Z}$	over $\mathbb{N}$
Linear equations	polynomial	polynomial	NP-complete
Linear inequalities	polynomial	NP-complete	NP-complete