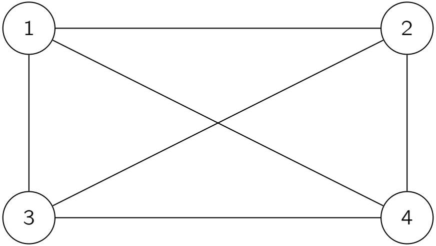
Formulations and Integer Programming

Jesper Larsen[1]

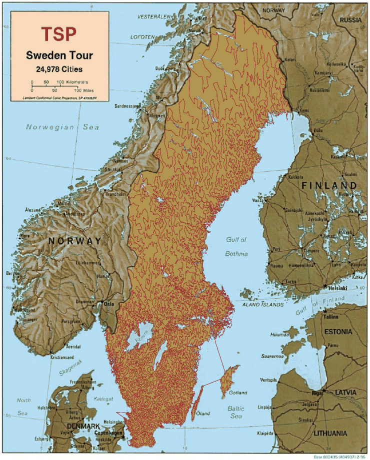
[1]Department of Management Engineering
Technical University of Denmark

The Travelling Salesman Problem

- Typically modelled as an undirected weighted graph
- **Vertices** model "cities", **edges** denote connections between "cities"
- The weight of an edge gives the "distance" (here denoted c_{ij})
- The graph is typically complete
- Comes an **assymetric** and **symmetric** variants

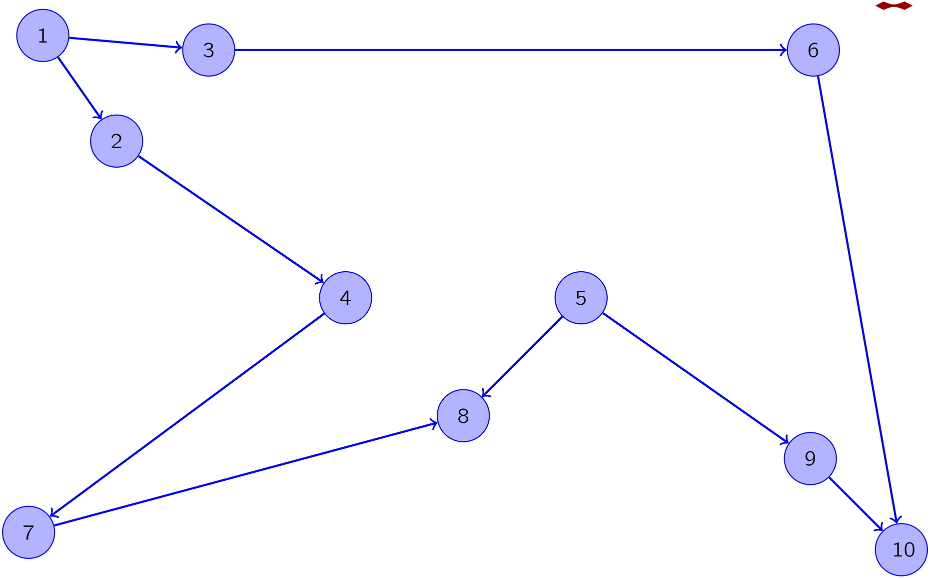


The Travelling Salesman



- Tour of Sweden have 24978 nodes.
- TSP record: 528,280,881 nodes.
- Real-life applications of TSP are VLSI design and DNA sequencing.
- For more info see `www.tsp.gatech.edu`

Example of a tour

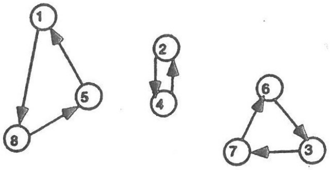


Modelling the TSP problem – in a directed graph

- What to decide? "Where do we go from city i?" or framed differently "Do we go directly from city i to city j?"
- Variables: $x_{ij} = 1$ if the tour goes directly from i to j.
- Objective function: Sum of all distances on the tour:
  - $\min \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} x_{ij}$

Further modelling

- Constraints: For every city i we enter the city from exactly one other city and we exit it towards exactly one other city.
  - $\sum_{i=1}^{n} x_{ij} = 1$ for each j and
  - $\sum_{j=1}^{n} x_{ij} = 1$ for each i
- Is that enough?
  - No, a "feasible" solution could be $1 \to 5 \to 8 \to 1$, $2 \to 4 \to 2$ and $3 \to 7 \to 6 \to 3$



Subtour elimination

- We need to ensure that we do not get these "subtours".
- For every time I define a set S of nodes, there has to be at least one edge with an endpoint in S and the other outside of S.
- $\sum_{i \in S} \sum_{j \notin S} x_{ij} \geq 1$ for $S \subset N, S \neq \emptyset$

TSP: "Classic" TSP model

In conclusion the model looks like. Let $N = \{1, 2, \ldots, n\}$.

$$\min \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} x_{ij} \tag{1}$$

$$\text{s.t.} \sum_{i=1}^{n} x_{ij} = 1 \qquad \forall j = 1, 2, \ldots, n \tag{2}$$

$$\sum_{j=1}^{n} x_{ij} = 1 \qquad \forall i = 1, 2, \ldots, n \tag{3}$$

$$\sum_{i \in S} \sum_{j \notin S} x_{ij} \geq 1 \qquad \forall S \subset N, S \neq \emptyset \tag{4}$$

$$x_{ij} \in \{0, 1\} \qquad \forall i, j \in N \tag{5}$$

Alternative model

- An exponential number of constraints is not very beneficial for finding the optimal solution quickly.
- Can we solve it in another way?
- Introduce integer variable s_i, $i = 1, 2, \ldots, n$. This variable indicates the sequence in which the cities are visited.

Use of sequence numbers

- $s_1 = 1$
- constraint:

$$x_{ij} = 1 \quad \Rightarrow \quad s_j = s_i + 1$$

- sufficient to write:

$$x_{ij} = 1 \quad \Rightarrow \quad s_j \geq s_i + 1$$

- MIP constraint:

$$s_j \geq s_i + 1 - M(1 - x_{ij})$$

- For all i, j where $j \neq 1$

Alternative TSP model
– Also known as the Miller-Tucker-Zemlin formulation

In conclusion the model looks like. Let $N = \{1, 2, \ldots, n\}$.

$$\min \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} x_{ij} \tag{6}$$

$$\text{s.t.} \sum_{i=1}^{n} x_{ij} = 1 \qquad \forall j = 1, 2, \ldots, n \tag{7}$$

$$\sum_{j=1}^{n} x_{ij} = 1 \qquad \forall i = 1, 2, \ldots, n \tag{8}$$

$$s_1 = 1 \tag{9}$$

$$2 \leq s_i \leq n \qquad \forall i = 2, 3, \ldots, n \tag{10}$$

$$s_j \geq s_i + 1 - M(1 - x_{ij}) \qquad \forall i, j \in N, i \neq 1, j \neq 1 \tag{11}$$

$$x_{ij} \in \{0, 1\} \qquad \forall i, j \in N \tag{12}$$

$$s_i \geq 0 \text{ and integer} \qquad \forall i \in N \tag{13}$$

Uncapacitated Facility Location (UFL)

- Given a set of *potential* **depots** $N = \{1, 2, \ldots, n\}$ and a set $M = \{1, 2, \ldots, m\}$ of **clients** (or customers), suppose there is a fixed cost f_j associated with the use of depot j, and a transportation cost c_{ij} if all of client i's order is delivered from depot j.
- The problem is to decide which depots to open, and which depots serves each client so as to minimize the sum of fixed and transportation cost.

An IP model for the UFL

- Definition of variables
  - Depot opening variable y_j (1 if depot is open, otherwise 0)
  - x_{ij} is the fraction of the demand client i gets from depot j
- Objective function is the sum of depot opening cost and transportation cost:
  - Depot opening cost: $\sum_{j \in N} f_j y_j$
  - Transportation cost: $\sum_{i \in M} \sum_{j \in N} c_{ij} x_{ij}$
  - In total: $\min \sum_{i \in M} \sum_{j \in N} c_{ij} x_{ij} + \sum_{j \in N} f_j y_j$

Constraints for the UFL

- Satisfaction of client demand (for all clients i):
  - $\sum_{j \in N} x_{ij} = 1$
- Link y_j and x_{ij} variables. We can only supply from a depot if it is open:
  - $\sum_{i \in M} x_{ij} \leq K y_j$ for $j \in N$
  - ... and we can set K ("big M" notation) to m.

First UFL model

In conclusion, we therefore have the following model:

$$\min \sum_{i \in M} \sum_{j \in N} c_{ij} x_{ij} + \sum_{j \in N} f_j y_j \tag{14}$$

$$\text{s.t.} \sum_{j \in N} x_{ij} = 1 \qquad \forall i \in M \tag{15}$$

$$\sum_{i \in M} x_{ij} \leq m y_j \qquad \forall j \in N \tag{16}$$

$$0 \leq x_{ij} \leq 1 \qquad \forall i \in M, j \in N \tag{17}$$

$$y_j \in \{0, 1\} \qquad \forall j \in N \tag{18}$$

An alternative model for the UFL

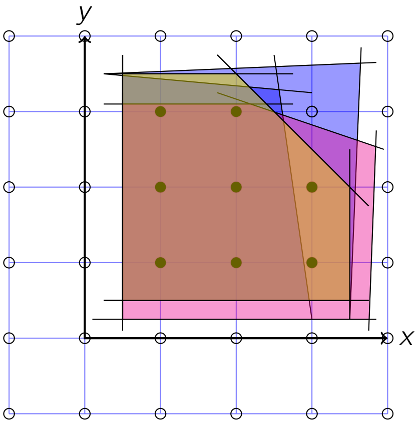
- We have the same variables as before, and the same objective function. And also the first constraint remains identical.

$$
\begin{aligned}
\min \; & \sum_{i \in M} \sum_{j \in N} c_{ij} x_{ij} + \sum_{j \in N} f_j y_j && && (19) \\
\text{s.t.} \; & \sum_{j \in N} x_{ij} = 1 && \forall i \in M && (20) \\
& \color{red}{x_{ij} \leq y_j} && \color{red}{\forall i \in N, j \in M} && (21) \\
& 0 \leq x_{ij} \leq 1 && \forall i \in M, j \in N && (22) \\
& y_j \in \{0, 1\} && \forall j \in N && (23)
\end{aligned}
$$

Polyhedron and Formulation

- A subset of R^n described by a finite set of linear constraints $P = \{x \in R^n : Ax \leq b\}$ is a **polyhedron**.
- A polyhedron $P \subset R^{n+p}$ is a **formulation** for a set $X \subset Z^n \times R^p$ if and only if $X = (Z^n \times R^p) \cap P$.
- Integer Program: A polyhedron $P \subset R^n$ is a **formulation** for a set $X \subset Z^n$ if and only if $X = Z^n \cap P$.

Understanding a formulation



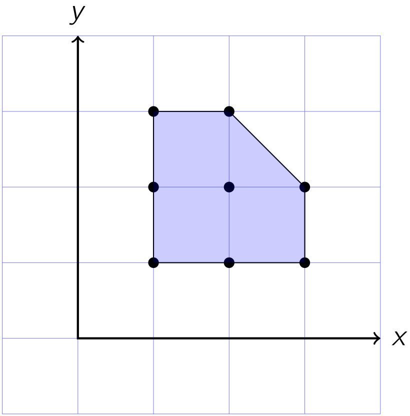
Convex Hull

- Given a set $x \subset R^n$ the **convex hull of** X, denoted conv(X) is defined as:

$$\text{conv}(X) = \{ \quad x : x = \sum_{i=1}^{t} \lambda_i x^i, \sum_{i=1}^{t} \lambda_i = 1, \lambda \geq 0 \text{ for } i = 1, \dots, t \text{ over all finite subsets } \{x^1, x^2, \dots, x^t\} \text{ of } X\}$$

- **Proposition:** conv(X) is a polyhedron
- **Proposition:** The extreme points of conv(X) all lie in X.

Ideal Formulation



Ideal Formulation

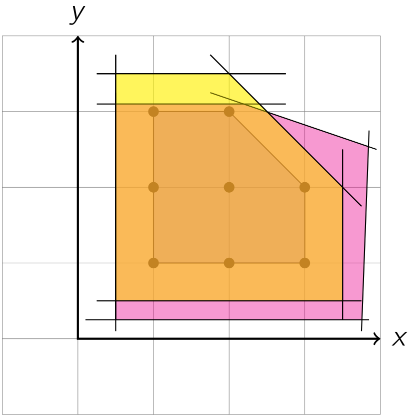
- In the ideal formulation our integer programming problem can be solved by solving a linear programming problem over the formulation.
- The **ideal formulation** in most cases consists of an enormous (exponential) number of inequalities needed to describe conv(X), and there is no simple characterization of them.

Better Formulations

Instead we could rather ask:

- Given two formulations P_1 and P_2 for X when can we say that one is better than the other?
- Given a set $X \subset R^n$ and two formulations P_1 and P_2 for X, P_1 is a **better formulation** than P_2 if $P_1 \subset P_2$.

Better Formulations



Formulations for a Knapsack Set

Look at the set
$X = \{(0,0,0,0),(1,0,0,0),(0,1,0,0),(0,0,1,0),(0,0,0,1),(0,1,0,1),(0,0,1,1)\}$
and the three formulations

$$P_1 = \{x \in R^4 : \quad 0 \le x \le 1, 83x_1 + 61x_2 + 49x_3 + 20x_4 \le 100\}$$
$$P_2 = \{x \in R^4 : \quad 0 \le x \le 1, 4x_1 + 3x_2 + 2x_3 + 1x_4 \le 4\}$$
$$P_3 = \{x \in R^4 : \quad 0 \le x \le 1, \begin{array}{lllll} 4x_1 & +3x_2 & +2x_3 & +1x_4 & \le 4 \\ 1x_1 & +1x_2 & +1x_3 & & \le 1 \\ 1x_1 & & & +1x_4 & \le 1 \end{array} \}$$

Wolsey writes: "It is easily seen that $P_3 \subset P_2 \subset P_1$, and it can be checked that $P_3 = \text{conv}(X)$".

Comparing formulations of UFL

Let P_1 be the formulation with the constraints

$$\sum_{i \in M} x_{ij} \leq m y_j$$

Let P_2 be the formulation with the constraints

$$x_{ij} \leq y_j$$

Now we want to show that P_2 is a better formulation than P_1.

P_2 is a better formulation than P_1

Basically we need to show:

1. $P_2 \subseteq P_1$
2. $P_2 \subset P_1$, that is, find a $(x, y) \in P_1$ but $(x, y) \notin P_2$

- The first item is relatively easy. Consider a solution (x, y) that is feasible in P_2.
- That means that $x_{ij} \leq y_j$ is fulfilled for all $i \in M$ and $j \in N$.
- Now for fixed j sum LHS and RHS. The resulting constraint is still valid for P_2.
  - LHS becomes $x_{1j} + x_{2j} + \ldots + x_{mj} = \sum_{i \in M} x_{ij}$
  - RHS becomes $y_j + y_j + y_j + \ldots + y_j = my_j$

$P_2 \subseteq P_1$

- This is actually exactly the constraint that regulates the relationship between x and y in P_1.
- So any feasible solution of P_2 is also a feasible solution of P_1.

P_2 is a better formulation than P_1 step 2

- We now need to find $(x, y) \in P_1$ but $(x, y) \notin P_2$.
- Suppose for simplicity that $m = 2n$, so twice as many customers as depots. And suppose that we have at least two depots.
- Now define a solution where each depot supplies exactly two customers. It could be:
  - $x_{11} = 1, x_{21} = 1$ and the rest of x_{i1} are zero
  - $x_{32} = 1, x_{42} = 1$ and the rest of x_{i2} are zero
  - $x_{ij} = 1$ for $i = 2j - 1$ and $j = 2j$
- Then we need to assign values to the y_i variables. Here we take $y_j = \frac{2}{m}$

$P_2 \subset P_1$

Now I have established a solution (x, y). I can check if it is feasible by inserting it into our two (sets of) constraints in each formulation.

- P_1: The constraints that ensures that demand is fulfilled is ok as each customer gets it deliveries from exactly one depot with value 1.
- $\sum_{i \in M} x_{ij} \leq m y_j$: On RHS I get $m \cdot \frac{2}{m} = 2$. On the LHS I get $\sum_{i \in M} x_{ij} = 2$. So that is also ok. So $(x, y) \in P_1$.
- P_2: First constraint is identical to the first constraint in P_1 so we do not need to check that. For the second constraint we take (i, j) where $x_{ij} = 1$, then we get:

  $$1 \leq \frac{2}{m}$$

  Since RHS is ≤ 1 this constraint is violated and so $(x, y) \notin P_2$

Projection

- first formulation: $\min\{cx : x \in P \cap Z^n\}$ with $P \subset R^n$.
- second formulation: $\min\{cx : (x, w) \in Q \cap (Z^n \times R^p)\}$ with $Q \subset R^n \times R^p$.
- Given a polyhedron $Q \subset R^n \times R^p$ the **projection of** Q onto the subspace R^n, denoted $\text{proj}_x Q$ is defined as:

$$\text{proj}_x Q = \{x \in R^n : (x, w) \in Q \text{ for some } w \in R^p\}$$

Take away points from today's lecture

Most important points from the lecture

- You have seen two different models for the TSP problem – one containing an exponential number of constraints and a second that replaces the exponential number of constraints with a extra set of variables.
- The definition of formulation gives us a way of comparing different mathematical models for the same problem.
- If we know the ideal formulation for a problem that enables us to solve the IP problem using linear programming.
- Better formulation does not say anything about how much better one formulation is compared to another, just that it is better.