Week 4

Integer Programming Formulations 2

Integer Programs

Integer programs: a linear program plus the additional constraints that some or all of the variables must be integer valued.

We also permit "$x_j \in \{0,1\}$," or equivalently, "x_j is binary"

This is a shortcut for writing the constraints:

$0 \le x_j \le 1$ and x_j integer.

2

This week's videos

Logical constraints and nonlinear functions

- More on "OR"
- Integer divisibility
- Union of polyhedra
- Fixed Charge Problems
- Piecewise Linear Costs

Combinatorial Problems

- Set covering problems
- Set packing problems
- Facility location problems
- Applications to statistics
- Graph coloring problems

3

More on "OR"

5

Modeling OR: 2 possible values

A variable takes on one of two values.

OR Constraint: $x_1 = 5 \text{ OR } 8$

MIP Constraints:
$x_1 = 5\, w_1 + 8\, (1 - w_1)$
$w_1 \in \{0, 1\}$

6

Modeling OR: 3 or more possible values

A variable takes on one of 3 or more values.

OR Constraint: $x_1 = 5 \text{ OR } 8 \text{ OR } 17 \text{ OR } 99$

MIP Constraints:
$x_1 = 5\, w_1 + 8\, w_2 + 17\, w_3 + 99\, w_4$
$w_1 + w_2 + w_3 + w_4 = 1$
$w_j \in \{0, 1\}$ for $j = 1, 2, 3, 4$

7

Modeling OR: $x = 0$ OR $L \leq x \leq U$

A variable is 0 or it is between two bounds.

"OR Constraint": $x_2 = 0$ OR $4 \leq x_2 \leq 9$.

Equivalent constraint: if $x_2 > 0$, then $4 \leq x_2 \leq 9$

MIP Constraints: $4\, w_5 \leq x_2 \leq 9\, w_5$

$w_5 \in \{0, 1\}$

8

Integer divisibility constraints

9

Modeling divisibility constraints

x_1 is even; x_2 is odd;

Add variables y_1 and y_2 that are required to be integers.

$x_1 = 2\, y_1$

$x_2 = 2\, y_2 + 1$

$y_j \geq 0$ integer for j = 1, 2

$x_j \geq 0$ for j = 1, 2

$y_1 , y_2 \in \mathbb{Z} \Rightarrow$
$x_1 , x_2 \in \mathbb{Z}$.

We don't need to add the constraint that x_j is integer.

10

Modeling non-divisibility constraints

x_1 is not divisible by 7;

We introduce two variables y_1 and y_2 that are required to be integers.

$x_1 = 7\, y_1 + y_2$

$1 \leq y_2 \leq 6$

$y_1 \geq 0$

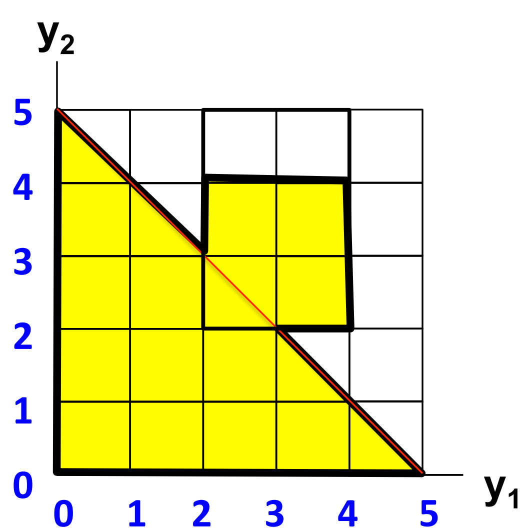
y_1, y_2 integer for j = 1, 2

We could have also written "y_2 = 1 or 2 or 3 or 4 or 5 or 6" and then modeled the OR constraint with six binary variables.

11

Modeling a union of polyhedra

Modeling union of polyhedra



$y_1 + y_2 \leq 5$

$y_1 \geq 0,\ y_2 \geq 0$

Polyhedron 1

OR

$2 \leq y_1 \leq 4$

$2 \leq y_2 \leq 4$

Polyhedron 2

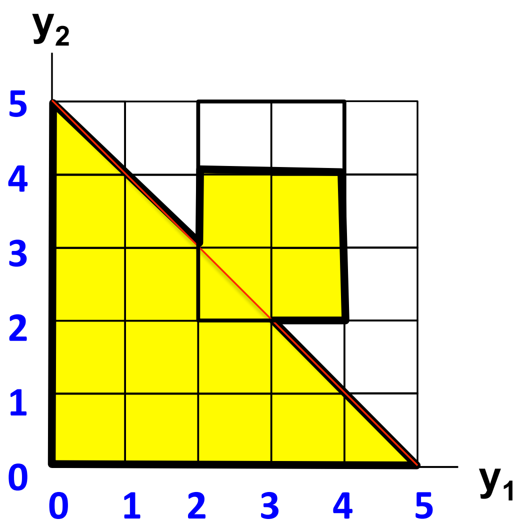
Note first that

$0 \leq y_1 \leq 5$

$0 \leq y_2 \leq 5$

Modeling union of polyhedra

$0 \le y_1 \le 5$
$0 \le y_2 \le 5$



$y_1 + y_2 \le 5$
$y_1 \ge 0,\ y_2 \ge 0$

OR

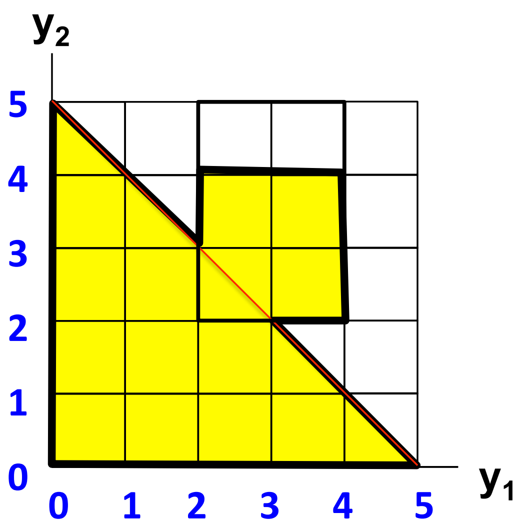
$2 \le y_1 \le 4$
$2 \le y_2 \le 4$

$$y_1 + y_2 \le 5 + 5(1 - w_1)$$
$$2 - 2\, w_1 \le y_1 \le 4 + 1\, w_1$$
$$2 - 2\, w_1 \le y_2 \le 4 + 1\, w_1$$
$$0 \le y_1 \le 5,\ \ 0 \le y_2 \le 5,$$
$$w_1 \in \{0,1\}$$

16

Modeling union of polyhedra

$0 \le y_1 \le 5$
$0 \le y_2 \le 5$



$y_1 + y_2 \le 5$
$y_1 \ge 0,\ y_2 \ge 0$

OR

$2 \le y_1 \le 4$
$2 \le y_2 \le 4$

$$y_1 + y_2 \le 5 + M_1(1 - w_1)$$
$$2 - M_2\, w_1 \le y_1 \le 4 + M_3\, w_1$$
$$2 - M_4\, w_1 \le y_2 \le 4 + M_5\, w_1$$
$$0 \le y_1 \le 5,\ \ 0 \le y_2 \le 5,$$
$$w_1 \in \{0,1\}$$

17

Two exercises

18

Exercises: Model the following inclusive OR constraints (5 minutes)

$0 \le x_1 \le 10$

$0 \le x_2 \le 15$

$0 \le x_3 \le 6$

$x_1 - x_2 \le 1$ OR $3x_1 - x_2 \le 5$ (1)

$x_1 \ge 7$ OR $x_2 \ge 10$ OR $x_1 + x_2 \ge 15$ (2)

[$1 \le x_1 \le 4$ AND $2 \le x_2 \le 5$] OR [$3 \le x_1 \le 7$ AND $4 \le x_2 \le 9$] (3)

If you have time, put in the least values for the M's.

19

Solution to exercises

$0 \le x_1 \le 10$

$0 \le x_2 \le 15$

$0 \le x_3 \le 6$

$$\begin{aligned} x_1 - x_2 &\le 1 + M_1 w_1 \\ 3x_1 - x_2 &\le 5 + M_2 (1 - w_1) \\ w_1 &\in \{0,1\} \end{aligned} \quad (1)$$

$$\begin{aligned} x_1 &\ge 7 - M_1 (1-w_1) \\ x_2 &\ge 10 - M_2 (1-w_2) \\ x_1 + x_2 &\ge 15 - M_3 (1-w_3) \\ w_1 + w_2 + w_3 &= 1 \\ w_j &\in \{0,1\} \text{ for } j = 1, 2, 3 \end{aligned} \quad (2)$$

$$\begin{aligned} 1 - M_1 w_1 &\le x_1 \le 4 + M_2 w_1 \\ 2 - M_3 w_1 &\le x_2 \le 5 + M_4 w_1 \\ 3 - M_5 (1- w_1) &\le x_1 \le 7 + M_6 (1- w_1) \\ 4 - M_7 (1- w_1) &\le x_2 \le 9 + M_8 (1- w_1) \end{aligned} \quad (3)$$

20

Fixed charge problems

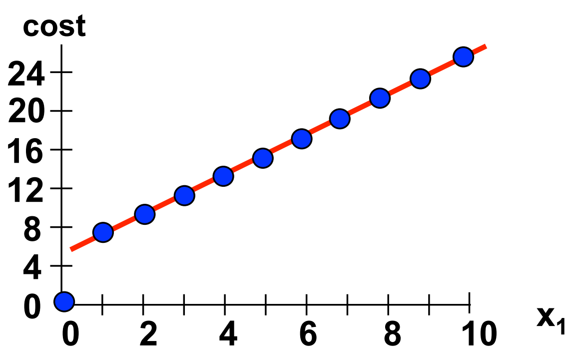
21

Fixed charge problems

- In a *fixed charge problem*, in order to produce a positive amount, the production process must be set up first, at a fixed cost.
- After the setup, the cost of production is linear.
- Note: you are permitted to pay for the set up even if production is 0, even though this would not be optimal.

22

Modeling fixed charges



$y_1 = 0$ if $x_1 = 0$

$y_1 = 4 + 4x_1$ if $x_1 > 0$

Assume $0 \le x_1 \le 100$

No cost for no production

Fixed cost for setting up production facilities
or a fixed cost for having goods delivered or ….

23

Modeling the fixed charge

Introduce a 0-1 variable w_1.

$w_1 = 1$ if there is a setup,
$= 0$ otherwise

$w_1 = 0 \Rightarrow x_1 = 0$

Assume $0 \le x_1 \le 100$

$$
\begin{array}{ll}
\text{Min} & y_1 + \ldots \\
\text{s.t.} & y_1 = 4w_1 + 4x_1 \\
 & 0 \le x_1 \le 100\, w_1 \\
 & \ldots \\
 & w_1 \in \{0,1\}
\end{array}
$$

24

A subtlety in modeling fixed charges

$$
\begin{array}{ll}
\text{s.t.} & y_1 = 4w_1 + 4x_1 \\
 & 0 \le x_1 \le 100\, w_1 \\
 & w_1 \in \{0,1\}
\end{array}
$$

It is permitted that $w_1 = 1$ and $x_1 = 0$.
But it is never optimal.
We do not need to forbid this possibility.

25

The alchemist's problem

26

The Alchemist's Problem

In 1502, the alchemist Zor Primal has set up shop creating gold, silver, and bronze medallions to celebrate the 10th anniversary of the discovery of America. His trainee alchemist (TA) makes the medallions out of lead and pixie dust. Here is the data table.

	Gold	Silver	Bronze	Available
TA labor (days)	2	4	5	100
lead (kilos)	1	1	1	30
pixie dust (grams)	10	5	2	204
Profit ($)	52	30	20	

Zor is unable to get any of his reactions going without an expensive set up.

Cost to set up	$500	$400	$300	

27

The LP Formulation without fixed costs

$$\begin{array}{ll} \text{Max} & 52\,x_1 + 30\,x_2 + 20\,x_3 \\ \text{s.t.} & 2\,x_1 + 4\,x_2 + 5\,x_3 \le 100 \\ & 1\,x_1 + 1\,x_2 + 1\,x_3 \le 30 \qquad \Rightarrow \; x_j \le 30 \text{ for all } j. \\ & 10\,x_1 + 5\,x_2 + 2\,x_3 \le 204 \\ & x_1 \ge 0\,; \quad x_2 \ge 0; \quad x_3 \ge 0; \end{array}$$

28

The IP Formulation with fixed costs

Gold	\$500
Silver	\$400
Bronze	\$300

$w_1 = 0 \Rightarrow x_1 = 0$

$w_2 = 0 \Rightarrow x_2 = 0$

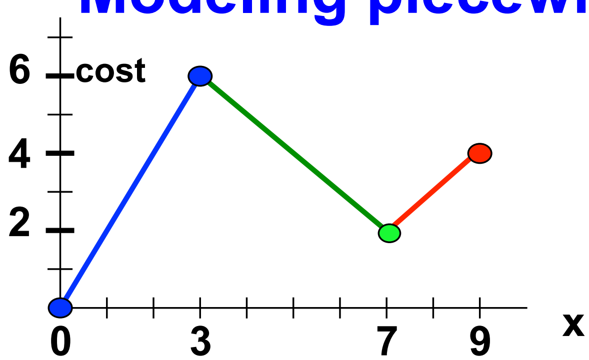
$w_3 = 0 \Rightarrow x_3 = 0$

$$\begin{array}{ll} \text{Max} & (-500\,w_1 + 52\,x_1) + (-400\,w_2 + 30\,x_2) + (-300\,w_3 + 20\,x_3) \\ \text{s.t.} & 2\,x_1 + 4\,x_2 + 5\,x_3 \le 100 \\ & 1\,x_1 + 1\,x_2 + 1\,x_3 \le 30 \\ & 10\,x_1 + 5\,x_2 + 2\,x_3 \le 204 \\ & 0 \le x_1 \le 30\,w_1; \quad 0 \le x_2 \le 30\,w_2; \quad 0 \le x_3 \le 30\,w_3; \\ & x_1, x_2, x_3 \text{ integer} \quad w_1, w_2, w_3 \in \{0,1\}. \end{array}$$

29

Modeling piecewise linear costs: an important special case of nonlinear costs.

Modeling piecewise linear functions.



$y = 2x$ if $0 \le x \le 3$

$y = 9 - x$ if $3 \le x \le 7$

$y = -5 + x$ if $7 \le x \le 9$

There are many ways of modeling this.
All of them seem complicated.

We use the following type of constraint:
if $x > 0$, then $L \le x \le U$.

Modeling three pieces

$y = 2x$ if $0 \le x \le 3$
$y = 9 - x$ if $3 \le x \le 7$
$y = -5 + x$ if $7 \le x \le 9$

Using a previous modeling trick: we create nine new variables. We create 0-1 variables w_1, w_2, w_3, and we create x_1, x_2, x_3, and y_1, y_2, y_3.

$w_1 + w_2 + w_3 = 1$;

if $w_1 = 1$, then $x = x_1$ and $y = y_1$, (and the other 6 variables are 0);

if $w_2 = 1$, then $x = x_2$ and $y = y_2$, (and the other 6 variables are 0);

if $w_3 = 1$, then $x = x_3$ and $y = y_3$, (and the other 6 variables are 0);

Therefore, $x = x_1 + x_2 + x_3$ and $y = y_1 + y_2 + y_3$

32

The formulation

$y = 2x$ if $0 \le x \le 3$
$y = 9 - x$ if $3 \le x \le 7$
$y = -5 + x$ if $7 \le x \le 9$

if $w_1 = 1$, then	if $w_2 = 1$, then	if $w_3 = 1$, then
$0 \le x_1 \le 3$	$3 \le x_2 \le 7$	$7 \le x_3 \le 9$
$y_1 = 2x_1$	$y_2 = 9 - x_2$	$y_3 = -5 + x_3$
if $w_1 = 0$, then	if $w_2 = 0$, then	if $w_3 = 0$, then
$x_1 = 0$	$x_2 = 0$	$x_3 = 0$
$y_1 = 0$	$y_2 = 0$	$y_3 = 0$

$w_1 + w_2 + w_3 = 1$; $x = x_1 + x_2 + x_3$; $y = y_1 + y_2 + y_3$

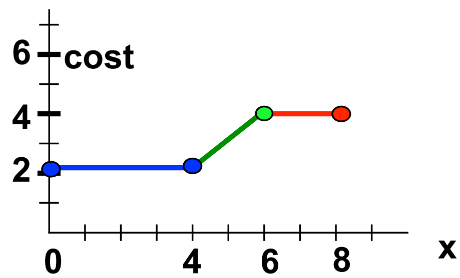
$0 \le x_1 \le 3w_1$
$y_1 = 2x_1$

$3w_2 \le x_2 \le 7w_2$
$y_2 = 9w_2 - x_2$
$w_j \in \{0, 1\}$ for $j = 1, 2, 3$

$7w_3 \le x_3 \le 9w_3$
$y_3 = -5w_3 + x_3$

Two exercises (5 minutes)

1. In the formulation for the alchemist's problem, we used the fact that $x_1 \le 30$ in the constraint "$0 \le x_1 \le 30\, w_1$". Find a better upper bound on x_1 and modify the constraint.
2. Model the following piecewise linear function y.



$y = 2$	if $0 \le x \le 4$
$y = 2 + x$	if $4 \le x \le 6$
$y = 4$	if $6 \le x \le 8$

34

Solution for the fixed charge problem

Max $(-500\, w_1 + 52\, x_1) + (-500\, w_2 + 30\, x_2) + (-300\, w_3 + 20\, x_3)$

s.t.

$$2\, x_1 + 4\, x_2 + 5\, x_3 \le 100$$

$$1\, x_1 + 1\, x_2 + 1\, x_3 \le 30$$

$$10\, x_1 + 5\, x_2 + 2\, x_3 \le 204$$

$$0 \le x_1 \le 30\, w_1; \quad 0 \le x_2 \le 30\, w_2; \quad 0 \le x_3 \le 30\, w_3;$$

x_1, x_2, x_3 integer $\quad w_1, w_2, w_3 \in \{0,1\}$.

$x_1 \le 20.4$

Since x_1 is integer, $x_1 \le 20$.

$0 \le x_1 \le 20\, w_1;$

35

The formulation

$$\begin{array}{lll} y = 2 & \text{if } 0 \le x \le 4 \\ y = 2 + x & \text{if } 4 \le x \le 6 \\ y = 4 & \text{if } 6 \le x \le 8 \end{array}$$

$w_1 + w_2 + w_3 = 1$; $x = x_1 + x_2 + x_3$; $y = y_1 + y_2 + y_3$

$0 \le x_1 \le 4w_1$ $\quad$ $4w_2 \le x_2 \le 6w_2$ $\quad$ $6w_3 \le x_3 \le 8w_3$

$y_1 = 2w_1$ $\quad$ $y_2 = 2w_2 + x_2$ $\quad$ $y_3 = 4w_3$

$w_j \in \{0, 1\}$ for $j = 1, 2, 3$

Overview of combinatorial problems

37

Combinatorial optimization and integer programming

In combinatorial optimization, the goal is to find an optimal object from a finite set of objects.

38

Choosing an optimal subset

- **The knapsack problem**
- **Set covering problems**
- **Set packing problems**
- **Facility location problems**

39

Choosing an optimal partition

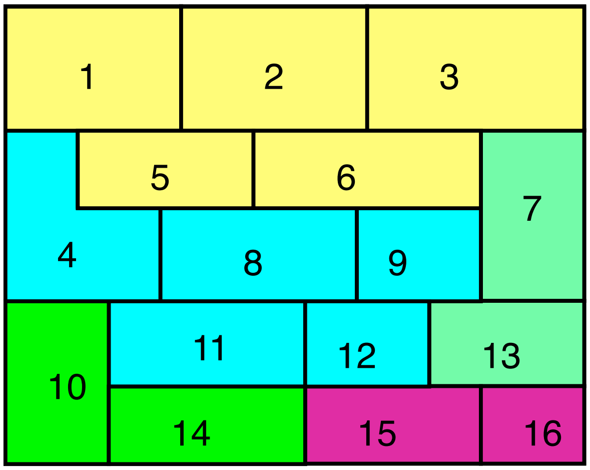
- **Graph coloring**

40

Set covering problems

41

053 Chocolates



Locate 053 Chocolate stores so that each district has a store in it or next to it.

Minimize the number of stores needed.

How do we represent this as an IP?

$$x_j = \begin{cases} 1 & \text{if store in district } j \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} \min \quad & \sum_{j=1}^{n} x_j \\ \text{s.t.} \quad & \sum_{j:i \in S_j} x_j \geq 1 \quad \text{for } i = 1,\ldots,m \\ & x_j \in \{0,1\} \quad \text{for } j = 1,\ldots,n \end{aligned}$$

A store covers its own district plus districts of neighbors.

Let S_j = districts covered by store j.

e.g., $S_2 = \{1, 2, 3, 5, 6\}$.

$i \in S_j$. Store j covers district i.

43

Set covering problems

- There are n binary variables, $x_1, \ldots, x_n$.
- There are m constraints. We let $\mathbf{S} = \{1, \ldots, m\}$, with one element for each constraint.
- For each variable x_j, there is a subset $S_j \subseteq \mathbf{S}$.
  If $x_j = 1$, and if $i \in S_j$, then i is *covered*
  (i.e., constraint i is satisfied).
  - $S_j \subseteq \mathbf{S}$ specifies which constraints are satisfied by letting $x_j = 1$.
  - It is possible that constraint i is covered (satisfied) in many ways.
  - It is permissible to cover constraint i more than once.

44

Set covering problem.

$$\text{Min} \quad \sum_{j=1}^{n} x_j$$

$$\text{s.t.} \quad \sum_{j:\, i \in S_j} x_j \geq 1 \qquad \text{for all } i = 1 \text{ to } m$$

$$x_j \in \{0,1\} \qquad \text{for all } j = 1 \text{ to } n$$

- Binary variables
- RHS: all 1's
- Coefficients in constraints: all 0's and 1's

45

Applications of the set cover problem

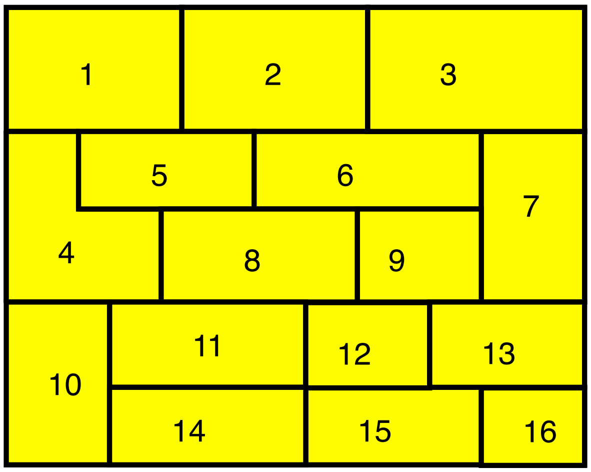
- Locating fire stations, or hospitals, or ambulances.
- Collecting information about a set of items from as few databases as possible.
- Selecting persons for a small start-up so that there is at least one person with each necessary skill.
- Identifying computer viruses (IBM example)
  - Out of 5000 viruses, they identified 9000 strings of 20 bytes not found in good code.
  - They found a set cover of 180 strings. This made searching easier.

46

Set packing problems

47

Set Packing Problems



Locate as many 053 Chocolate stores as possible so that no two stores are in adjacent districts.

How many stores can you find?

48

Set packing problems

- **There are n possible items to choose to pack. For each item j, there is a binary variables x_j.**
  - **x_j = 1 if the item is selected for packing, x_j = 0 otherwise.**
- **We say that items i and j are incompatible if one cannot select both of them. (They cannot both be in a packing.)**
  - **We let A denote the set of incompatible pairs.**
  - **(i, j) $\in$ A $\Rightarrow$ it is not feasible to have x_i = 1 and x_j = 1.**

Set Packing Problem: **find a maximum size (or max value) packing of items.**

49

Set packing as an integer program

$$x_j = \begin{cases} 1 & \text{if item } j \text{ is selected} \\ 0 & \text{otherwise} \end{cases}$$

$$\max \quad \sum_{j=1}^{n} x_j$$

$$\text{s.t.} \quad x_i + x_j \leq 1 \quad \text{if } (i, j) \in A.$$

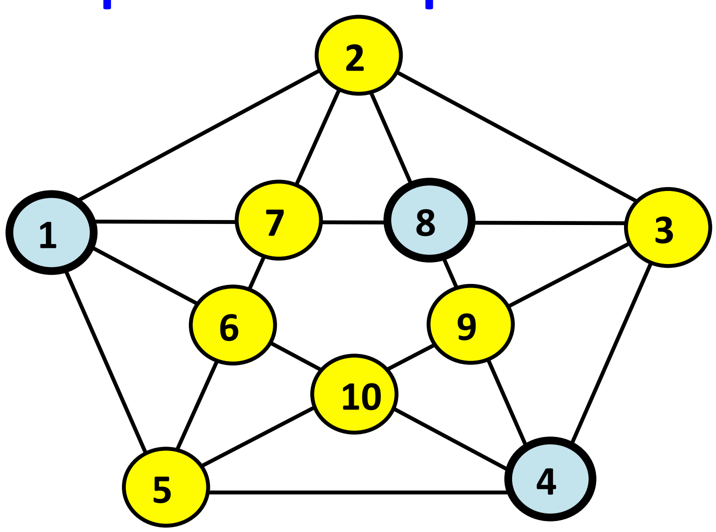
$$x_j \in \{0,1\} \quad \text{for } j = 1, \ldots, n$$

50

Also known as independent set problem

Two vertices are independent if they are not adjacent.

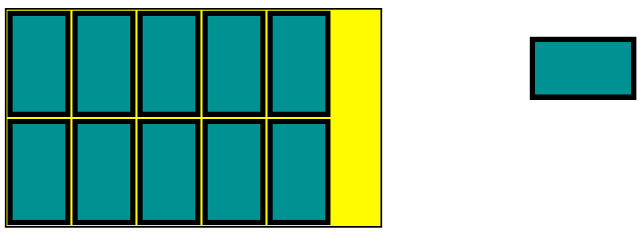
The independent set problem is to find a maximum size subset of (pairwise) independent vertices.



51

Applications of the set packing problem

- In an exam, no two students should sit "too close." How many students can be "packed" into an exam room?
- Combinatorial auction problem (on problem set)
- Cutting doors (for manufacturing of cars) out of large metal sheets.



52

These are not set packing problems
(But they are other types of packing problems)

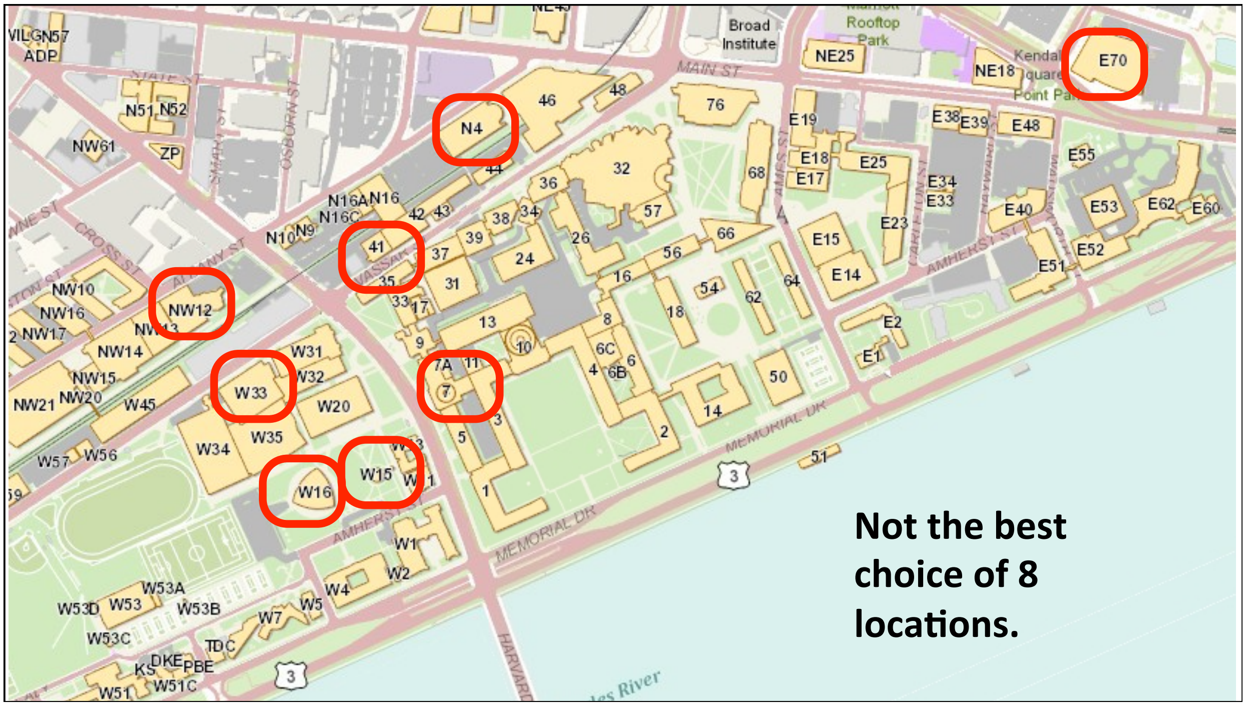
- The knapsack problem.
- Packing items into a suitcase.
- Minimize the number of bins that can hold a collection of items.

53

Facility location problems

A facility location problem.

Suppose that MIT wants to choose where to locate 8 dining facilities on campus. What is the best choice so as to minimize the cost of locating and running the facilities plus the walking inconvenience for students?



Modeling the facility location problem

Sets

- S = locations of "customers"; that is, students.
- D = possible locations for dining facilities

Decision variables

- y_d = 1 if a dining facility is placed at location $d \in D$.
- x_{sd} = # of students at location $s \in S$ who dine at $d \in D$.

Data

- f_d = cost of dining facility at $d \in D$.
- u_d = max number of students who can dine at $d \in D$.
- n_s = number of students at location $s \in S$.
- c_{sd} = disutility per student at $s \in S$ for dining at $d \in D$.

Aspects of this facility location problem

Sets

Two different objectives

Model student behavior

An IP model of the facility location problem

The objective function

$$\begin{aligned} Min \quad & F_1 + F_2 \\ s.t. \quad & F_1 = \sum_{d \in D} f_d y_d; \\ & F_2 = \sum_{s \in S} \sum_{d \in D} c_{sd} x_{sd}. \end{aligned}$$

F_1 = cost of facilities

F_2 = disutility for students.

An alternative: let p be a percentage between 0 and 100.

$$Min \quad pF_1 + (100 - p)F_2$$

e.g., 10% F_1 + 90% F_2

or 50% F_1 + 50% F_2

or any other choice for p.

Modeling the facility location problem

$$\begin{aligned} \text{Min} \quad & \sum_{d \in D} f_d y_d + \sum_{s \in S} \sum_{d \in D} c_{sd} x_{sd} \\ \text{s.t} \quad & \sum_{d \in D} y_d = 8 \\ & \sum_{s \in S} x_{sd} \le u_d y_d \quad \text{for } d \in D. \\ & \sum_{d \in D} x_{sd} = n_s \quad \text{for } s \in S \\ & y_d \in \{0,1\} \ \forall d \in D, \\ & x_{sd} \ge 0, \text{integer} \ \forall s \in S, d \in D. \end{aligned}$$

"forcing constraint"

min total cost

choose 8 facilities

x_{sd} = 0 if y_d = 0

& facility capacity

each student at s dines

Uncapacitated facilities

- Suppose that in the previous model, there was no explicit bound on how many students could dine at a facility.
- How would the model change?

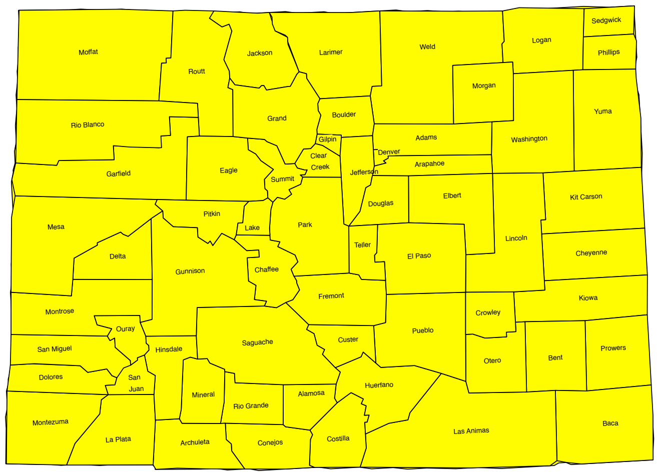
$$\sum_{s \in S} x_{sd} \leq M y_d \quad \text{for } d \in D.$$

62

Map and graph coloring problems

63

Map Coloring

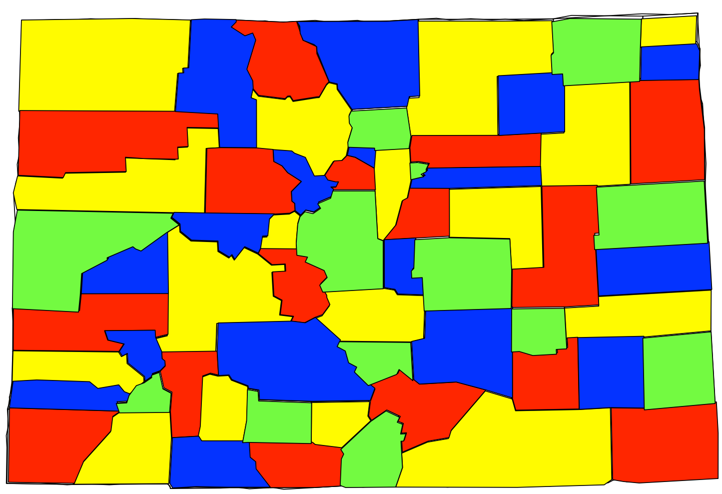


What is the fewest number of colors need to color all of the counties so that no counties with a common border have the same color?

The counties in Colorado.

64

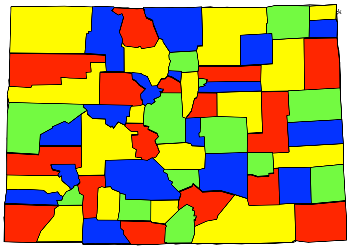
Map Coloring



Here is a four coloring of the map.

65

Graph Coloring Problem



$G = (N, A)$

$N = \{1, 2, 3, \dots n\}$ set of counties.

$A = (i, j) \in A$ if counties i and j are adjacent.

Graph coloring problem: color nodes of N with as few colors as possible so that adjacent nodes have different colors.

$$y_k = \begin{cases} 1 & \text{if color } k \text{ is used} \\ 0 & \text{if color } k \text{ is not used} \end{cases}$$

$$x_{ik} = \begin{cases} 1 & \text{if county } i \text{ is given color } k \\ 0 & \text{otherwise} \end{cases}$$

66

The Integer Programming Formulation

Min $\sum_{k=1}^{C} y_k$ — Minimize the number of colors that are used.

s.t $\sum_{k=1}^{C} x_{ik} = 1 \quad \forall i \in N$ — Each county is given a color.

$x_{ik} + x_{jk} \leq 1 \quad \forall (i, j) \in A,$ and $\forall k = 1$ to C — If counties i and j share a boundary, then they are not both assigned color k.

$x_{ik} \leq y_k \quad \forall i \in N, \forall k$ — Forcing constraint

$x_{ik} \in \{0, 1\} \qquad y_k \in \{0, 1\}$

67

An Exam Scheduling Problem (coloring)

The University of Waterloo has to schedule 500 exams in 28 exam periods so that there are no exam conflicts.

G = (N, A)

N = {1, 2, 3, … n} set of exams. 28 periods.

A = set of arcs. (i, j) ∈ A if exams i and j conflict.

$$x_{ik} = \begin{cases} 1 & \text{if exam } i \text{ is assigned in period } k \\ 0 & \text{otherwise} \end{cases}$$

$$x_{ik} + x_{jk} \leq 1 \quad \text{for } (i, j) \in A \text{ and } \forall k \in [1, 28]$$

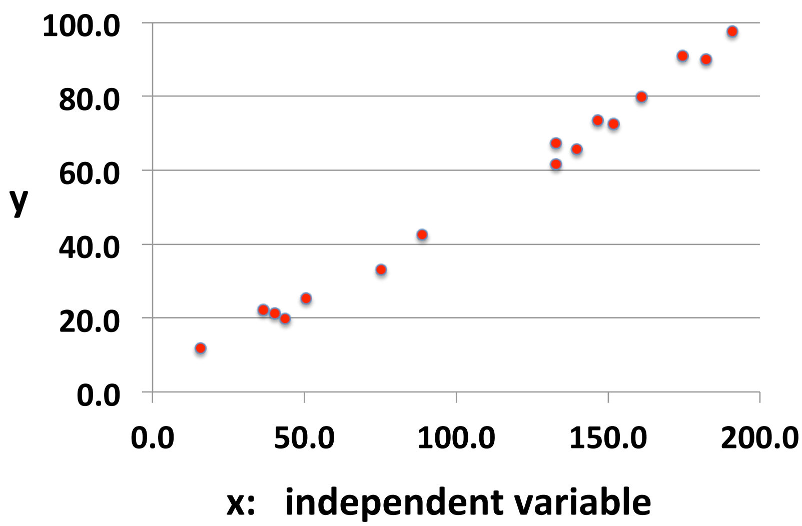
Two exams conflict if they cannot be scheduled in the same period.

Equivalently, can the exam conflict graph be colored with 28 colors?

Linear regression: applications of optimization to statistics.

70

Simple linear regression



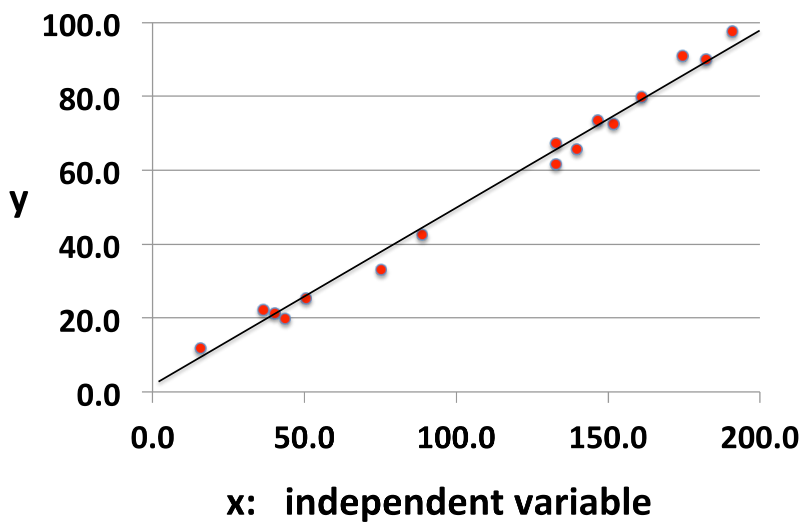
71

Examples

y: dependent variable	x: independent variable
Number of persons taking MOOC for a grade	Number of persons signed up for a MOOC
Grade on homework	Minutes spent on homework
Age at death for smokers	number of cigarettes smoked per day
GPA in college	GPA in high school
Degrees Centigrade	Degrees Fahrenheit

72

Simple linear regression



Estimate y as a function of x.

Choose β_0 and β_1 and estimate y as follows:

$$\hat{y} = \beta_0 + \beta_1 x$$

$$\min \sum_{j=1}^{N} (y_i - \hat{y}_i)^2$$

73

Linear regression with 2 independent variables

Data

1 dependent variable y

2 independent variables x_1, x_2

N observations of the data: the i-th observation is:

$$y_i, x_{i1}, x_{i2}$$

Decision variables

β_0 = constant in linear regression

β_1, β_2 = coefficients of 1st and 2nd covariate

Estimate: $\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$

$$\text{minimize} \sum_{j=1}^{N} (y_i - \hat{y}_i)^2$$

Multiple linear regression

Data

1 dependent variable y
p independent variables $x_1, x_2, \ldots, x_p$
N observations of the data: the i-th observation is:

$$y_i, x_{i1}, x_{i2}, x_{i3}, \ldots, x_{iN}$$

Decision variables

β_0 = constant in linear regression
β_i = coefficient of i-th covariate

Estimate: $\hat{y}_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik}$

minimize $\sum_{j=1}^{N} (y_i - \hat{y}_i)^2$

What if p is very large?

e.g., How much will you like a given movie, based on every other movie that you have ever watched or rated?

e.g., Use of genetic data (there are over 20,000 genes)

e.g., Estimate the number of bicycles at a given location at the end of the day, based on Hubway data.

e.g., Predict whether a person will buy a product based on all products the person has purchased recently.

Regression will appear to be good but it is called "overfitting."

An IP based approach when p is large

- At most K (e.g., 10) non-zero coefficients for β
- Assume all non-zero coefficients are between -100 and 100.

Define w_j as follows:

$$w_j = \begin{cases} 1 & \text{if } \beta_j \neq 0 \\ 0 & \text{if } \beta_j = 0 \end{cases}$$

$$\min \quad \sum_{j=1}^{N} (y_i - \hat{y}_i)^2$$

$$\text{s.t.} \quad \hat{y}_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik}$$

$$\sum_{j=1}^{p} w_j \leq K$$

$$-100 w_j \leq \beta_j \leq 100 w_j \quad \text{for } j = 1 \text{ to } p$$

$$w_j \in \{0,1\} \quad \text{for } j = 1 \text{ to } p.$$

77

Exercise (3 minutes)

- In the previous problem, we bounded the number of non-zero coefficients.
- Introduce new binary variables (call them v_j) and write a constraint that the number of negative coefficients of β is at most 5.

78

Solution

Define v_j as follows:

$$v_j = \begin{cases} 1 & \text{if } \beta_j < 0 \\ 0 & \text{if } \beta_j \geq 0 \end{cases}$$

$$-100 w_j \leq \beta_j \leq 100 w_j \quad \text{for } j = 1 \text{ to } p$$

$$-100 v_j \leq \beta_j \qquad \text{for } j = 1 \text{ to } p$$

$$\sum_{i=1}^{p} v_j \leq 5$$

$$w_j \in \{0,1\} \quad \text{for } j = 1 \text{ to } p.$$

$$v_j \in \{0,1\} \quad \text{for } j = 1 \text{ to } p.$$

79

Modeling the facility location problem

$$\text{Min} \quad \sum_{d \in D} f_d y_d + \sum_{s \in S} \sum_{d \in D} c_{sd} x_{sd} + \sum_{s \in S} g_s v_s$$

min total cost

$$\text{s.t} \qquad \sum_{d \in D} y_d = 8$$

choose 8 facilities

"forcing constraint"

$$\sum_{s \in S} x_{sd} \leq u_d y_d \quad \text{for } d \in D.$$

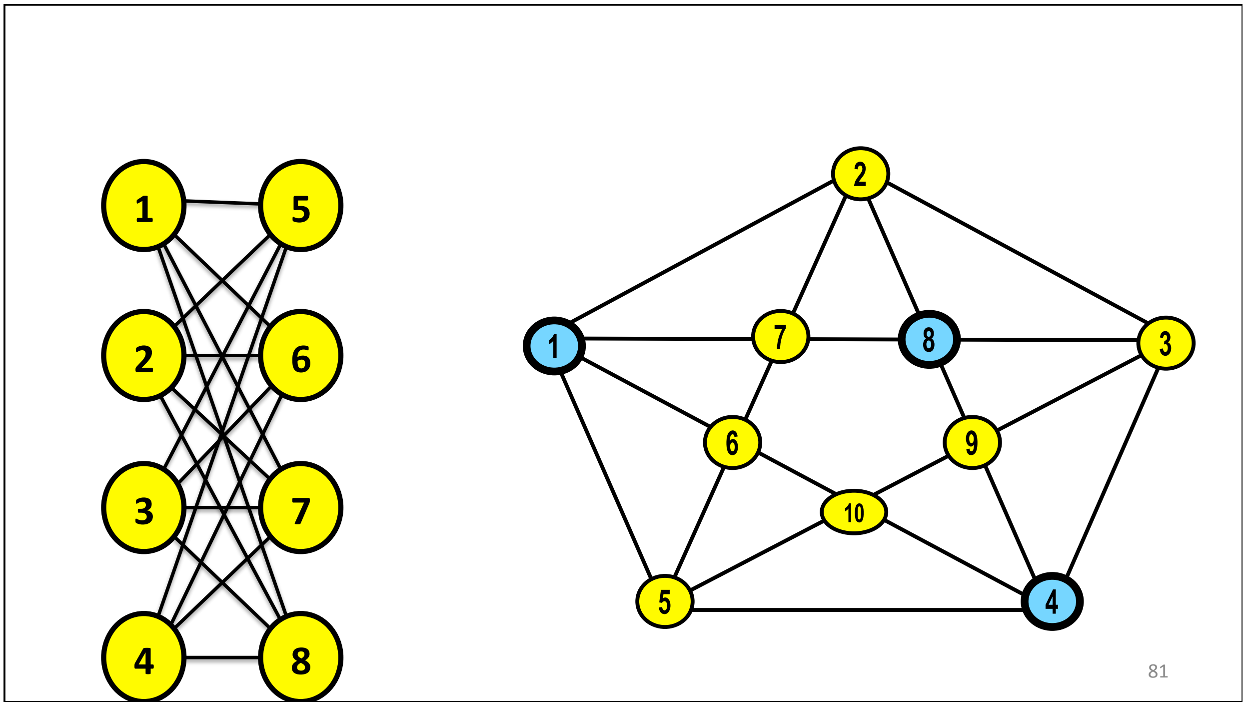
$x_{sd} = 0$ if $y_d = 0$
& facility capacity

$$\sum_{d \in D} x_{sd} + v_s = n_s \quad \text{for } s \in S$$

each student at s dines

$$y_d \in \{0,1\} \ \forall d \in D,$$

$$x_{sd} \geq 0, \text{integer} \quad \forall s \in S, d \in D.$$


1
5
2
6
3
7
4
8
2
1
7
8
3
6
9
10
5
4
81