

CSC373

Week 2: Greedy Algorithms

Announcements

- First tutorial tomorrow!
- Details on Piazza
- First Assignment to be posted tomorrow (May 19) after tutorial
- Due **June 1**

Recap

- Divide & Conquer

- Master theorem
- Counting inversions in $O(n \log n)$
- Finding closest pair of points in \mathbb{R}^2 in $O(n \log n)$
- Fast integer multiplication in $O(n^{\log_2 3})$
- Fast matrix multiplication in $O(n^{\log_2 7})$
- Finding k^{th} smallest element (in particular, median) in $O(n)$

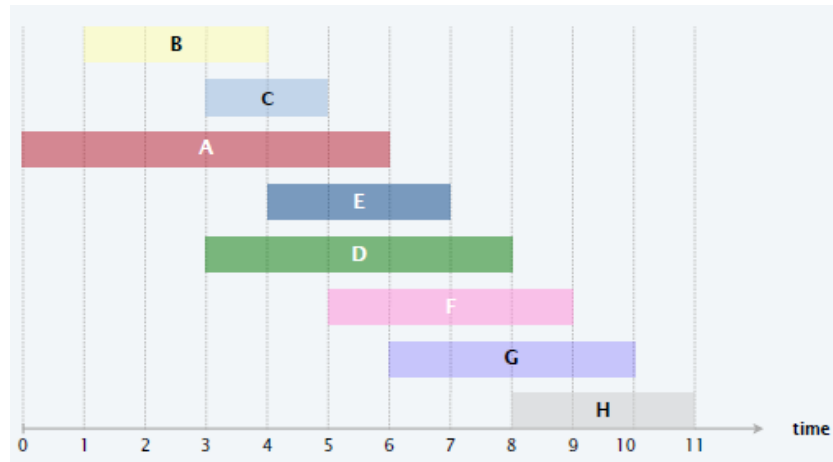
Greedy Algorithms

- Greedy/myopic algorithm outline
 - **Goal:** find a solution x maximizing/minimizing objective function f
 - **Challenge:** space of possible solutions x is too large
 - **Insight:** x is composed of several parts (e.g., x is a set or a sequence)
 - **Approach:** Instead of computing x directly...
 - Compute it one part at a time
 - Select the next part “greedily” to get the most immediate “benefit” (this needs to be defined carefully for each problem)
 - Polynomial running time is typically guaranteed
 - Need to prove that this will always return an optimal solution despite having no foresight

Interval Scheduling

- **Problem**

- Job j starts at time s_j and finishes at time f_j
- Two jobs i and j are compatible if $[s_i, f_i)$ and $[s_j, f_j)$ don't overlap
 - Note: we allow a job to start right when another finishes
- **Goal:** find maximum-size subset of mutually compatible jobs



Interval Scheduling

- Greedy template

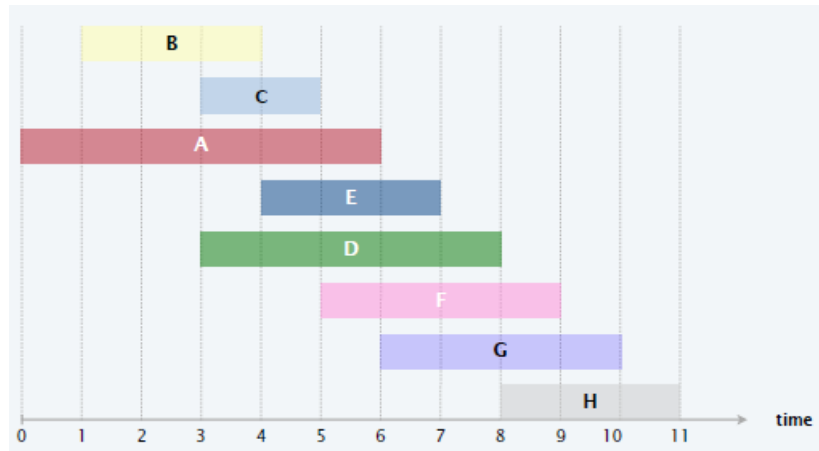
- Consider jobs in some “natural” order
- Take a job if it’s compatible with the ones already chosen

- What order?

- Earliest start time: ascending order of s_j
- Earliest finish time: ascending order of f_j
- Shortest interval: ascending order of $f_j - s_j$
- Fewest conflicts: ascending order of c_j , where c_j is the number of remaining jobs that conflict with j

Example

- **Earliest start time:** ascending order of s_j
- **Earliest finish time:** ascending order of f_j
- **Shortest interval:** ascending order of $f_j - s_j$
- **Fewest conflicts:** ascending order of c_j , where c_j is the number of remaining jobs that conflict with j



Interval Scheduling

- Does it work?



Counterexamples for

earliest start time



shortest interval



fewest conflicts

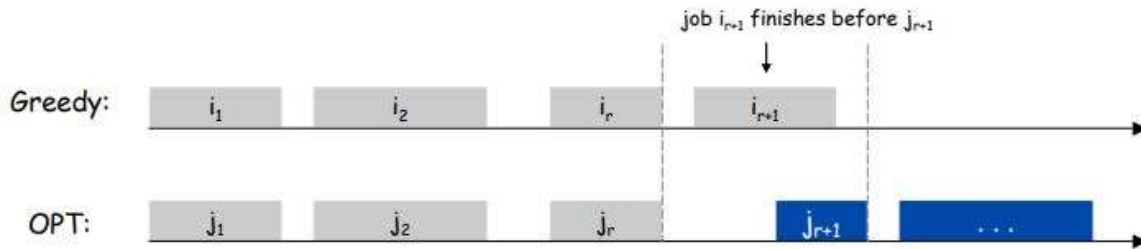
Interval Scheduling

- Implementing greedy with earliest finish time (EFT)
 - Sort jobs by finish time, say $f_1 \leq f_2 \leq \dots \leq f_n$
 - $O(n \log n)$
 - For each job j , we need to check if it's compatible with *all* previously added jobs
 - Naively, this can take $O(n)$ time per job j , so $O(n^2)$ total time
 - We only need to check if $s_j \geq f_{i^*}$, where i^* is the *last added job*
 - For any jobs i added before i^* , $f_i \leq f_{i^*}$
 - By keeping track of f_{i^*} , we can check job j in $O(1)$ time
 - Running time: $O(n \log n)$

Interval Scheduling

- **Proof of optimality by contradiction**

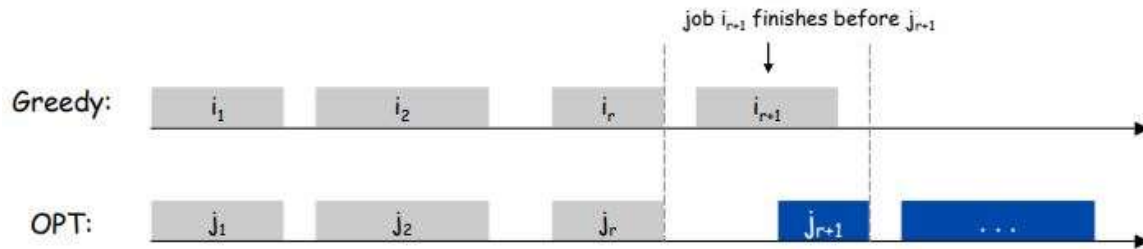
- Suppose for contradiction that greedy is not optimal
- Say greedy selects jobs i_1, i_2, \dots, i_k sorted by finish time
- Consider an optimal solution j_1, j_2, \dots, j_m (also sorted by finish time) which matches greedy for as many indices as possible
 - That is, we want $j_1 = i_1, \dots, j_r = i_r$ for the greatest possible r
- Both i_{r+1} and j_{r+1} must be compatible with the previous selection ($i_1 = j_1, \dots, i_r = j_r$)



Interval Scheduling

- **Proof of optimality by contradiction**

- Consider a new solution $i_1, i_2, \dots, i_r, i_{r+1}, j_{r+2}, \dots, j_m$
 - We have replaced j_{r+1} by i_{r+1} in our reference optimal solution
 - This is still feasible because $f_{i_{r+1}} \leq f_{j_{r+1}} \leq s_{j_t}$ for $t \geq r + 2$
 - This is still optimal because m jobs are selected
 - But it matches the greedy solution in $r + 1$ indices
 - This is the desired contradiction



Interval Scheduling

- **Proof of optimality by induction**

- Let S_j be the subset of jobs picked by greedy after considering the first j jobs in the increasing order of finish time
 - Define $S_0 = \emptyset$
- We call this partial solution **promising** if there is a way to extend it to an optimal solution by picking some subset of jobs $j + 1, \dots, n$
 - $\exists T \subseteq \{j + 1, \dots, n\}$ such that $O_j = S_j \cup T$ is optimal
- **Inductive claim:** For all $t \in \{0, 1, \dots, n\}$, S_t is promising
- If we prove this, then we are done!
 - For $t = n$, if S_n is promising, then it must be optimal (**Why?**)
 - We chose $t = 0$ as our base case since it is “trivial”

Interval Scheduling

- **Proof of optimality by induction**

- S_j is *promising* if $\exists T \subseteq \{j+1, \dots, n\}$ such that $O_j = S_j \cup T$ is optimal
- **Inductive claim:** For all $t \in \{0, 1, \dots, n\}$, S_t is promising
- **Base case:** For $t = 0$, $S_0 = \emptyset$ is clearly promising
 - Any optimal solution extends it
- **Induction hypothesis:** Suppose the claim holds for $t = j - 1$ and optimal solution O_{j-1} extends S_{j-1}
- **Induction step:** At $t = j$, we have two possibilities:
 - 1) Greedy did not select job j , so $S_j = S_{j-1}$
 - Job j must conflict with some job in S_{j-1}
 - Since $S_{j-1} \subseteq O_{j-1}$, O_{j-1} also cannot include job j
 - $O_j = O_{j-1}$ also extends $S_j = S_{j-1}$

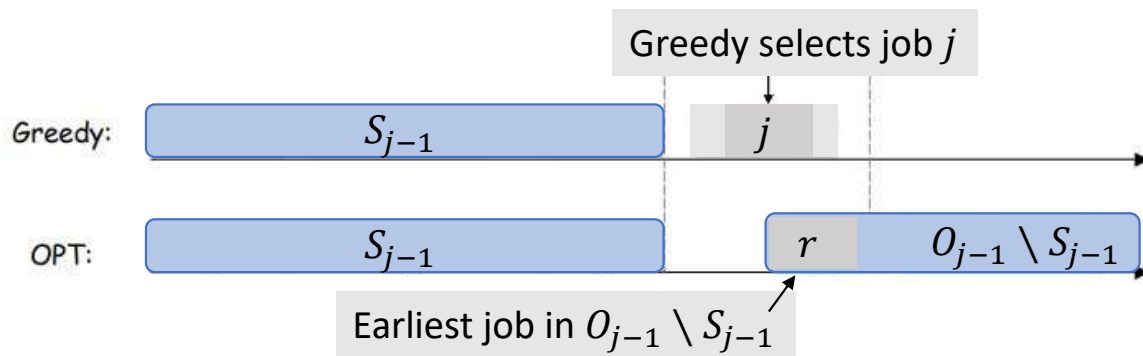
Interval Scheduling

- Proof of optimality by induction

➤ Induction step: At $t = j$, we have two possibilities:

2) Greedy selected job j , so $S_j = S_{j-1} \cup \{j\}$

- Consider the earliest job r in $O_{j-1} \setminus S_{j-1}$
- Consider O_j obtained by replacing r with j in O_{j-1}
- Prove that O_j is still feasible
- O_j extends S_j , as desired!



Contradiction vs Induction

- Both methods make the same claim
 - “The greedy solution after j iterations can be extended to an optimal solution, $\forall j$ ”
- They also use the same key argument
 - “If the greedy solution after j iterations can be extended to an optimal solution, then the greedy solution after $j + 1$ iterations can be extended to an optimal solution as well”
 - For proof by induction, this is the key induction step
 - For proof by contradiction, we take the greatest j for which the greedy solution can be extended to an optimal solution, and derive a contradiction by extending the greedy solution after $j + 1$ iterations

Interval Partitioning

- **Problem**

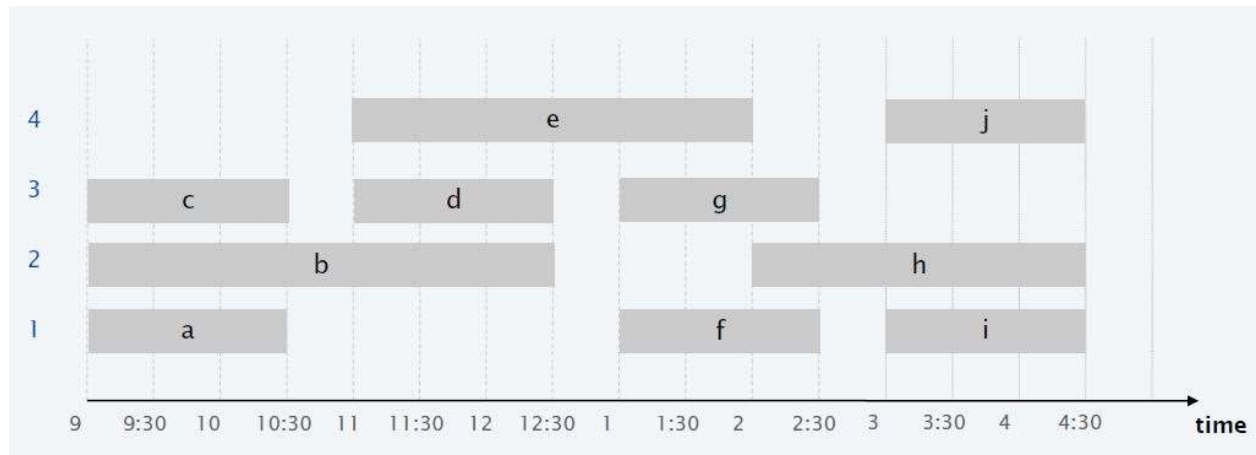
- Job j starts at time s_j and finishes at time f_j
- Two jobs are compatible if they don't overlap
- **Goal:** group jobs into fewest partitions such that jobs in the same partition are compatible

- **One idea**

- Find the maximum compatible set using the previous greedy EFT algorithm, call it one partition, recurse on the remaining jobs.
- Doesn't work (check by yourselves)

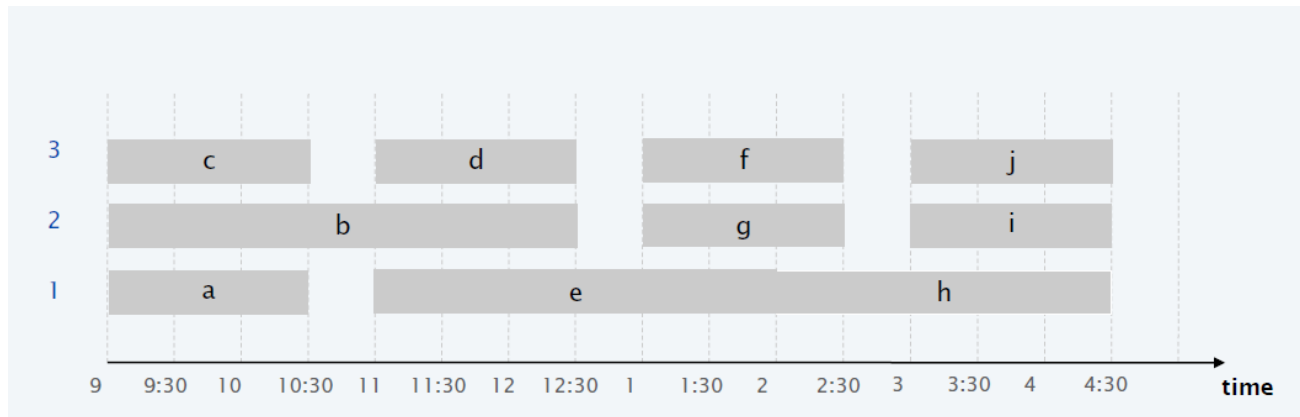
Interval Partitioning

- Think of scheduling lectures for various courses into as few classrooms as possible
- This schedule uses **4** classrooms for scheduling 10 lectures



Interval Partitioning

- Think of scheduling lectures for various courses into as few classrooms as possible
- This schedule uses **3** classrooms for scheduling 10 lectures



Interval Partitioning

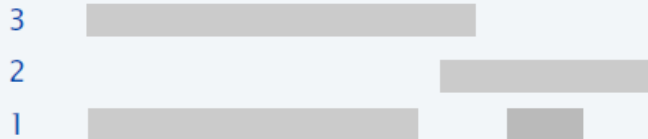
- Let's go back to the **greedy template**!
 - Go through lectures in some “natural” order
 - Assign each lecture to an (**arbitrary?**) compatible classroom, and create a new classroom if the lecture conflicts with every existing classroom
- **Order of lectures?**
 - **Earliest start time**: ascending order of s_j
 - **Earliest finish time**: ascending order of f_j
 - **Shortest interval**: ascending order of $f_j - s_j$
 - **Fewest conflicts**: ascending order of c_j , where c_j is the number of remaining jobs that conflict with j

Interval Partitioning

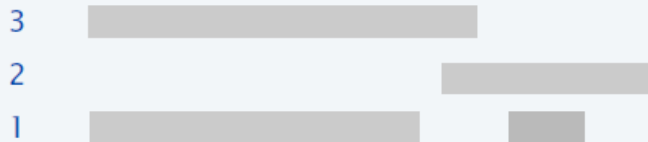
counterexample for earliest finish time



counterexample for shortest interval



counterexample for fewest conflicts



- At least when you assign each lecture to an arbitrary compatible classroom, three of these heuristics do not work.
- The fourth one works! (next slide)

Interval Partitioning

EARLIESTSTARTTIMEFIRST($n, s_1, s_2, \dots, s_n, f_1, f_2, \dots, f_n$)

SORT lectures by start time so that $s_1 \leq s_2 \leq \dots \leq s_n$.

$d \leftarrow 0$  number of allocated classrooms

FOR $j = 1$ **TO** n

IF lecture j is compatible with some classroom

 Schedule lecture j in any such classroom k .

ELSE

 Allocate a new classroom $d + 1$.

 Schedule lecture j in classroom $d + 1$.

$d \leftarrow d + 1$

RETURN schedule.

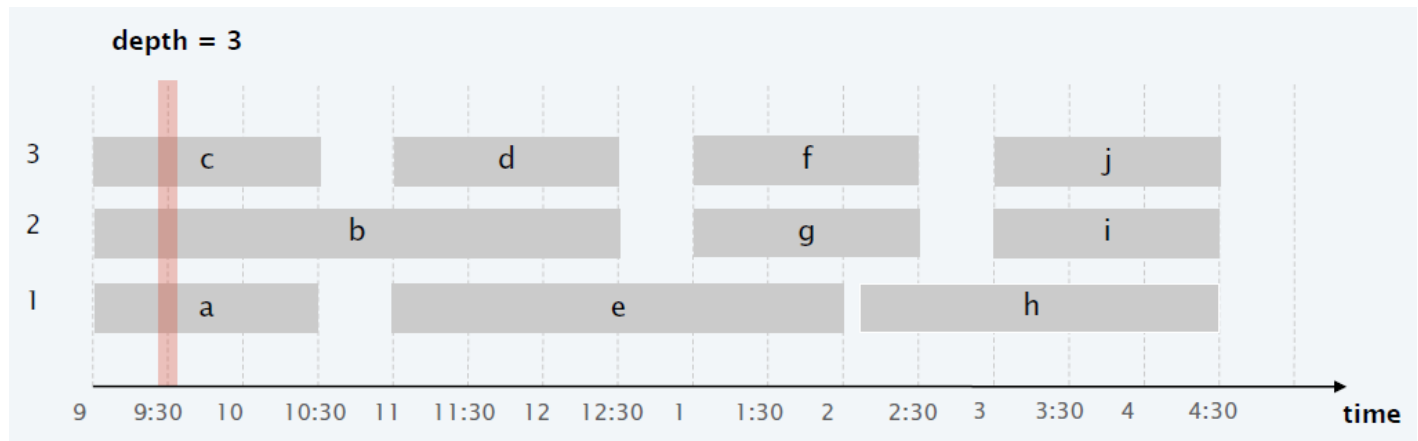
Interval Partitioning

- Running time

- **Key step:** check if the next lecture can be scheduled at some classroom
- Store classrooms in a priority queue
 - key = latest finish time of any lecture in the classroom
- Is lecture j compatible with some classroom?
 - Same as “Is s_j at least as large as the minimum key?”
 - If yes: add lecture j to classroom k with minimum key, and increase its key to f_j
 - Otherwise: create a new classroom, add lecture j , set key to f_j
- $O(n)$ priority queue operations, $O(n \log n)$ time

Interval Partitioning

- **Proof of optimality (lower bound)**
 - # classrooms needed \geq “depth”
 - depth = maximum number of lectures running at any time
 - Recall, as before, that job i runs in $[s_i, f_i)$
 - Claim: our greedy algorithm uses only these many classrooms!



Interval Partitioning

- **Proof of optimality (upper bound)**

- Let $d = \#$ classrooms used by greedy
- Classroom d was opened because there was a lecture j which was incompatible with some lectures already scheduled in each of $d - 1$ other classrooms
- All these d lectures end after s_j
- Since we sorted by start time, they all start at/before s_j
- So, at time s_j , we have d mutually overlapping lectures
- Hence, $\text{depth} \geq d = \# \text{classrooms used by greedy}$ ■
- Note: before we proved that $\# \text{classrooms used by any algorithm (including greedy)} \geq \text{depth}$, so greedy uses exactly as many classrooms as the depth.

Interval Graphs

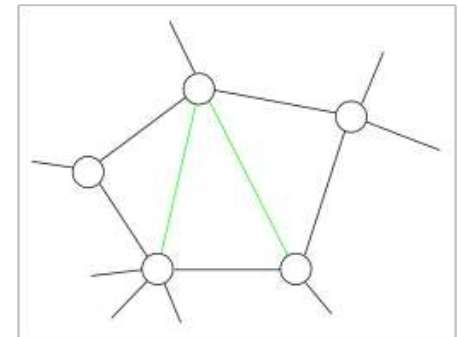
NOT IN SYLLABUS

- Interval scheduling and interval partitioning can be seen as graph problems
- **Input**
 - Graph $G = (V, E)$
 - Vertices V = jobs/lectures
 - Edge $(i, j) \in E$ if jobs i and j are incompatible
- Interval scheduling = **maximum independent set (MIS)**
- Interval partitioning = **graph coloring**

Interval Graphs

NOT IN SYLLABUS

- MIS and graph coloring are NP-hard for general graphs
- But they're efficiently solvable for “**interval graphs**”
 - Graphs which can be obtained from incompatibility of intervals
 - In fact, this holds even when we are not given an interval representation of the graph
- Can we extend this result further?
 - Yes! Chordal graphs
 - Every cycle with 4 or more vertices has a chord



Minimizing Lateness

- **Problem**

- We have a single machine
- Each job j requires t_j units of time and is due by time d_j
- If it's scheduled to start at s_j , it will finish at $f_j = s_j + t_j$
- Lateness: $\ell_j = \max\{0, f_j - d_j\}$
- **Goal:** minimize the maximum lateness, $L = \max_j \ell_j$

- Contrast with interval scheduling

- We can decide the start time
- There are soft deadlines

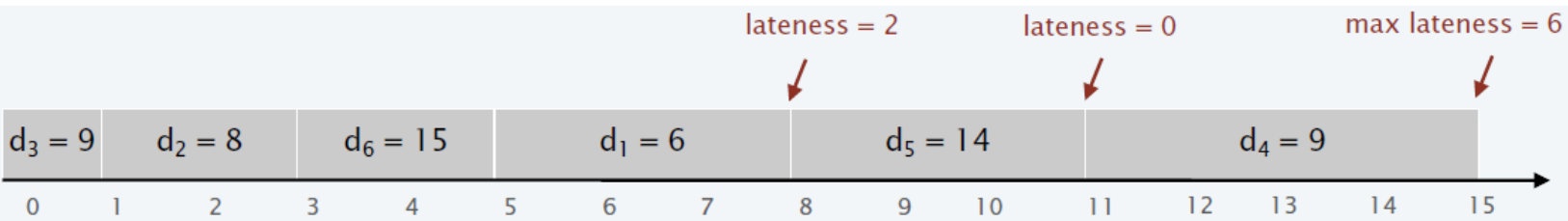
Minimizing Lateness

- Example

Input

	1	2	3	4	5	6
t_j	3	2	1	4	3	2
d_j	6	8	9	9	14	15

An example schedule



Minimizing Lateness

- Let's go back to greedy template
 - Consider jobs one-by-one in some “natural” order
 - Schedule jobs in this order (nothing special to do here, since we have to schedule all jobs and there is only one machine available)
- Natural orders?
 - Shortest processing time first: ascending order of processing time t_j
 - Earliest deadline first: ascending order of due time d_j
 - Smallest slack first: ascending order of $d_j - t_j$

Minimizing Lateness

- Counterexamples

- Shortest processing time first
 - Ascending order of processing time t_j
- Smallest slack first
 - Ascending order of $d_j - t_j$

	1	2
t_j	1	10
d_j	100	10

	1	2
t_j	1	10
d_j	2	10

Minimizing Lateness

- By now, you should know what's coming...

EARLIESTDEADLINEFIRST($n, t_1, t_2, \dots, t_n, d_1, d_2, \dots, d_n$)

SORT n jobs so that $d_1 \leq d_2 \leq \dots \leq d_n$.

$t \leftarrow 0$

FOR $j = 1$ TO n

 Assign job j to interval $[t, t + t_j]$.

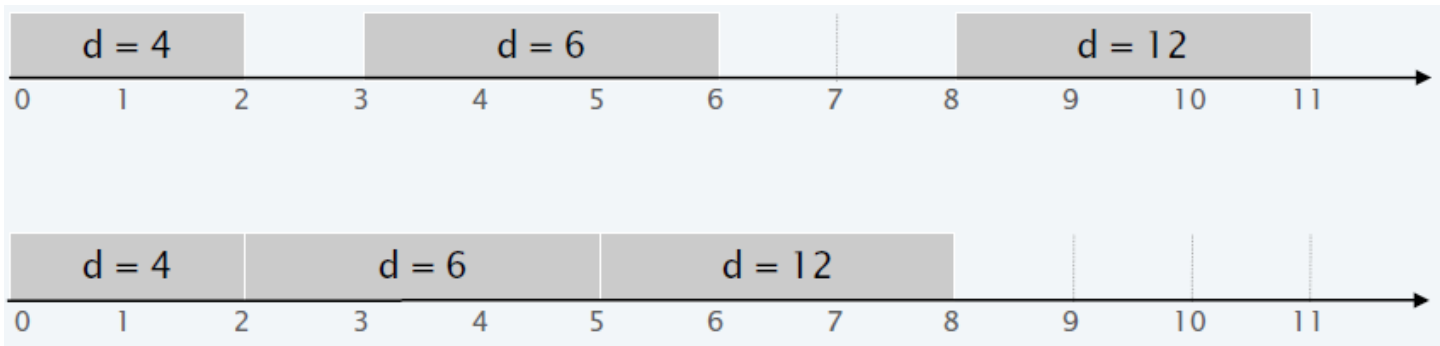
$s_j \leftarrow t$; $f_j \leftarrow t + t_j$

$t \leftarrow t + t_j$

RETURN intervals $[s_1, f_1], [s_2, f_2], \dots, [s_n, f_n]$.

Minimizing Lateness

- Observation 1
 - There is an optimal schedule with **no idle time**



Minimizing Lateness

- Observation 2
 - Earliest deadline first has no idle time
- Let us define an “inversion”
 - (i, j) such that $d_i < d_j$ but j is scheduled before i
- Observation 3
 - By definition, earliest deadline first has no inversions
- Observation 4
 - If a schedule with no idle time has at least one inversion, it has a pair of inverted jobs scheduled consecutively

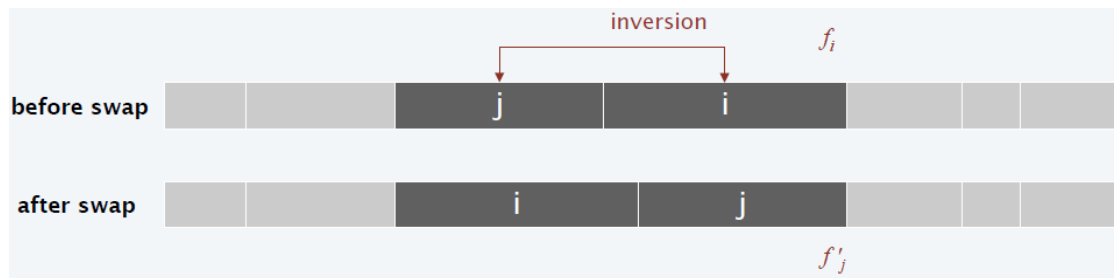
Minimizing Lateness

- Observation 5

- Swapping adjacently scheduled inverted jobs doesn't increase lateness but reduces #inversions by one

- Proof

- Check that swapping an adjacent inverted pair reduces the total #inversions by one



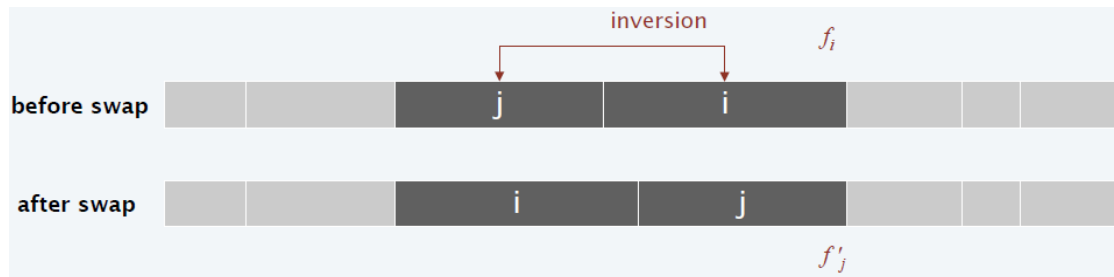
Minimizing Lateness

- Observation 5

- Swapping adjacently scheduled inverted jobs doesn't increase lateness but reduces #inversions by one

- Proof

- Let ℓ_k and ℓ'_k denote the lateness of job k before & after swap
- Let $L = \max_k \ell_k$ and $L' = \max_k \ell'_k$
- 1) $\ell_k = \ell'_k$ for all $k \neq i, j$ (no change in their finish time)
- 2) $\ell'_i \leq \ell_i$ (i is moved early)



Minimizing Lateness

- **Observation 5**

- Swapping adjacently scheduled inverted jobs doesn't increase lateness but reduces #inversions by one

- **Proof**

- 3) $\ell'_j = f'_j - d_j = f_i - d_j \leq f_i - d_i = \ell_i$
 - This uses the fact that, due to the inversion, $d_j \geq d_i$
- $L' = \max\{\ell'_i, \ell'_j, \max_{k \neq i, j} \ell'_k\} \leq \max\{\ell_i, \ell_i, \max_{k \neq i, j} \ell_k\} \leq L$

- **Observation 5**

- Swapping adjacently scheduled inverted jobs doesn't increase lateness but reduces #inversions by one

- **Proof**

- 3) $\ell'_j = f'_j - d_j = f_i - d_j \leq f_i - d_i = \ell_i$
 - This uses the fact that, due to the inversion, $d_j \geq d_i$
- $L' = \max\{\ell'_i, \ell'_j, \max_{k \neq i, j} \ell'_k\} \leq \max\{\ell_i, \ell_i, \max_{k \neq i, j} \ell_k\} \leq L$

Minimizing Lateness

- Observations 4+5 are the key!
- Recall the proof of optimality of the greedy algorithm for interval scheduling:
 - Took an optimal solution matching greedy for r steps, and produced another optimal solution matching greedy for $r + 1$ steps
 - “Wrapped” this in a proof by contradiction or a proof by induction
 - Observations 4+5 provide something similar
 - If optimal solution doesn’t fully match greedy ($\text{\#inversions} \geq 1$), we can swap an adjacent inverted pair and reduce \#inversions by one

Minimizing Lateness

- **Proof of optimality by contradiction**
 - Suppose for contradiction that the greedy EDF solution is not optimal
 - Consider an optimal schedule S^* with the fewest inversions
 - Without loss of generality, suppose it has no idle time
 - Because EDF is not optimal, S^* has at least one inversion
 - By Observation 4, it has an adjacent inversion (i, j)
 - By Observation 5, swapping the adjacent pair keeps the schedule optimal but reduces the #inversions by 1
 - Contradiction! ■

Minimizing Lateness

- **Proof of optimality by (reverse) induction**

- **Claim:** For each $r \in \{0, 1, \dots, \binom{n}{2}\}$, there is an optimal schedule with *at most* r inversions
- **Base case of $r = \binom{n}{2}$:** trivial, any optimal schedule works
- **Induction hypothesis:** Suppose the claim holds for $r = t + 1$
- **Induction step:** Take an optimal schedule with at most $t + 1$ inversions
 - If it has at most t inversions, we're done!
 - If it has exactly $t + 1 \geq 1$ inversions...
 - Assume no idle time WLOG
 - Find and swap an adjacent inverted pair (Observations 4 & 5)
 - #inversions reduces by one to t , so we're done!
- **QED!**
- Claim for $r = 0$ shows optimality of EDF

Contradiction vs Induction

- Choose the method that feels natural to you
- It may be the case that...
 - For some problems, a proof by contradiction feels more natural
 - But for other problems, a proof by induction feels more natural
 - No need to stick to one method
- As we saw for interval partitioning, sometimes you may require an entirely different kind of proof

Lossless Compression

- **Problem**

- We have a document that is written using n distinct labels
- Naïve encoding: represent each label using $\log n$ bits
- If the document has length m , this uses $m \log n$ bits

- English document with no punctuations etc.

- $n = 26$, so we can use 5 bits

- $a = 00000$

- $b = 00001$

- $c = 00010$

- $d = 00011$

- ...


Lossless Compression


- Is this optimal?
 - What if a, e, r, s are much more frequent in the document than x, q, z ?
 - Can we assign shorter codes to more frequent letters?
- Say we assign...
 - $a = 0, b = 1, c = 01, \dots$
 - See a problem?
 - What if we observe the encoding '01'?
 - Is it 'ab'? Or is it 'c'?

Lossless Compression

- To avoid conflicts, we need a *prefix-free encoding*
 - Map each label x to a bit-string $c(x)$ such that for all distinct labels x and y , $c(x)$ is not a prefix of $c(y)$
 - Then it's impossible to have a scenario like this

.....


 $c(x)$


 $c(y)$

- Now, we can read left to right
 - Whenever the part to the left becomes a valid encoding, greedily decode it, and continue with the rest

Lossless Compression

- **Formal problem**

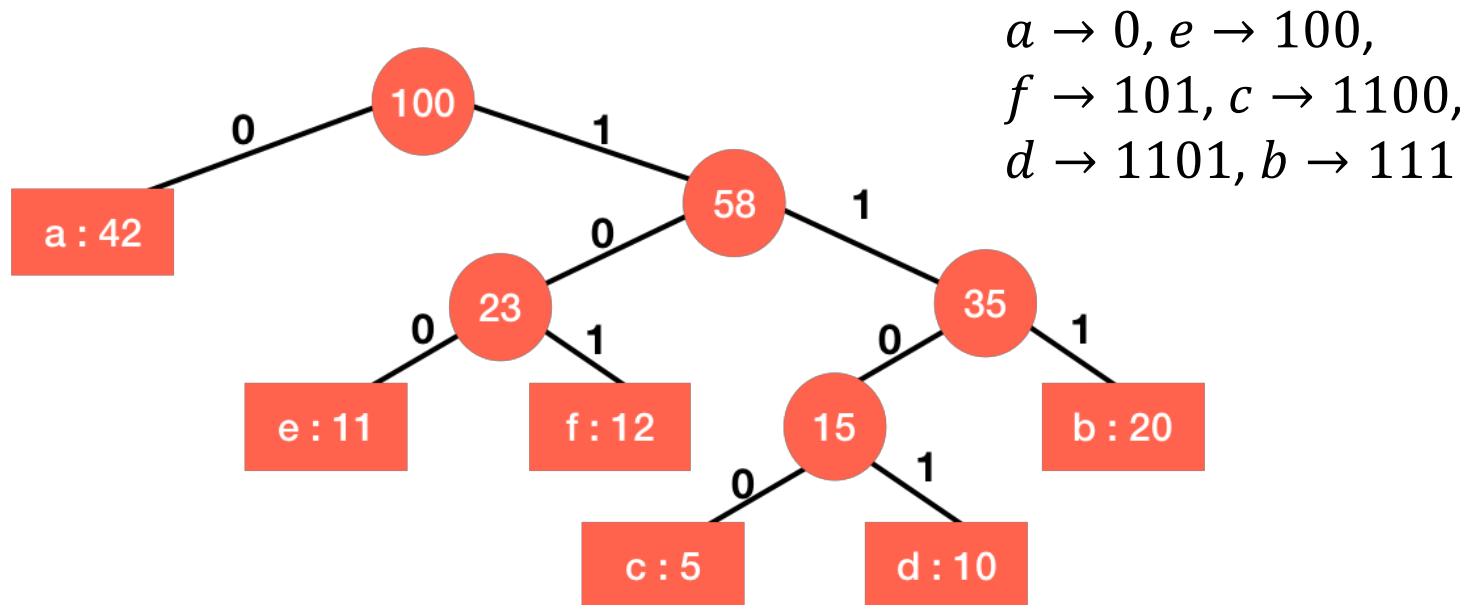
- Given n symbols and their frequencies (w_1, \dots, w_n) , find a prefix-free encoding with lengths (ℓ_1, \dots, ℓ_n) assigned to the symbols which minimizes $\sum_{i=1}^n w_i \cdot \ell_i$
 - Note that $\sum_{i=1}^n w_i \cdot \ell_i$ is the length of the compressed document

- **Example**

- $(w_a, w_b, w_c, w_d, w_e, w_f) = (42, 20, 5, 10, 11, 12)$
- No need to remember the numbers 😊

Lossless Compression

- **Observation:** prefix-free encoding = tree



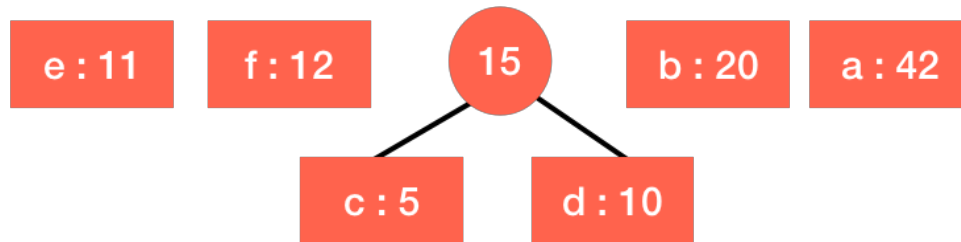
Lossless Compression

- Huffman Coding

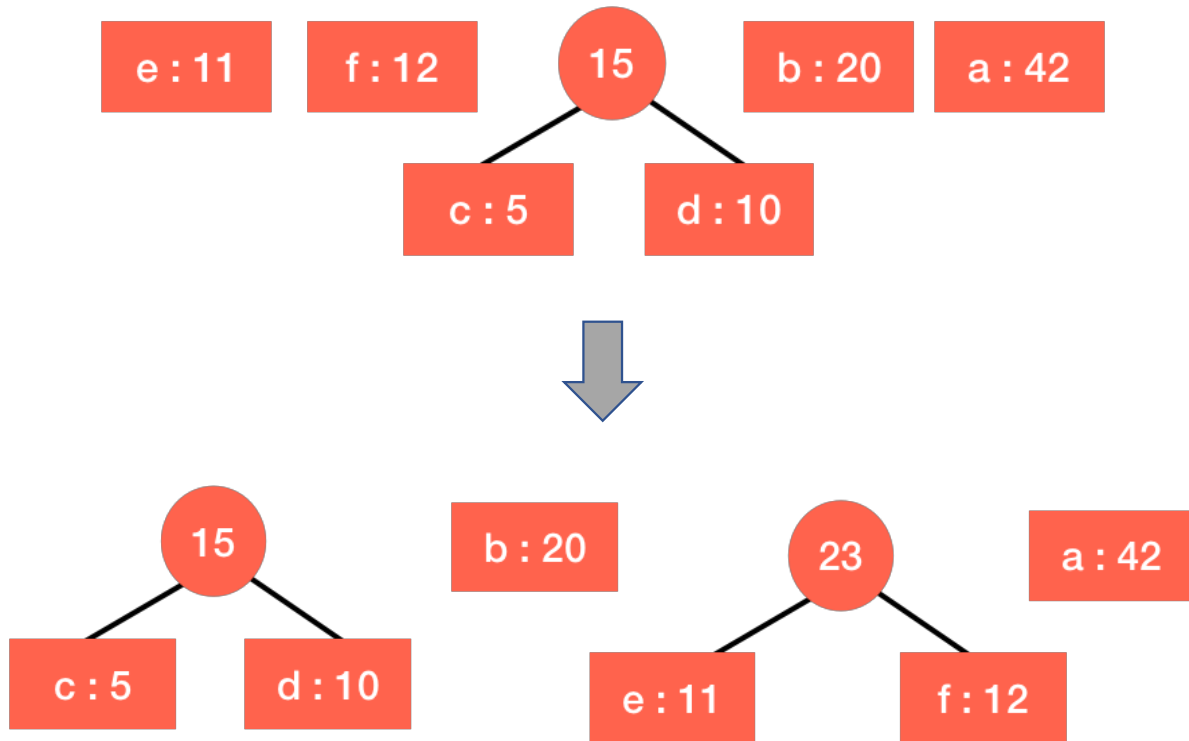
- Build a priority queue by adding (x, w_x) for each symbol x
- While $|\text{queue}| \geq 2$
 - Take the two symbols with the lowest weight (x, w_x) and (y, w_y)
 - Merge them into one symbol with weight $w_x + w_y$

- Let's see this on the previous example

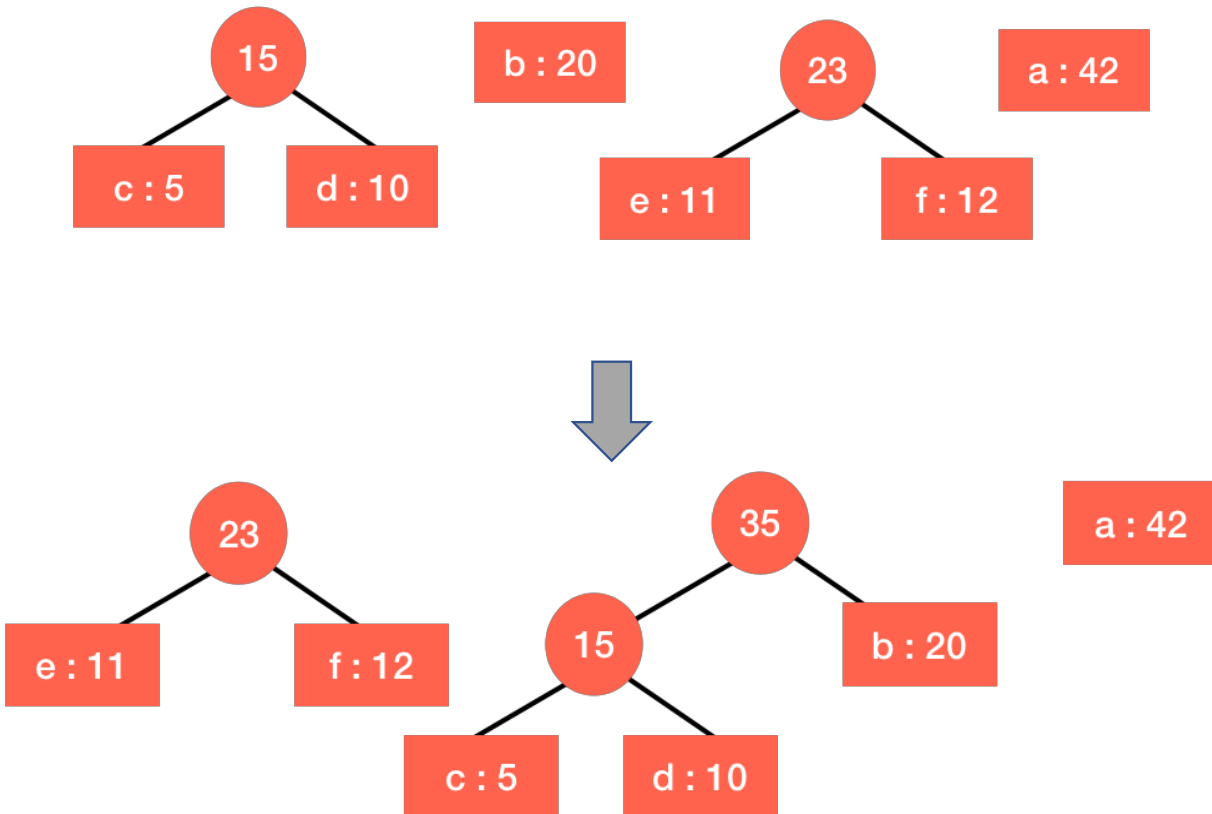
Lossless Compression



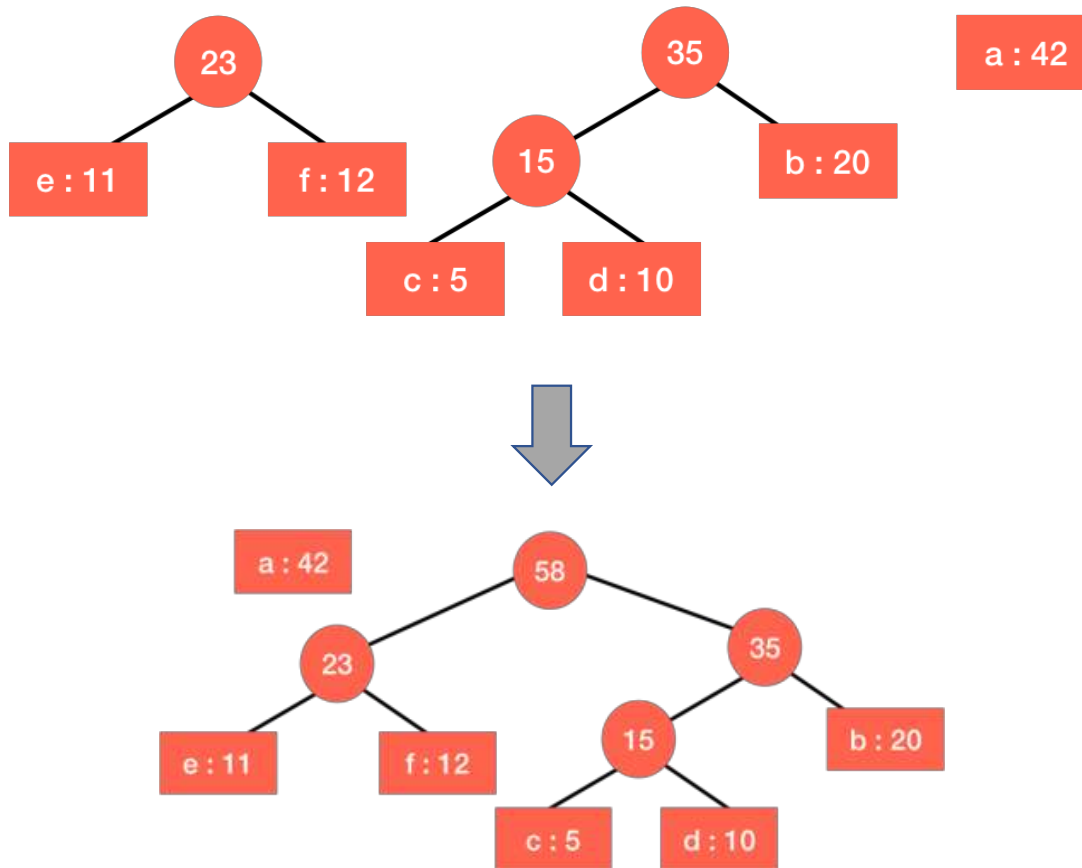
Lossless Compression



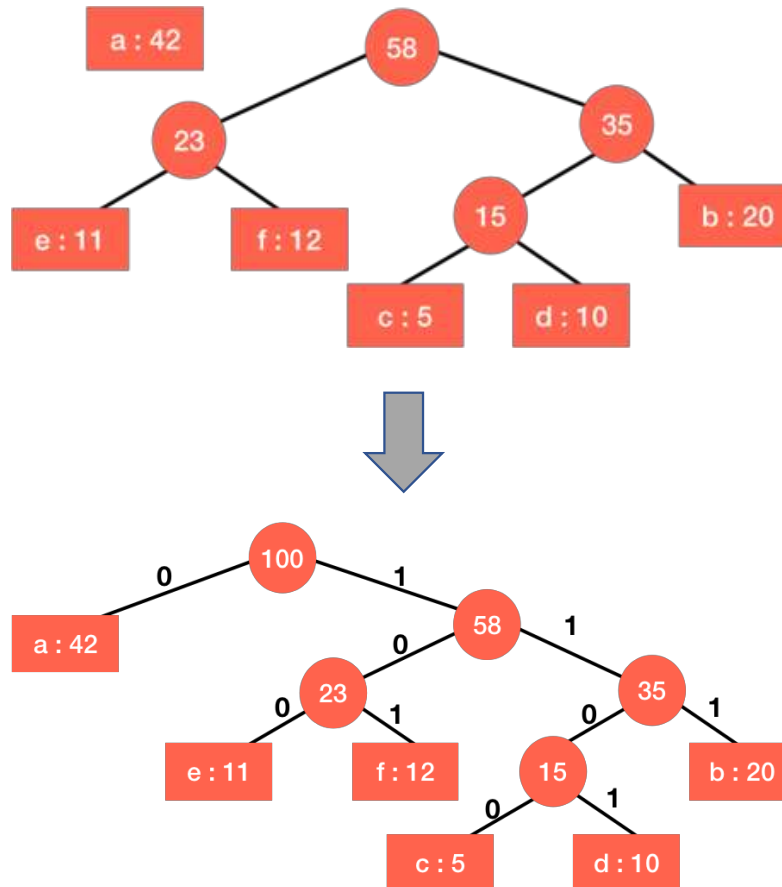
Lossless Compression



Lossless Compression



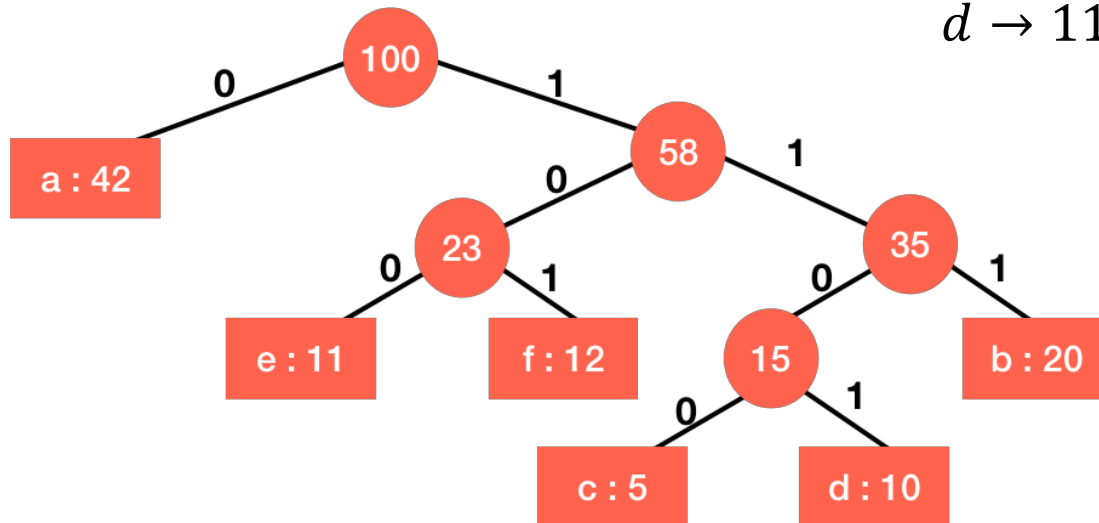
Lossless Compression



Lossless Compression

- Final Outcome

$a \rightarrow 0, e \rightarrow 100,$
 $f \rightarrow 101, c \rightarrow 1100,$
 $d \rightarrow 1101, b \rightarrow 111$



Lossless Compression

- Running time

- $O(n \log n)$
- Can be made $O(n)$ if the labels are given to you sorted by their frequencies
 - Exercise! Think of using two queues...

- Proof of optimality

- Induction on the number of symbols n
- **Base case:** For $n = 2$, both encodings which assign 1 bit to each symbol are optimal
- **Hypothesis:** Assume it returns an optimal encoding with $n - 1$ symbols

Lossless Compression

- **Proof of optimality**

- Consider the case of n symbols

- **Lemma 1:** If $w_x < w_y$, then $\ell_x \geq \ell_y$ in any optimal tree.

- **Proof:**

- Suppose for contradiction that $w_x < w_y$ and $\ell_x < \ell_y$.
 - Swapping x and y strictly reduces the overall length as $w_x \cdot \ell_y + w_y \cdot \ell_x < w_x \cdot \ell_x + w_y \cdot \ell_y$ (check!)
 - QED!

Lossless Compression

- **Proof of optimality**

- Consider the two symbols x and y with lowest frequency which Huffman combines in the first step
- **Lemma 2:** \exists optimal tree T in which x and y are siblings (i.e., for some p , they are assigned encodings $p0$ and $p1$).
- **Proof:**
 1. Take any optimal tree
 2. Let x be the label with the lowest frequency.
 3. If x doesn't have the longest encoding, swap it with one that has
 4. Due to optimality, x must have a sibling (check!)
 5. If it's not y , swap it with y
 6. Check that Steps 3 and 5 do not change the overall length. ■

Lossless Compression

- Proof of optimality

- Let x and y be the two least frequency symbols that Huffman combines in the first step into “ xy ”
- Let H be the Huffman tree produced
- Let T be an optimal tree in which x and y are siblings
- Let H' and T' be obtained from H and T by treating xy as one symbol with frequency $w_x + w_y$
- Induction hypothesis: $Length(H') \leq Length(T')$
- $Length(H) = Length(H') + (w_x + w_y) \cdot 1$
- $Length(T) = Length(T') + (w_x + w_y) \cdot 1$
- So $Length(H) \leq Length(T)$ ■

Other Greedy Algorithms

- If you aren't familiar with the following algorithms, spend some time checking them out!
 - Dijkstra's shortest path algorithm
 - Kruskal and Prim's minimum spanning tree algorithms