Lab Worksheet 8: NP-Completeness (Solution)

Subset Sum

We are given a multiset of integers $Z = \{z_1, \ldots, z_n\}$ and a target $T \in \mathbb{Z}$. The question is whether there exists an index set $I \subseteq \{1, \ldots, n\}$ such that $\sum_{i \in I} z_i = T$. Despite its simple statement, this problem is computationally intractable in general.

Proof of NP-completeness

Our goal is to prove that Subset Sum is NP-complete. The proof has two steps.

1. Show that Subset Sum is in NP.

   A certificate is a subset I. In $O(n)$ time we can sum $\sum_{i \in I} z_i$ and check if it equals T. Hence the problem is in NP.

2. Show Subset Sum is NP-hard via a polynomial-time reduction from the NP-complete problem 3-SAT.

3-SAT Problem. Recal from lecture: Given a 3-CNF formula with variables $x_1, \ldots, x_n$ and clauses $C_1, \ldots, C_m$, each clause having exactly three literals, decide whether there exists a truth assignment satisfying all clauses.

Connecting 3-SAT and Subset Sum. The strength of Subset Sum is that many different constraints can be encoded simultaneously by using *different digit positions* in a large-base number system. Each digit represents one constraint.

For example, suppose we include in our multiset two numbers $z_1 = 1000$ and $z_2 = 1000$, and assume all other z_i are either at least 10000 or together sum to less than 1000. If we make the thousands digit of the target T equal to exactly 1000, then any solution I to the Subset Sum instance must include **exactly one** of z_1 or z_2: including none gives 0 in that digit; including both gives 2000. This ability to force "choose exactly one" is precisely what allows us to reduce 3-SAT to Subset Sum. We can use this structure to ensure literals are selected consistently, meaning that x_i and $\neg x_i$ are in fact not equal.

A second key idea is that we can also allow *multiple acceptable values* in a digit, by adding *slack numbers*. For instance, if a certain digit of the target is 3, and our construction allows a chosen subset to contribute 1, 2, or 3 from the "meaningful" numbers, we can add two slack numbers that each contribute 1 in that digit so that 1, 2, or 3 can all be "completed" up to the target value 3. This flexibility lets us express constraints of the form "at least one of these contributions must be present", because although slack numbers can fill in missing contributions, they can only fill in a limited amount. In the 3-SAT reduction, this is exactly how we ensure that every clause has at least one satisfied literal: each clause digit can be brought up to its target value using at most two slack numbers, so the chosen subset must provide at least one literal-contribution.

These two mechanisms are what make it possible to reduce 3-SAT to Subset Sum.

Construction for the 3-SAT to Subset Sum Reduction

We now outline the structure of the reduction and ask you to complete the missing pieces. Consider a 3-SAT instance with literals$\{x_1, \ldots, x_n, \neg x_1, \ldots, \neg x_n\}$ and clauses $C_1, \ldots, C_m$. We will construct a Subset Sum instance in base $B = 10$.[1]

Digits. Each constructed number will have exactly $n + m$ digits. Each digit encodes one constraint:

- **Variable digits**: the first n digits ensure that for every variable x_i, we choose *exactly one* of x_i or $\neg x_i$ to be true.
- **Clause digits**: the last m digits ensure that each clause has *at least one satisfied literal.*

Numbers. For each literal ℓ, we construct a number $z(\ell)$. Denote

$$z_{2i} := z(x_i), \qquad z_{2i-1} := z(\neg x_i).$$

For each clause C_j, we also create two slack numbers $s_{j,1}$ and $s_{j,2}$. These allow clause digits to reach their target value even when the selected literals contribute less than the target.

Interpretation. We construct a target number T with $n + m$ digits. A subset I is a solution if and only if

$$\sum_{k \in I} z_k = T.$$

If $z(\ell)$ is selected in the sum (its index is in I), then we interpret ℓ as being set to TRUE; otherwise it is FALSE.

[1] A larger base works as well; the crucial requirement is that no carries occur.

Part 1: Enforcing the variable constraint (digit i).

For each $i \in \{1, \dots, n\}$, digit i enforces that: "Exactly one of $z(x_i)$ or $z(\neg x_i)$ is selected". This implies that between x_i and $\neg x_i$ one of them is TRUE and the other one is FALSE.

To enforce this, fill in the i-th digit entries below.

Hint: Try assigning values to z_{2i} and z_{2i-1}. What happens if the subset includes: neither, exactly one, or both? Which of these should match the target digit?

Solution:

Number	**Digit i (variable digit)**
$z(x_i)$	1
$z(\neg x_i)$	1
Any other number	0
Target T	1

Part 2: Enforcing the clause constraint (digit $n + j$).

For each clause $C_j = (\ell_1 \vee \ell_2 \vee \ell_3)$, digit $(n+j)$ ensures: "At least one literal in C_j is selected." Each literal ℓ_r that appears in the clause contributes to digit $(n + j)$. Complete the table:

Hint: Set the i-th digit of T such that selecting **no true literal** makes it impossible to hit the target digit. Using the slack variables if any of the literals are True, it should be possible to hit the target digit.

Solution: Without slack variables, the clause digit might sum to 1, 2, or 3 depending on how many literals of the clause are chosen. But we want a *fixed target digit*, say 3. Slack numbers help "fill in" the difference: by one or two. However, they cannot fill in all three that is needed to hit the target because only two slack numbers exist.

Number	**Digit $n + j$ (clause digit)**
$z(\ell_1)$	1
$z(\ell_2)$	1
$z(\ell_3)$	1
$s_{j,1}$ (slack 1)	1
$s_{j,2}$ (slack 2)	1
Any other number	0
Target T	3

Mini Example: Fill in the literal numbers and the target. We use base $B = 10$ with $n = 2$ variables (x_1, x_2) and a single clause $C_1 = (x_1 \lor x_2 \lor \neg x_2)$. There are $n + m = 3$ digits per number: $(d_1,\ d_2 \mid d_3)$, where d_1 enforces "exactly one of x_1 or $\neg x_1$," d_2 enforces "exactly one of x_2 or $\neg x_2$," and d_3 enforces that C_1 is satisfied (with two slack numbers $s_{1,1}, s_{1,2}$). **Solution:**

Number	d_1	d_2	d_3 (clause C_1)
$z(x_1)$	1	0	1
$z(\neg x_1)$	1	0	0
$z(x_2)$	0	1	1
$z(\neg x_2)$	0	1	1
$s_{1,1}$	0	0	1
$s_{1,2}$	0	0	1
Target T	1	1	3

Given one satisfying assignment here and the corresponding subset-sum solution I:

Assignment: $x_1 = \boxed{\text{T}}$, $x_2 = \boxed{\text{F}}$, $I = \{z(x_1), z(\neg x_2), s_{1,1}\}$

Proof of correctness

Next we focus on the proof of correctness. We show a one-to-one correspondence between satisfying assignments of the 3-SAT instance and solutions to the constructed SUBSET SUM instance.

($\Rightarrow$ direction) Assume the 3-SAT instance is a YES instance; that is, there exists a truth assignment that satisfies all clauses. Show that the constructed Subset Sum instance has a subset I whose numbers sum exactly to the target T.

Solution: For each variable x_i, if $x_i = \text{TRUE}$ we include $z(x_i)$ in I, and if $x_i = \text{FALSE}$ we include $z(\neg x_i)$ in I. Thus, in variable digit i, the sum is exactly 1, matching $T_i = 1$.

Now consider any clause $C_j = (\ell_1 \lor \ell_2 \lor \ell_3)$. Because the assignment satisfies the formula, at least one literal is TRUE. Each TRUE literal contributes 1 in clause digit $(n + j)$. If exactly $r \in \{1, 2, 3\}$ of the clause's literals are true, then we add $3 - r$ of the clause's slack numbers, each contributing 1 in digit $(n + j)$. Since at most two slacks exist, this fully reaches the required $T_{n+j} = 3$.

Thus, in every digit, $\sum_{k \in I} z_k$ matches T, and no carries occur since $B \geq 10$. Therefore the Subset Sum instance is a YES instance.

($\Leftarrow$ direction) Assume the constructed Subset Sum instance admits a subset I such that $\sum_{k \in I} z_k = T$. Prove that the 3-SAT instance is a YES instance.

Solution: Consider digit i (a variable digit). Only $z(x_i)$ and $z(\neg x_i)$ contribute a 1 in this digit. Since $T_i = 1$, the sum in digit i must also be 1. Thus exactly one of $z(x_i)$ or $z(\neg x_i)$ is selected. We define the corresponding truth assignment: x_i = TRUE if $z(x_i) \in I$ and x_i = FALSE if $z(\neg x_i) \in I$.

Next consider clause digit $(n + j)$. The slack numbers contribute at most 2 in this digit. Because $T_{n+j} = 3$, the clause literals must contribute at least 1. Thus at least one literal in clause C_j has its corresponding number selected into I, meaning that literal is TRUE under the assignment. Hence every clause contains at least one TRUE literal, so the truth assignment satisfies the 3-SAT formula.

Conclusion. You have shown that a satisfying assignment of 3-SAT corresponds exactly to a subset summing to the target T, and vice versa. Thus, 3-SAT reduces to Subset Sum, proving the problem is NP-hard. Since Subset Sum is also in NP, it is NP-complete.

Job Scheduling on Identical Machines

You run a small compute cluster with two identical machines. A queue of jobs $J = \{1, \ldots, n\}$ has processing times $p_1, \ldots, p_n$. You want to assign each job to one of the two machines so that the *finishing time* (the time when the last machine stops) is as small as possible. This objective is called the *makespan*. Even this simple-looking problem hides computational hardness.

Decision version of two-machine job scheduling We are given positive integers $p_1, \ldots, p_n$, and a target makespan $K \in \mathbb{Z}_{>0}$. Can we schedule the jobs on two identical machines so that the makespan is at most K? Equivalently, is there a partition of the jobs into two sets A and B such that

$$\sum_{j \in A} p_j \leq K \quad \text{and} \quad \sum_{j \in B} p_j \leq K?$$

Proof of NP-completeness

Our goal is to prove that the decision version of two-machine job scheduling is NP-complete. The proof has two steps.

1. Show that job scheduling on two machines is in NP.

   To prove NP-completeness we need to show a certificate for a YES instance of the problem exists for which we can be verified in polynomial time. If set B and A are provided, we can compute the makespan in $O(n)$ time, and check whether it is at most K or not. Thus, the problem is in NP.

2. The problem is NP-hard via a polynomial-time reduction from the NP-complete problem PARTITION.

Partition Problem. Given positive integers $a_1, \ldots, a_n$ with total sum $S = \sum_i a_i$, decide whether there exists an index set $I \subseteq \{1, \ldots, n\}$ such that $\sum_{i \in I} a_i = S/2$. For the purpose of this worksheet, we assume this problem is NP-complete. This can be proved by showing a polynomial reduction from SUBSET-SUM.

Construct the reduction. Given an instance $(a_1, \ldots, a_n)$ of PARTITION, define the corresponding scheduling instance $(p_1, \ldots, p_n; K)$ for two machines. Specify each p_j and K. Briefly argue that mapping is computable in polynomial time.

Solution: Set $p_j := a_j$ for all j. Let $S = \sum_j a_j$ and set $K := S/2$. We copy the n numbers and compute S (a single pass), then set $K = S/2$. This can be done in linear time in n.

Proof of correctness.

Next we focus on the proof of correctness. We show a one and one correspondence between the solutions of these two problems.

($\Rightarrow$ direction) Assume the PARTITION instance is a YES instance; that is, there exists I with $\sum_{i \in I} a_i = S/2$. Show that the constructed scheduling instance has a schedule of makespan $\leq K$.

Solution: Assign the jobs corresponding to indices in I to machine 1 and the remaining jobs to machine 2. Their loads are $S/2$ and $S/2$, so the makespan is $S/2 = K$.

($\Leftarrow$ direction). Assume the scheduling instance admits a schedule with makespan $\leq K$. Prove that the PARTITION instance is a YES instance.

Solution: Let the two machine loads be L_1 and L_2 with $L_1, L_2 \leq K$. Since $L_1 + L_2 = \sum_j p_j = S$, we must have $L_1 = L_2 = K = S/2$. The set of jobs on either machine then gives a subset summing to $S/2$, which is a valid partition.

Conclusion. You have shown an NP-complete problem (PARTITION) can be reduced to the decision version of job scheduling. Hence, job scheduling is NP-hard. Since the problem is also in NP, it is NP-complete.