CS120: Intro. to Algorithms and their Limitations Hesterberg & Vadhan

Lecture 18: Computational Complexity

Harvard SEAS - Fall 2022 *2022-11-02*

1 Announcements

Recommended Reading:

- MacCormick §5.3–5.5, Ch. 10, 11

2 Computational Complexity

A common category error when discussing computational problems is to talk about the "runtime of the problem", when runtime is a property of an algorithm, not of a problem. For instance, sorting is a problem which we've seen solved by algorithms whose runtimes are $O(n \log n)$ (Merge Sort), $O(n+U)$ (Radix Sort), and $O(n!n)$ (brute force). (Note that in this case there isn't even a single best runtime!)

There is a sense in which we can talk about runtime (or space, or probability of correctness) of a problem: we say that a problem is solvable in time $O(T(n))$ if there exists an algorithm which solves it that quickly. Note that proving that a problem *is* solvable in time $O(T(n))$ is straightforward (give a single algorithm solving it in time $O(T(n))$), but saying that a problem *is not* solvable in time $O(T(n))$ requires knowing that *no* algorithm with runtime $O(T(n))$ solves it, a much harder claim.

Computational complexity aims to classify problems according to the amount of resources (e.g. time) that they require.

For example, we've seen algorithms that are:

- Linear time: Shortest Paths, 2-Coloring in time $O(n+m)$.
- Nearly linear time: Sorting, Interval Scheduling (Decision, Optimization, Coloring) in time $O(n \log n)$.
- Polynomial time: Bipartite Matching in time $O(nm)$, 2-SAT in time $O(n^3)$.[1]
- Exponential time: k-Coloring for $k \geq 3$, k-SAT for $k \geq 3$, Independent Set, and Longest Path in time $O(c^n)$ for constants $c > 1$.

[1] A linear-time algorithm for 2-SAT is actually known, based on DFS (which is covered in CS 124) rather than BFS/Reachability.

To develop a robust and clean theory for classifying problems according to computational complexity, we make two choices:

- A problem-independent size measure. Recall that we allowed ourselves to use different size parameters for different problems (array length n and universe size U for sorting; number n of vertices and number m of edges for graphs, number n of variable and number m of clauses for Satisfiability). To classify problems, it is convenient to simply measure the size of the input by its *length N in bits*. For example:
  - Array of n numbers from universe size U:
  - Graphs on n vertices and m edges in adjacency list notation:
  - 3-SAT formulas with n variables and m clauses:
- Polynomial slackness in running time: We will only try to make coarse distinctions in running time, e.g. polynomial time vs. super-polynomial time. If the Extended Church-Turing Thesis is correct, the theory we develop will be independent of changes in computing technology. It is possible to make finer distinctions, like linear vs. nearly linear vs. quadratic, if we fix a model (like the Word-RAM), and a newer subfield called *Fine-Grained Complexity* does this.

To this end, we define the following *complexity classes*.

Definition 2.1.

- For a function $T : \mathbb{N} \to \mathbb{R}^+$, $\mathsf{TIME}_{\mathsf{search}}(T(N))$ is:

  $\mathsf{TIME}(T(N))$ is

- (Polynomial time)

  $$\mathsf{P}_{\mathsf{search}} = \qquad\qquad\qquad \mathsf{P} = \qquad\qquad\qquad .$$

- (Exponential time)

  $$\mathsf{EXP}_{\mathsf{search}} = \qquad\qquad\qquad \mathsf{EXP} = \qquad\qquad\qquad .$$

(Remark on terminology: what we call $\mathsf{P}_{\mathsf{search}}$ is called Poly in the MacCormick text, and is often called FP elsewhere in the literature.)

By this definition, Shortest Paths, 2-Coloring, Sorting, Interval Scheduling, Bipartite Matching, and 2-SAT are all in $\mathsf{P}_{\mathsf{search}}$ (as well as P for decision versions of the problems). However, all we know to say about 3-Coloring, 3-SAT, Independent Set, or Longest Path is that they are in $\mathsf{EXP}_{\mathsf{search}}$. Can we prove that they are not in $\mathsf{P}_{\mathsf{search}}$?

The following seems to give some hope:

Theorem 2.2.

We won't give a proof of this theorem (take CS 121 for that), but we'll see similar proofs in the last unit of the course.

We even know (again, without proof) an example of a problem in $\mathsf{EXP}_{\mathsf{search}} \setminus \mathsf{P}_{\mathsf{search}}$ (in fact $\mathsf{EXP} - \mathsf{P}$): the problem of deciding whether a Word-RAM program halts on an input x of length n within 2^n steps, called the "Bounded Halting" problem.

Next we might try to obtain more intractable problems via reductions.

Definition 2.3. For computational problems Π and Γ, we write $\Pi \leq_p \Gamma$ if

Some examples of polynomial time reduction that we've seen include:

- GraphColoring $\leq_p$ SAT
- LongPath $\leq_p$ SAT

Lemma 2.4. *Let Π and Γ be computational problems such that $\Pi \leq_p \Gamma$. Then:*

1.

2.

This is the same lemma as we introduced in lecture 3 and recalled yesterday, but keeping track only of whether there are polynomial-time algorithms solving the problems, not more precise runtimes.

Proof. $\square$

So, we have a procedure for proving that problems are not in $\mathsf{P}_{\mathsf{search}}$:

1. Identify a particular problem Π in $\mathsf{EXP}_{\mathsf{search}} \setminus \mathsf{P}_{\mathsf{search}}$. One example is deciding whether a Word-RAM program halts within 2^n steps on an input x of length n.
2. Show that Π reduces to the problems we are interested in, via a *polynomial-time reduction*.

Unfortunately, we don't know how to reduce the problems we know in $\mathsf{EXP}_{\mathsf{search}} \setminus \mathsf{P}_{\mathsf{search}}$ (like Bounded Halting) to many of the problems we care about (like Independent Set, 3-Coloring, and Longest Path), so we can only conjecture that those problems are in $\mathsf{EXP}_{\mathsf{search}} \setminus \mathsf{P}_{\mathsf{search}}$.

So we have many possible worlds:

These problems have additional structure, which will require us to define and study a different complexity class, NP, next time.

Lemma 2.5. *We can compose reductions - if* $\Pi \leq_p \Gamma$ *and* $\Gamma \leq_p \Theta$ *then* $\Pi \leq_p \Theta$.

The proof of this is similar to the proof of Lemma 2.4: run a reduction from Π to Γ, but whenever an oracle call to Γ is made, substitute in the reduction from Γ to Θ.

3 Optional reading: Turing Machines

Most courses on the theory of computation (like CS121) use Turing Machines as their main model of computation, whereas we use the (Word-)RAM model because it better suited for measuring the efficiency of algorithms. However, Turing machines can be understood as a small variant of Word-RAM programs, where we make the word size *constant*:

Definition 3.1 (TM-RAM programs)**.** A *TM-RAM* program P is like a RAM program with the following modifications:

1. *Finite Alphabet:*

2. *Memory Pointer:*

3. *Read/write:*

4. *Moving Pointer:*

See Figure 1

Figure 1: A TM RAM machine, with memory pointer and commands.

Philosophically, TM-RAM programs are appealing because one step of computation only operates on constant-sized objects (ones with domain $[q]$). However, as we will discuss below, the ability to only increment and decrement `mem_ptr` by 1 does make TM-RAM programs somewhat slow compared to Word-RAM programs.

Note that the number of possibilities for the state of a TM-RAM's computation, excluding the memory contents is:

Thus, the computation can be more concisely described as follows:

Definition 3.2 (Turing machine)**.** A *Turing machine* $M = (Q, \Sigma, \delta, q_0, H)$ is specified by:

1. A finite set Q of states.
2. A finite alphabet Σ (e.g. $[q]$).
3. A transition function $\delta : Q \times \Sigma \to Q \times \Sigma \times \{L, R, S\}$.
4. An initial state $q_0 \in Q$.
5. A set $H \subseteq Q$ of halting states.

Semantics of δ:

Theorem 3.3 (Equivalence of TMs and TM-RAMs)**.** *1. There is an algorithm that given TM-RAM program P, constructs a Turing Machine M such that $M(x) = P(x)$ for all inputs x and $\mathrm{Time}_M(x) = O(\mathrm{Time}_P(x))$.*

2. There is an algorithm that given a Turing Machine M, constructs a TM-RAM program P such that $P(x) = M(x)$ for all inputs x and $\mathrm{Time}_P(x) = O(\mathrm{Time}_M(x))$.

Thus Turing Machines are indeed equivalent to a restricted form of RAM programs. The appeal of Turing machines is their mathematically simple description, with no arbitrary set of operations being chosen (allowing any "constant-sized" computation to happen in one step).

What about Turing Machines vs. Word-RAM Programs?

Theorem 3.4. *There is an algorithm that given a Word-RAM Program P constructs a TM-RAM program P' such that $P'(x) = P(x)$ for all inputs x and*

$$\mathrm{Time}_{P'}(x) =$$

provided that $\mathrm{Time}_P(x)$ is at least $n \cdot \max_i x[i]$ for an input array x of length n.

Figure 2: The requirement to move the memory pointer step by step in TM-RAM induces an up to quadratic slowdown vs RAM.

Proof Sketch.

□

So TM-RAMs and Turing Machines can simulate Word-RAM programs, but with a bit more than a quadratic slowdown in runtime. This is a lot better than the relation between RAM programs and Word-RAM programs, which incurs an exponential slowdown in simulating the former by the latter (as demonstrated by your experiments on PS3).