1

Prologo

"This is a rocket science but you
don't need to be a rocket
scientist to use it"
The Economist, September 2007

The main actor of this book is *the Algorithm* so, in order to dig into the beauty and challenges that pertain with its ideation and design, we need to start from one of its many possible definitions. The Oxford English Dictionary reports that an algorithm is, informally, *"a process, or set of rules, usually one expressed in algebraic notation, now used esp. in computing, machine translation and linguistics"*. The modern meaning for Algorithm is quite similar to that of *recipe, method, procedure, routine* except that the word Algorithm in Computer Science connotes something more precisely described. In fact many authoritative researchers have tried to pin down the term over the last 200 years [3] by proposing definitions which became more complicated and detailed nonetheless, hopefully in the minds of their proponents, more precise and elegant. As algorithm designers and engineers we will follow the definition provided by Donald Knuth at the end of the 60s [7, pag 4]: an Algorithm is *a finite, definite, effective procedure, with some output.* Although these five features may be intuitively clear and are widely accepted as requirements for a sequence-of-steps to be an Algorithm, they are so dense of significance that we need to look into them with some more detail, even because this investigation will surprisingly lead us to the scenario and challenges posed nowadays by algorithm design and engineering, and to the motivation underling these lectures.

Finite: *"An algorithm must always terminate after a finite number of steps ... a very finite number, a reasonable number."* Clearly, the term "reasonable" is related to the *efficiency* of the algorithm: Knuth [7, pag. 7] states that *"In practice, we not only want algorithms, we want good algorithms"*. The "goodness" of an algorithm is related to the use that the algorithm makes of some precious *computational resources* such as: time, space, communication, I/Os, energy, or just simplicity and elegance which both impact onto the coding, debugging and maintenance costs!

Definite: *"Each step of an algorithm must be precisely defined; the actions to be carried out must be rigorously and unambiguously specified for each case"*. Knuth made an effort in this direction by detailing what he called the "machine language" for his *"mythical MIX...the world's first polyunsaturated computer"*. Today we know of many other programming languages such as `C/C++`, `Java`, `Python`, etc. etc. All of them specify a set of instructions that the programmer may use to describe the procedure underlying his/her algorithm in an unambiguous way: "unambiguity" here is granted by the formal semantics that researchers have attached to each of these instructions. This eventually means that anyone reading the algorithm's description will interpret it in a precise way: nothing will be left to personal mood!

Effective: *"... all of the operations to be performed in the algorithm must be sufficiently basic that they can in principle be done exactly and in a finite length of time by a man using paper and pencil"*. Therefore the notion of "step" invoked in the previous item implies that one has to dig into a complete and deep understanding of the problem to be solved, and then into logical well-definite structuring of a step-by-step solution.

Procedure: *"... the sequence of specific steps arranged in a logical order"*.

Input: *"... quantities which are given to it initially before the algorithm begins. These inputs are taken from specified sets of objects"*.

Output: *"... quantities which have a specified relation to the inputs"*.

In this booklet we will not use a *formal* approach to algorithm description, because we wish to concentrate on the theoretically elegant and practically efficient ideas which underlie the algorithmic solution of some interesting problems, without being lost in the maze of programming technicalities. So in every lecture we will take an interesting *problem* coming out from a *real/useful application* and then propose *deeper and deeper* solutions of increasing sophistication and improved efficiency, taking care that this will not necessarily correspond to increasing the complexity of algorithm's description. Actually, problems were selected to admit *surprisingly* elegant solutions that can be described in few lines of code! So we will opt for the current practice of algorithm design and describe our algorithms either *colloquially* or by using *pseudo-code* that mimics the most famous `C` and `Java` languages. In any case we will not renounce to be as much rigorous as it needs an algorithm description to match the five features above.

Elegance will not be the only feature of our algorithm design, of course, we will also aim for *efficiency* which commonly relates to the *time/space complexity* of the algorithm. Traditionally time complexity has been evaluated as a function of the input size n by counting the (maximum) number of steps, say $T(n)$, an algorithm takes to complete its computation over an input of n items. Since the maximum is taken over all inputs of that size, the time complexity is named *worst case* because it concerns with the input that induces the worst behavior in time for the algorithm. Of course, the larger is n the larger is $T(n)$, which is therefore non decreasing and positive. In a similar way we can define the (worst-case) *space complexity* of an algorithm, as the maximum number of memory cells it uses for its computation over an input of size n. This approach to the *design* and *analysis* of algorithms assumes a very simple model of computation, known as *model of Von Neumann* (aka Random Access Machine, *RAM* model). This model consists of a CPU and a memory of infinite size and constant-time access to each one of its cells. Here we argue that every step takes a fixed amount of time on a PC, which is the same for any operation: being it arithmetic, logical, or just a memory access (read/write). Here one postulates that it is enough to *count* the number of steps executed by the algorithm in order to have an "accurate" estimate of its execution time on a real PC. Two algorithms can then be *compared* according to the *asymptotic behavior* of their time-complexity functions as $n \longrightarrow +\infty$, the faster is growing the time complexity over inputs of larger and larger size, the worse is its corresponding algorithm. The robustness of this approach has been debated for a long time but, eventually, the RAM model dominated the algorithmic scene for decades (and is still dominating it!) because of its simplicity, which impacts on algorithm design and evaluation, and its ability to estimate the algorithm performance "quite accurately" on (old) PCs. Therefore it is not surprising that most introductory books on Algorithms take the RAM model as a reference.

But in the last ten years things have changed significantly, thus highlighting the need for a *shift* in algorithm design and analysis! Two main changes occurred: the architecture of modern PCs became more and more sophisticated (not just one CPU and one monolithic memory!), and input data have exploded in size ("$n \longrightarrow +\infty$" does not live only in the theory world!) because they are abundantly generated by many sources: such as DNA sequencing, bank transactions, mobile communications, Web navigation and searches, auctions, etc. etc.. The first change turned the RAM model into an unsatisfactory abstraction of modern PCs; whereas the second change made the design

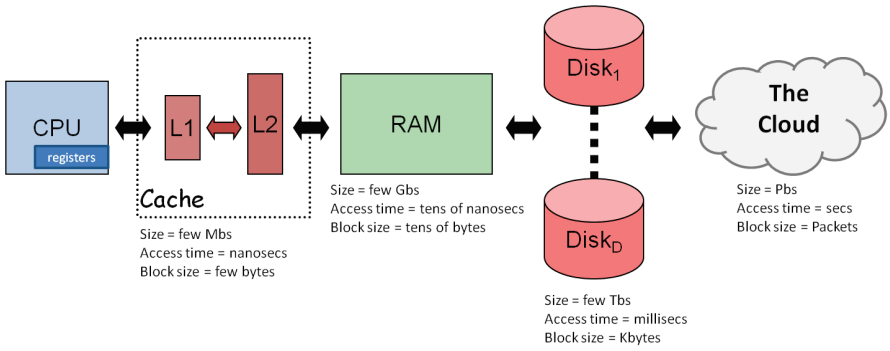

FIGURE 1.1: An example of memory hierarchy of a modern PC.

of *asymptotically good* algorithms ubiquitous and fruitful not only for dome-headed mathematicians but also for a much larger audience because of their impact on business [2], society [1] and science in general [4]. The net consequence was a revamped scientific interest in algorithmics and the spreading of the word "Algorithm" to even colloquial speeches!

In order to make algorithms effective in this new scenario, researchers needed new models of computation able to abstract in a better way the features of modern computers and applications and, in turn, to derive more accurate estimates of algorithm performance from their complexity analysis. Nowadays a modern PC consists of one or more CPUs (multi-core?) and a very complex hierarchy of memory levels, all with their own technological specialties (Figure 1.1): L1 and L2 caches, internal memory, one or more mechanical or SSDisks, other external storage devices (like flash memories, DVDs, tapes), and possibly other (hierarchical-)memories of multiple hosts distributed over a (possibly geographic) network, the so called "Cloud". Each of these memory levels has its own cost, capacity, latency, bandwidth and access method. The closer a memory level is to the CPU, the smaller, the faster and the more expensive it is. Currently nanoseconds suffice to access the caches, whereas milliseconds are yet needed to fetch data from disks (aka I/O). This is the so called *I/O-bottleneck* which amounts to the astonishing factor of $10^5 - 10^6$, nicely illustrated by Tomas H. Cormen with his quote:

"The difference in speed between modern CPU and (mechanical) disk technologies is analogous to the difference in speed in sharpening a pencil using a sharpener on one's desk or by taking an airplane to the other side of the world and using a sharpener on someone else's desk".

Engineering research is trying nowadays to improve the input/output subsystem to reduce the impact of the I/O-bottleneck onto the efficiency of applications managing large datasets; but, on the other hand, we are aware that the improvements achievable by means of a good algorithm design abundantly surpass the best expected technology advancements. Let us see the why with a simple example![1]

Assume to take three algorithms having increasing I/O-complexity: $C_1(n) = n$, $C_2(n) = n^2$ and $C_3(n) = 2^n$. Here $C_i(n)$ denotes the number of disk accesses executed by the ith algorithm to process n input data (stored e.g. in n/B disk pages). Notice that the first two algorithms execute a *polynomial* number of I/Os (in the input size), whereas the last one executes an *exponential* number of I/Os. Moreover we note that the above complexities have a very simple (and thus unnatural) mathematical

[1] This is paraphrased from [8], now we talk about I/Os instead of steps.

form because we want to simplify the calculations without impairing our final conclusions. Let us now ask for how many data each of these algorithms is able to process in a fixed time-interval of size t, given that each I/O takes c time. The answer is obtained by solving the equation $C_i(n) \times c = t$ with respect to n: so we get t/c data are processed by the first algorithm in t time, $\sqrt{t/c}$ data are processed by the second algorithm, and only $\log_2(t/c)$ data are processed by the third algorithm. These values are already impressive by themselves, and provide a robust understanding of why polynomial-time algorithms are called *efficient*, whereas exponential-time algorithms are called *inefficient*: a large change in the length t of the time-interval induces just a tiny change in the amount of data that exponential-time algorithms can process. Of course, this distinction admits many exceptions when the problem instances of interest have limited size; furthermore, there are exponential-time algorithms that have been quite useful in practice (think e.g. to the simplex algorithm). On the other hand, these examples are quite rare, and the much more stringent bounds on execution time satisfied by polynomial-time algorithms make them considered *provably* efficient and the preferred way to solve problems. Algorithmically speaking, most exponential-time algorithms are merely implementations of the approach based on *exhaustive* search, whereas polynomial-time algorithms are generally made possible only through the gain of some deeper insight into the structure of a problem. So polynomial-time algorithms are the *right* choice from many points of view. Let us now assume to run the above algorithms with a better I/O-subsystem, say one that is k times faster, and ask: How many data can be managed by this new PC? To address this question we solve the previous equations with the time-interval set to the length $k \times t$, thus implicitly assuming to run the algorithms over the old PC but providing itself with k times more time. We get that the first algorithm perfectly scales by a factor k, the second algorithm scales by a factor $\sqrt{k}$, whereas the last algorithm scales *only of an additional term* $\log_2 k$. Noticeably the improvement induced by a k-times more powerful PC for an exponential-time algorithm is totally negligible even in the presence of impressive (and thus unnatural) technology advancements! Super-linear algorithms, like the second one, are positively affected by technology advancements but their performance improvement decreases as the degree of the polynomial-time complexity grows: more precisely, if $C(n) = n^\alpha$ then a k-times more powerful PC induces a speed-up of a factor $\sqrt[\alpha]{k}$. Overall, it is not hazardous to state that the impact of a good algorithm is far beyond any optimistic forecasting for the performance of future (mechanical or SSD) disks.[2]

Given this *appetizer* about the "Power of Algorithms", let us now turn back to the problem of analyzing the performance of algorithms in modern PCs by considering the following simple example: Compute the sum of the integers stored in an array $A[1, n]$. The simplest idea is to scan A and accumulate in a temporary variable the sum of the scanned integers. This algorithm executes n sums, accesses each integer in A once, and thus takes n steps. Let us now generalize this approach by considering a family of algorithms, denoted by $\mathcal{A}_{s,b}$, which differentiate themselves according to the pattern of accesses to A's elements, as driven by the parameters s and b. In particular $\mathcal{A}_{s,b}$ looks at array A as logically divided into blocks of b elements each, say $A_j = A[j * b, (j+1) * b - 1]$ for $j = 0, 1, 2, \ldots, n/b - 1$.[3] Then it sums all items in one block before moving to the next block that is s blocks apart on the right. Array A is considered as cyclic so that, when the next block lies out of A, the algorithm wraps around it starting again from its beginning. Clearly not all values of s allow to take into account all of A's blocks (and thus sum all of A's integers). Nevertheless we know that if s is co-prime with n/b then $s \times i \mod (n/b)$ generates a permutation of the integers $\{0, 1, \ldots, n/b - 1\}$, and thus $\mathcal{A}_{s,b}$ touches all blocks in A and hence sums all of its integers. But the specialty of this parametrization is that by varying s and b we can sum according to different patterns

[2] See [11] for an extended treatment of this subject.

[3] For the sake of presentation we assume that n and b are powers of two, so b divides n.

of memory accesses: from the sequential scan indicated above (setting $s = b = 1$), to a block-wise access (set a large b) and/or a random-wise access (set a large s). Of course, all algorithms $\mathcal{A}_{s,b}$ are equivalent from a computational point of view, since they read exactly n integers and thus take n steps; but from a practical point of view, they have different time performance which becomes more and more significant as the array size n grows. The reason is that, for a growing n, data will be spread over more and more memory levels, each one with its own capacity, latency, bandwidth and access method. So the "equivalence in efficiency" derived by adopting the RAM model, and counting the number-of-steps executed by $\mathcal{A}_{s,b}$, is not an accurate estimate of the real time required by that algorithms to sum A's elements.

We need a different model that grasps the essence of real computers and is simple enough to not jeopardize algorithm design and analysis. In a previous example we already argued that the number of I/Os is a good estimator for the time complexity of an algorithm, given the large gap existing between disk- and internal-memory accesses. This is indeed what is captured by the so called *2-level memory model* (aka. disk-model, or external-memory model [11]) which abstracts the computer as composed by only *two memory levels*: the internal memory of size M, and the (unbounded) disk memory which operates by reading/writing data via blocks of size B (called *disk pages*). Sometimes the model consists of D disks, each of unbounded size, so that each I/O reads or writes a total of $D \times B$ items coming from D pages, each one residing on a different disk. For the sake of clarity we remark that the *two-level view* must not suggest to the reader that this model is restricted to abstracts disk-based computations; in fact, we are actually free to choose any two levels of the memory hierarchy, with their M and B parameters properly set. The algorithm performance is evaluated in this model by counting: (a) the number of accesses to disk pages (hereafter *I/O*s), (b) the internal running time (CPU time), and (c) the number of disk pages used by the algorithm as its working space. This suggests *two golden rules* for the design of "good" algorithms operating on large datasets: they must exploit *spatial locality* and *temporal locality*. The former imposes a data organization onto the disk(s) that makes each accessed disk-page as much useful as possible; the latter imposes to execute as much useful work as possible onto data fetched in internal memory, before they are written back to disk.

In the light of this new model, let us re-analyze the time complexity of algorithms $\mathcal{A}_{s,b}$ by taking into account I/Os, given that the CPU time is still n and the space occupancy is n/B pages. We start from the simplest settings for s and b in order to gain some intuitions about the general formulas. The case $s = 1$ is obvious, algorithms $\mathcal{A}_{1,b}$ scan A rightward by taking n/B I/Os, independently of the value of b. The case $b = n$ coincides with the previous one (independently of s), because we are considering just one big block that coincides with A. As s and b change the situation complicates, but by not much. Fix $s = 2$ and pick some $b < B$ that, for simplicity, is assumed to divide the block-size B. Every block of size B consists of B/b smaller (logical) blocks of size b, and the algorithm $\mathcal{A}_{2,b}$ examines only half of them because of the jump $s = 2$. This actually means that each B-sized page is half utilized in the summing process, thus inducing a total of $2n/B$ I/Os. It is then not difficult to generalize this formula to any s by writing a cost of sn/B I/Os, which correctly gives n/B in the simplest cases dealt with above. This formula provides a better approximation of the real time complexity of the algorithm, although it does not capture all features of the disk. In fact, it considers all I/Os as *equal*, independently of their distribution. This is clearly unprecise because on real disks the *sequential* I/Os are faster than the *random* I/Os.[4] Referring to the previous example, the algorithms $\mathcal{A}_{s,B}$ have still I/O-complexity n/B, independently of s, although their behavior is

[4]Conversely, this difference will be almost negligible in an (electronic) memory, such as the DRAM or the modern Solid-State disks, where the distribution of the memory accesses does not significantly impact onto the throughput of the memory/SSD.

rather different if executed on a (mechanical) disk because of the disk seeks induced by larger and larger s. As a result, we can conclude that even the 2-level memory model is an approximation of the behavior of algorithms on real computers, although it results sufficiently good that it has been widely adopted in the literature to evaluate their performance on massive datasets. So that, in order to be as much precise as possible, we will evaluate in these notes our algorithms by specifying not only the number of executed I/Os but also characterizing their *distribution* (random vs contiguous) over the disk.

At this point one could object that given the impressive technological advancements of the last years, the internal-memory size M is so large that most of the working set of an algorithm (roughly speaking, the set of pages it will reference in the near future) can be fit into it, thus reducing significantly the case of an I/O-fault. We will argue that an even small portion of data resident to disk makes the algorithm significantly slower than expected, and so, data organization cannot be neglected even in these extremely favorable situations.

Let us see why, by means of a "back of the envelope" calculation! Assume that the input size $n = (1+\epsilon)M$ is larger than the internal-memory size of a factor $\epsilon > 0$. The question is how much ϵ impacts onto the average cost of an algorithm step, given that it may access a datum located either in internal memory or on disk. To simplify our analysis, without renouncing to a meaningful conclusion, we assume that $p(\epsilon)$ is the probability of an I/O-fault. As an example, if $p(\epsilon) = 1$ then the algorithm always accesses its data on disk (i.e. one of the ϵM items); if $p(\epsilon) = \frac{\epsilon}{1+\epsilon}$ then the algorithm has a fully-random behavior in accessing its input data (since, from above, it is $p(\epsilon) == \frac{\epsilon}{1+\epsilon} = \frac{\epsilon M}{(1+\epsilon)M} = \frac{\epsilon M}{n}$); finally, if $p(\epsilon) = 0$ then the algorithm has a working set smaller than the internal memory size, and thus it does not execute any I/Os. Overall $p(\epsilon)$ measures the *un-locality* of the memory references of the analyzed algorithm.

To complete the notation, let us indicate with c the time needed for 1 I/O— we have $c \approx 10^5 - 10^6$, see above— and we set a to be the fraction of steps that induce a memory access in the running algorithm (this is typically $30\% - 40\%$, according to [6]). Now we are ready to estimate the *average cost of the step* for an algorithm working in this scenario:

$$1 \times \mathcal{P}(\text{computation step}) + t_m \times \mathcal{P}(\text{memory-access step}),$$

where t_m is the average cost of a memory access. To compute t_m we have to distinguish two cases: an in-memory access (occurring with probability $1 - p(\epsilon)$) or a disk access (occurring with probability $p(\epsilon)$). So we have $t_m = 1 \times (1 - p(\epsilon)) + c \times p(\epsilon)$. Observing that $\mathcal{P}(\text{memory-access step}) + \mathcal{P}(\text{computation step}) = 1$, and plugging the fraction of memory accesses $\mathcal{P}(\text{memory access step}) = a$, we derive the final formula:

$$(1-a) \times 1 + a \times [1 \times (1 - p(\epsilon)) + c \times p(\epsilon)] = 1 + a \times (c-1) \times p(\epsilon) \geq 3 \times 10^4 \times p(\epsilon).$$

This formula clearly shows that, even for algorithms exploiting locality of references (i.e. a small $p(\epsilon)$), the slowdown may be significant and actually it turns out to be four order of magnitudes larger than what might be expected (i.e. $p(\epsilon)$). Just as an example, take an algorithm that exploits locality of references in its memory accesses, say 1 out of 1000 memory accesses is on disk (i.e. $p(\epsilon) = 0.001$). Then, its performance on a massive dataset that is stored on disk would be slowed down by a factor > 30 with respect to a computation executed completely in internal-memory.

It goes without saying that this is just the tip of the iceberg, because the larger is the amount of data to be processed by an algorithm, the higher is the number of memory levels involved in the storage of these data and, hence, the more variegate are the types of "memory faults" (say cache-faults, memory-faults, etc.) to cope with for achieving efficiency. The overall message is that neglecting questions pertaining to the cost of memory references in a hierarchical-memory system may *prevent* the use of an algorithm on large input data.

Motivated by these premises, these notes will provide few examples of challenging problems which admit elegant algorithmic solutions whose efficiency is crucial to manage the large datasets that occur in many real-world applications. Algorithm design will be accompanied by several comments on the difficulties that underlie the *engineering* of those algorithms: how to turn a "theoretically efficient" algorithm into a "practically efficient" code. Too many times, as a theoretician, I got the observation that "your algorithm is far from being amenable to an efficient implementation!". By following the recent surge of investigations in *Algorithm Engineering* [10] (to be not confused with the "practice of Algorithms"), we will also dig into the deep computational features of some algorithms by resorting few other successful models of computations— mainly the streaming model [9] and the cache-oblivious model [5]. These models will allow us to capture and highlight some interesting issues of the underlying computation: such as disk passes (streaming model), and universal scalability (cache-oblivious model). We will try our best to describe all these issues in their simplest terms but, nonetheless to say, we will be unsuccessful in turning this "rocket science for non-boffins" into a "science for dummies" [2]. In fact lots of many more things have to fall into place for algorithms to work: top-IT companies (like Google, Yahoo, Microsoft, IBM, AT&T, Oracle, Facebook, Twitter, etc.) are perfectly aware of the difficulty to find people with the right skills for developing and refining "good" algorithms. This booklet will scratch just the surface of Algorithm Design and Engineering, with the main goal of spurring inspiration into your daily job as software designer or engineer.

References

[1] Person of the Year. *Time Magazine*, 168(27–28), December 2006.

[2] Business by numbers. *The Economist*, September 2007.

[3] Wikipedia's entry: "Algorithm characterizations", 2009. At `http://en.wikipedia.org/wiki/Algorithm_characterizations`

[4] Declan Butler. *2020 computing: Everything, everywhere*, volume 440, chapter 3, pages 402–405. Nature Publishing Group, March 2006.

[5] Rolf Fagerberg. Cache-oblivious model. In Ming-Yang Kao, editor, *Encyclopedia of Algorithms*. Springer, 2008.

[6] John L. Hennessy and David A. Patterson. *Computer Architecture, Fourth Edition: A Quantitative Approach*. Morgan Kaufmann, September 2006.

[7] Donald Knuth. *The Art of Computer Programming: Fundamental Algorithms*, volume 1. Addison-Wesley, 1973.

[8] Fabrizio Luccio. *La struttura degli algoritmi*. Boringhieri, 1982.

[9] S. Muthukrishnan. Data streams: Algorithms and applications. *Foundations and Trends in Theoretical Computer Science*, 1(2), 2005.

[10] Peter Sanders. Algorithm engineering - an attempt at a definition. In Susanne Albers, Helmut Alt, and Stefan Näher, editors, *Efficient Algorithms, Essays Dedicated to Kurt Mehlhorn on the Occasion of His 60th Birthday*, volume 5760 of *Lecture Notes in Computer Science*, pages 321–340. Springer, 2009.

[11] Jeffrey S. Vitter. External memory algorithms and data structures. *ACM Computing Surveys*, 33(2):209–271, 2001.

2

A warm-up!

"Everything should be made as simple as possible, but not simpler."
Albert Einstein

Let us consider the following problem, surprisingly simple in its statement but not that much for what concerns the design of its optimal solution.

Problem. *We are given the performance of a stock at* NYSE *expressed as a sequence of* day-by-day differences *of its quotations. We wish to determine the best buy-&-sell strategy for that stock, namely the pair of days $\langle b, s\rangle$ that would have maximized our revenues if we would have bought the stock at (the beginning of) day b and sold it at (the end of) day s.*

The specialty of this problem is that it has a simple formulation, which finds many other useful variations and applications. We will comment on them at the end of this lecture, now we content ourselves by mentioning that we are interested in this problem because it admits a sequence of algorithmic solutions of increasing sophistication and elegance, which imply a significant reduction in their time complexity. The ultimate result will be a linear-time algorithm, i.e. linear in the number n of stock quotations. This algorithm is *optimal* in terms of the number of executed steps, because all day-by-day differences must be looked at in order to determine if they must be included or not in the optimal solution, actually, one single difference could provide a one-day period worth of investment! Surprisingly, the optimal algorithm will exhibit the simplest pattern of memory accesses— it will execute a single scan of the available stock quotations— and thus it will offer a *streaming behavior*, particularly useful in a scenario in which the granularity of the buy-&-sell actions is not restricted to full-days and we must possibly compute the optimal time-window *on-the-fly* as quotations oscillate. More than this, as we commented in the previous lecture, this algorithmic scheme is optimal in terms of I/Os and *uniformly* over all levels of the memory hierarchy. In fact, because of its streaming behavior, it will execute n/B I/Os independently of the disk-page size B, which may be thus unknown to the underlying algorithm. This is the typical feature of the so called *cache-oblivious algorithms* [4], which we will therefore introduce at the right point of this lecture.

This lecture will be the prototype of what you will find in the next pages: a simple problem to state, with few elegant solutions and challenging techniques to teach and learn, together with several intriguing extensions that can be posed as exercises to the students or as puzzles to tempt your mathematical skills!

Let us now dig into the technicalities, and consider the following example. Take the case of 11 days of exchange for a given stock, and assume that $D[1, 11] = [+4, -6, +3, +1, +3, -2, +3, -4, +1,$

$-9, +6]$ denotes the day-by-day differences of quotations of that stock. It is not difficult to convince yourself that the gain of buying the stock at day x and selling it at day y is equal to the sum of the values in the sub-array $D[x, y]$, namely the sum of all its fluctuations. As an example, take $x = 1$ and $y = 2$, the gain is $+4 - 6 = -2$, and indeed we would loose 2 dollars in buying the morning of the first day and selling the stock at the end of the second day. Notice that the starting value of the stock is not crucial for determining the best time-interval of our investment, what is important are its variations. In other words, the problem stated above boils down to determine the sub-array of $D[1, n]$ which maximizes the sum of its elements. In the literature this problem is indeed known as the *maximum sub-array sum* problem.

Problem Abstraction. *Given an array $D[1, n]$ of positive and negative numbers, we want to find the sub-array $D[b, s]$ which maximizes the sum of its elements.*

It is clear that if all numbers are positive, then the optimal sub-array is the entire D: this is the case of an always increasing stock price, and indeed there is no reason to sell it before the last day! Conversely, if all numbers are negative, then we can pick the one-element window containing the largest (negative) value: if you are imposed to buy this poor stock, then do it in the day it looses the smallest value and sell it soon! In all other cases, it is not at all clear where the optimum sub-array is located. In the example, the optimum spans $D[3, 7] = [+3, +1, +3, -2, +3]$ and has gain +8 dollars. This shows that the optimum neither includes the best exploit of the stock (i.e. +6) nor it consists of positive values only. The *structure* of the optimum sub-array is not simple but, surprisingly enough, not very complicated as we will show in Section 2.3.

2.1 A cubic-time algorithm

We start by considering an inefficient solution which translates in pseudo-code the formulation of the problem given above. This algorithm is detailed in Figure 2.1, where the pair of variables $<b_o, s_o>$ identifies the current sub-array of maximum sum, whose value is stored in `MaxSum`. Initially `MaxSum` is set to the dummy value $-\infty$, so that it is immediately changed whenever the algorithm executes Step 8 for the first time. The core of the algorithm examines all possible sub-arrays $D[b, s]$ (Steps 2-3) computing for each of them the sum of their elements (Steps 4-7). If a sum larger than the current maximal value is found (Steps 8-9), then `TmpSum` and its corresponding sub-array are stored in `MaxSum` and $<b_o, s_o>$, respectively.

Algorithm 2.1 The cubic-time algorithm

```
1:  MaxSum = -∞
2:  for (b = 1; b ≤ n; b++) do
3:      for (s = b; s ≤ n; s++) do
4:          TmpSum = 0
5:          for (i = b; i ≤ s; i++) do
6:              TmpSum+ = D[i];
7:          end for
8:          if (MaxSum < TmpSum) then
9:              MaxSum = TmpSum; b_o = b; s_o = s;
10:         end if
11:     end for
12: end for
13: return ⟨MaxSum, b_o, s_o⟩;
```

The correctness of the algorithm is immediate, since it checks all possible sub-arrays of $D[1,n]$ and selects the one whose sum of its elements is the largest (Step 8). The time complexity is cubic, i.e. $\Theta(n^3)$, and can be evaluated as follows. Clearly the time complexity is upper bounded by $O(n^3)$ because we can form no more than $\frac{n^2}{2}$ pairs <b,s> out of n elements,[1] and n is an upper-bound to the cost of computing the sum of each sub-array. Let us now show that the time cost is also $\Omega(n^3)$, so concluding that the time complexity is strictly cubic. To show this lower bound, we observe that $D[1,n]$ contains $(n-L)$ sub-arrays of length L, and thus the cost of computing the sum for all of their elements is $(n-L)\times L$. Summing over all values of L, we would get the exact time complexity. But here we are interested in a lower bound, so we can evaluate that cost just for the subset of sub-arrays whose length L is in the range $[n/4, n/2]$. For each such L, we have that $n - L \geq n/2$ and $L \geq n/4$, so the cost above is $(n-L)\times L \geq n^2/8$. Since we have $n/4$ of those Ls, the total cost for analyzing that subset of sub-arrays is lower bounded by $n^3/32 = \Omega(n^3)$.

It is natural now to ask ourselves how much fast in practice is the designed algorithm. We implemented it in Java and tested on a commodity PC. As n grows, its time performance reflects in practice its cubic time complexity, evaluated in the RAM model. More precisely, it takes about 20 seconds to solve the problem for $n = 10^3$ elements, and about 30 hours for $n = 10^5$ elements. Too much indeed if we wish to scale to very large sequences (of quotations), as we are aiming for in these lectures.

2.2 A quadratic-time algorithm

The key inefficiency of the cubic-time algorithm resides in the execution of Steps 4-7 of the pseudo-code in Figure 2.1. These steps re-compute from scratch the sum of the sub-array $D[b,s]$ each time its extremes change in Steps 2-3. But if we look carefully at the **for**-cycle of Step 3 we notice that the size s is incremented by one unit at a time from the value b (one element sub-array) to the value n (the longest possible sub-array that starts at b). Therefore, from one execution to the next one of Step 3, the sub-array to be summed changes from $D[b,s]$ to $D[b,s+1]$. It is thus immediate to conclude that the new sum for $D[b,s+1]$ does not need to be recomputed from scratch, but can be computed *incrementally* by just adding the value of the new element $D[s+1]$ to the current value of `TmpSum` (which inductively stores the sum of $D[b,s]$). This is exactly what the pseudo-code of Figure 2.2 implements: its two main changes with respect to the cubic algorithm of Figure 2.1 are in Step 3, that nulls `TmpSum` every time b is changed (because the sub-array starts again from length 1, namely $D[b,b]$), and in Step 5, that implements the incremental update of the current sum as commented above. Such small changes are worth of a saving of $\Theta(n)$ additions per execution of Step 2, thus turning the new algorithm to have quadratic-time complexity, namely $\Theta(n^2)$.

More precisely, let us concentrate on counting the number of additions executed by the algorithm of Figure 2.2; this is the prominent operation of this algorithm so that its evaluation will give us an estimate of its total number of steps. This number is[2]

$$\sum_{b=1}^{n}(1+\sum_{s=b}^{n}1) = \sum_{b=1}^{n}(1+(n-b+1)) = n\times(n+2) - \sum_{b=1}^{n} b = n^2 + 2n - \frac{n(n-1)}{2} = O(n^2).$$

This improvement is effective also in practice. Take the same experimental scenario of the previous section, this new algorithm requires less than 1 second to solve the problem for $n = 10^3$

[1] For each pair $< b, s >$, with $b \leq s$, $D[b,s]$ is a possible sub-array, but $D[s,b]$ is not.

[2] We use below the famous formula, discovered by the young Gauss, to compute the sum of the first n integers.

Algorithm 2.2 The quadratic-time algorithm

```
1:  MaxSum = −∞;
2:  for (b = 1; b ≤ n; b++) do
3:      TmpSum = 0;
4:      for (s = b; s ≤ n; s++) do
5:          TmpSum += D[s];
6:          if (MaxSum < TmpSum) then
7:              MaxSum = TmpSum; b_o = b; s_o = s;
8:          end if
9:      end for
10: end for
11: return ⟨MaxSum, b_o, s_o⟩;
```

elements, and about 28 minutes to manage 10^6 elements. This means that the new algorithm is able to manage more elements in "reasonable" time. Clearly, these timings and these numbers could change if we use a different programming language (Java, in the present example), operating system (Windows, in our example), and processor (the old Pentium IV, in our example). Nevertheless we believe that they are interesting anyway because they provide a concrete picture of what it does mean a theoretical improvement like the one we showed in the above paragraphs on a real situation. It goes without saying that the life of a coder is typically not so easy because theoretically-good algorithms many times hide so many details that their engineering is difficult and big-O notation often turn out to be not much "realistic". Do not worry, we will have time in these lectures to look at these issues in more detail.

2.3 A linear-time algorithm

The final step of this lecture is to show that the maximum sub-array sum problem admits an elegant algorithm that processes the elements of $D[1,n]$ is a streaming fashion and takes the *optimal* $O(n)$ time. We could not aim for more!

To design this algorithm we need to dig into the structural properties of the optimal sub-array. For the purpose of clarity, we refer the reader to Figure 2.1 below, where the optimal sub-array is assumed to be located at two positions $b_o \le s_o$ in the range $[1,n]$.

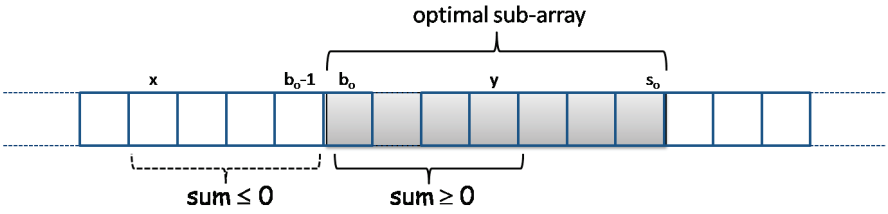

FIGURE 2.1: An illustrative example of Properties 1 and 2.

Let us now take a sub-array that starts before b_o and ends at position $b_o - 1$, say $D[x, b_o - 1]$. The sum of the elements in this sub-array cannot be positive because, otherwise, we could merge it with the (adjacent) optimal sub-array and thus get the longer sub-array $D[x, s_o]$ having sum *larger than*

the one obtained with the (claimed) optimal $D[b_o, s_o]$. So we can state the following:

Property 1. The sum of the elements in a sub-array $D[x, b_o - 1]$, with $x < b_o$, cannot be (strictly) positive.

Via a similar argument, we can consider a sub-array that is a prefix of the optimal $D[b_o, s_o]$, say $D[b_o, y]$ with $y \leq s_o$. This sub-array cannot have negative sum because, otherwise, we could drop it from the optimal solution and get a shorter array, namely $D[y + 1, s_o]$ having sum larger than the one obtained by the (claimed) optimal $D[b_o, s_o]$. So we can state the following other property:

Property 2. The sum of the elements in a sub-array $D[b_o, y]$, with $y \leq s_o$, cannot be (strictly) negative.

We remark that any one of the sub-arrays considered in the above two properties might have sum equal to zero. This would not affect the optimality of $D[b_o, s_o]$, it could only introduce other optimal solutions being either longer or shorter than $D[b_o, s_o]$.

Let us illustrate these two properties on the array $D[1, 11] = [+4, -6, +3, +1, +3, -2, +3, -4, +1, -9, +6]$. Here the optimum sub-array is $D[3, 7] = [+3, +1, +3, -2, +3]$. We note that $D[x, 2]$ is always negative (Prop. 1), in fact for $x = 1$ the sum is $+4 - 6 = -2$ and for $x = 2$ the sum is -6. On the other hand the sum of all elements in $D[3, y]$ is positive for all prefixes of the optimum sub-array (Prop. 2), namely $y \leq 7$. We also point out that the sum of $D[3, y]$ is positive even for some $y > 7$, take for example $D[3, 8]$ for which the sum is 4 and $D[3, 9]$ for which the sum is 5. Of course, this does not contradict Prop. 2.

Algorithm 2.3 The linear-time algorithm

```
1:  MaxSum = −∞
2:  TmpSum = 0; b = 1;
3:  for (s = 1; s ≤ n; s++) do
4:      TmpSum += D[s];
5:      if (MaxSum < TmpSum) then
6:          MaxSum = TmpSum; b_o = b; s_o = s;
7:      end if
8:      if (TmpSum < 0) then
9:          TmpSum = 0; b = s + 1;
10:     end if
11: end for
12: return ⟨MaxSum, b_o, s_o⟩;
```

The two properties above lead to the simple Algorithm 2.3. It consists of one unique **for**-cycle (Step 3) which keeps in `TmpSum` the sum of a sub-array ending in the currently examined position s and starting at some position $b \leq s$. At any step of the **for**-cycle, the candidate sub-array is extended one position to the right (i.e. s++), and its sum `TmpSum` is increased by the value of the current element $D[s]$ (Step 4). Since the current sub-array is a candidate to be the optimal one, its sum is compared with the current optimal value (Step 5). Then, according to Prop. 1, if the sub-array sum is negative, the current sub-array is discarded and the process "restarts" with a new sub-array beginning at the next position $s + 1$ (Steps 8-9). Otherwise, the current sub-array is extended to the right, by incrementing s. The tricky issue here is to show that the optimal sub-array is checked in Step 5, and thus stored in $< b_o, s_o >$. This is not intuitive at all because the algorithm is checking n sub-arrays out of the $\Theta(n^2)$ possible ones, and we want to show that this (minimal) subset of

candidates actually contains the optimal solution. This subset is *minimal* because these sub-arrays form a *partition* of $D[1, n]$ so that every element belongs to one, and only one checked sub-array. Moreover, since every element must be analyzed, we cannot discard any sub-array of this partition without checking its sum!

Before digging into the formal proof of correctness, let us follow the execution of the algorithm over the array $D[1, 11] = [+4, -6, +3, +1, +3, -2, +3, -4, +1, -9, +6]$. Remember that the optimum sub-array is $D[3, 7] = [+3, +1, +3, -2, +3]$. We note that $D[x, 2]$ is negative for $x = 1, 2$, so the algorithm surely zeroes the variable `TmpSum` when $s = 2$ in Steps 8-9. At that time, b is set to 3 and `TmpSum` is set to 0. The subsequent scanning of the elements $s = 3, \ldots, 7$ will add their values to `TmpSum` which is always positive (see above). When $s = 7$, the examined sub-array coincides with the optimal one, we thus have `TmpSum` = 8, so Step 5 stores the optimum location in $< b_o, s_o >$. It is interesting to notice that, in this example, the algorithm does not re-start the value of `TmpSum` at the next position $s = 8$ because it is still positive (namely, `TmpSum` = 4); this means that the algorithm will examine sub-arrays longer than the optimal one, but all having a smaller sum, of course. The next re-starting will occur at position $s = 10$ where `TmpSum` $= -4$.

It is easy to realize that the time complexity of the algorithm is $O(n)$ because every element is examined just once. More tricky is to show that the algorithm is correct, which actually means that Steps 4 and 5 eventually compute and check the optimal sub-array sum. To show this, it suffices to prove the following two facts: (i) when $s = b_o - 1$, Step 8 resets b to b_o; (ii) for all subsequent positions $s = b_o, \ldots, s_o$, Step 8 never resets b so that it will eventually compute in `TmpSum` the sum of all elements in $D[b_o, s_o]$, whenever $s = s_o$. It is not difficult to see that Fact (i) derives from Property 1, and Fact (ii) from Property 2.

This algorithm is very fast in the same experimental scenario mentioned in the previous sections, it takes less than 1 second to process millions of quotations. A truly scalable algorithm, indeed, with many nice features that make it appealing also in a hierarchical-memory setting. In fact, this algorithm scans the array D from left to right and examines each of its elements just once. If D is stored on disk, these elements are fetched in internal memory one page at a time. Hence the algorithm executes n/B I/Os, which is *optimal*. It is interesting also to note that the design of the algorithm does not depend on B (which indeed does not appear in the pseudo-code), but we can evaluate its I/O-complexity in terms of B. Hence the algorithm takes n/B optimal I/Os independently of the the page size B, and thus subtly on the hierarchical-memory levels interested by the algorithm execution. Decoupling the use of the parameter B between algorithm design and algorithm analysis is the key issue of the so called *cache-oblivious algorithms*, which are a hot topic of algorithmic investigation nowadays. This feature is achieved here in a basic (trivial) way by just adopting a scan-based approach. The literature [4] offers more sophisticated results regarding the design of cache-oblivious algorithms and data structures.

2.4 Another linear-time algorithm

There exists another optimal solution to the maximum sub-array sum problem which hinges on a different algorithm design. For simplicity of exposition, let us denote by $Sum_D[y', y'']$ the sum of the elements in the sub-array $D[y', y'']$. Take now a selling time s and consider all sub-arrays that end at s: namely we are interested in sub-arrays having the form $D[x, s]$, with $x \leq s$. The value $Sum_D[x, s]$ can be expressed as the difference between $Sum_D[1, s]$ and $Sum_D[1, x - 1]$. Both of these sums are indeed *prefix*-sums over the array D and can be computed in linear time. As a result, we can rephrase our maximization problem as follows:

$$\max_s \max_{b \leq s} \mathrm{Sum}_D[b, s] = \max_s \max_{b \leq s} (\mathrm{Sum}_D[1, s] - \mathrm{Sum}_D[1, b - 1]).$$

We notice that if $b = 1$ the second term refers to the empty sub-array $D[1, 0]$; so we can assume that $Sum_D[1, 0] = 0$. This is the case in which $D[1, s]$ is the sub-array of maximum sum among all the sub-arrays ending at s (so no prefix sub-array $D[1, b-1]$ is dropped from it).

The next step is to pre-compute all prefix sums $P[i] = Sum_D[1, i]$ in $O(n)$ time and $O(n)$ space via a scan of the array D: Just notice that $P[i] = P[i-1] + D[i]$, where we set $P[0] = 0$ in order to manage the special case above. Hence we can rewrite the maximization problem in terms of the array P, rather than Sum_D: $\max_{b \le s}(P[s] - P[b-1])$. The cute observation now is that we can decompose the max-computation into a max-min calculation over the two variables b and s

$$\max_s \max_{b \le s}(P[s] - P[b-1]) = \max_s(P[s] - \min_{b \le s} P[b-1]).$$

The key idea is that we can move $P[s]$ outside the inner max-calculation because it does not depend on the variable b, and then change a max into a min because of the negative sign. The final step is then to pre-compute the minimum $\min_{b \le s} P[b-1]$ for all positions s, and store it in an array $M[0, n-1]$. We notice that, also in this case, the computation of $M[i]$ can be performed via a single scan of P in $O(n)$ time and space: set $M[0] = 0$ and then derive $M[i]$ as $\min\{M[i-1], P[i]\}$. Given M, we can rewrite the formula above as $\max_s(P[s] - M[s-1])$ which can be clearly computed in $O(n)$ time given the two arrays P and M. Overall this new approach takes $O(n)$ time and $O(n)$ extra space.

As an illustrative example, consider again the array $D[1, 11] = [+4, -6, +3, +1, +3, -2, +3, -4, +1, -9, +6]$. We have that $P[0, 11] = [0, +4, -2, +1, +2, +5, +3, +6, +2, +3, -6, 0]$ and $M[0, 10] = [0, 0, -2, -2, -2, -2, -2, -2, -2, -2, -6]$. If we compute the difference $P[s] - M[s-1]$ for all $s = 1, \ldots, n$, we obtain the sequence of values $[+4, -2, +3, +4, +7, +5, +8, +4, +5, -4, +6]$, whose maximum (sum) is $+8$ that occurs (correctly) at the (final) position $s = 7$. It is interesting to note that the left-extreme b_o of the optimal sub-array could be derived by finding the position $b_o - 1$ where $P[b_o - 1]$ is minimum: in the example, $P[2] = -2$ and thus $b_o = 3$.

Algorithm 2.4 Another linear-time algorithm

```
 1: MaxSum = −∞; b_o = 1;
 2: TmpSum = 0; MinTmpSum = 0;
 3: for (s = 1; s ≤ n; s++) do
 4:     TmpSum += D[s];
 5:     if (MaxSum < TmpSum − MinTmpSum) then
 6:         MaxSum = TmpSum; s_o = s;
 7:     end if
 8:     if (TmpSum < MinTmpSum) then
 9:         MinTmpSum = TmpSum; b_o = s + 1;
10:     end if
11: end for
12: return ⟨MaxSum, b_o, s_o⟩;
```

We conclude this section by noticing that the proposed algorithm executes three passes over the array D, rather than the single pass of Algorithm 2.3. It is not difficult to turn this algorithm to make *one-pass* too. It suffices to deploy the associativity of the min/max functions, and use two variables that inductively keep the values of $P[s]$ and $M[s-1]$ as the array D is scanned from left to right. Algorithm 2.4 implements this idea by using the variable `TmpSum` to store $P[s]$ and the variable `MinTmpSum` to store $M[s-1]$. This way the formula $\max_s(P[s] - M[s-1])$ is evaluated incrementally for $s = 1, \ldots, n$, thus avoiding the two passes for pre-calculating the arrays P and

M and the extra-space needed to store them. One pass over D is then enough, and so we have re-established the nice algorithmic properties of Algorithm 2.3 but with a completely different design!

2.5 Few interesting variants$^{\infty}$

As we promised at the beginning of this lecture, we discuss now few interesting variants of the maximum sub-array sum problem. For further algorithmic details and formulations we refer the interested reader to [1, 2]. Note that this is a challenging section, because it proposes an algorithm whose design and analysis are sophisticated!

Sometimes in the bio-informatics literature the term "sub-array" is substituted by "segment", and the problem takes the name of "maximum-sum segment problem". In the bio-context the goal is to identify segments which occur inside DNA sequences (i.e. strings of four letters A, T, G, C) and are *rich* of G or C nucleotides. Biologists believe that these segments are biologically significant since they predominantly contain genes. The mapping from DNA sequences to *arrays of numbers*, and thus to our problem abstraction, can be obtained in several ways depending on the objective function that models the *GC-richness* of a segment. Two interesting mappings are the following ones:

- Assign a penalty $-p$ to the nucleotides A and T of the sequence, and a reward $1-p$ to the nucleotides C and G. Given this assignment, the sum of a segment of length l containing x occurrences of C+G is equal to $x - p \times l$. Biologists think that this function is a good measure for the CG-richness of that segment. Interestingly enough, all algorithms described in the previous sections can be used to identify the CG-rich segments of a DNA sequence in linear time, according to this objective function. Often, however, biologists prefer to define a cutoff range on the length of the segments for which the maximum sum must be searched, in order to avoid the reporting of extremely short or extremely long segments. In this new scenario the algorithms of the previous sections cannot be applied, but yet linear-time optimal solutions are known for them (see e.g. [2]).
- Assign a value 0 to the nucleotides A and T of the sequence, and a value 1 to the nucleotides C and G. Given this assignment, the density of C+G nucleotides in a segment of length l containing x occurrences of C and G is x/l. Clearly $0 \leq x/l \leq 1$ and every single occurrence of a nucleotide C or G provides a segment with maximum density 1. Biologists consider this as an interesting measure of CG-richness for a segment, provided that a cutoff range on the length of the searched segments is imposed. This problem is more difficult than the one stated in the previous item, nevertheless it posses optimal (quasi-)linear time solutions which are much sophisticated and for which we refer the interested reader to the pertinent bibliography (e.g. [1, 3, 5]).

These examples are useful to highlight a *dangerous trap* that often occurs when abstracting a real problem: apparently small changes in the problem formulation lead to big jumps in the complexity of designing efficient algorithms for them. Think for example to the density function above, we needed to introduce a cutoff lower-bound to the segment length in order to avoid the trivial solution consisting of *single* nucleotides C or G! With this "small" change, the problem results more challenging and its solutions sophisticated.

Other subtle traps are more difficult to be discovered. Assume that we decide to circumvent the single-nucleotide outcome by searching for the the *longest* segment whose density is *larger than* a fixed value d. This is, in some sense, a complementary formulation of the problem stated in the second item above, because maximization is here on the segment length and a (lower) cut-off is imposed on the density value. Surprisingly it is possible to *reduce* this density-based problem to a

sum-based problem, in the spirit of the one stated in the first item above, and solved in the previous sections. Algorithmic reductions are often employed by researchers to re-use known solutions and thus do not re-discover again and again the "hot water". To prove this reduction it is enough to notice that:

$$\frac{\text{Sum}_D[x,y]}{y-x+1} = \sum_{k=x}^{y} \frac{D[k]}{y-x+1} \geq t \iff \sum_{k=x}^{y} (D[k]-t) \geq 0.$$

Therefore, subtracting to all elements in D the density-threshold t, we can turn the problem stated in the second item above into the one that asks for the *longest segment that has sum larger than* 0. Be careful that if you change the request from the *longest segment* to the *shortest one* whose density is larger than a threshold t, then the problem becomes trivial again: Just take the single occurrence of a nucleotide C or G. Similarly, if we fix an upper bound S to the segment's sum (instead of a lower bound), then we can change the sign to all D's elements and thus turn the problem again into a problem with a lower bound $t = -S$. So let us stick on the following general formulation:

Problem. *Given an array $D[1,n]$ of positive and negative numbers, we want to find the* longest *segment in D whose sum of its elements is* larger than *a fixed threshold t.*

We notice that this formulation is in some sense a complement of the one given in the first item above. Here we maximize the segment length and pose a lower-bound to the sum of its elements; there, we maximized the sum of the segment provided that its length was within a given range. It is nice to observe that the structure of the algorithmic solution for both problems is similar, so we detail only the former one and refer the reader to the literature for the latter.

The algorithm proceeds inductively by assuming that, at step $i = 1, 2, \ldots, n$, it has computed the longest sub-array having sum larger than t and occurring within $D[1, i-1]$. Let us denote the solution available at the beginning of step i with $D[l_{i-1}, r_{i-1}]$. Initially we have $i = 1$ and thus the inductive solution is the empty one, hence having length equal to 0. To move from step i to step $i+1$, we need to compute $D[l_i, r_i]$ by possibly taking advantage of the currently known solution.

It is clear that the new segment is either inside $D[1, i-1]$ (namely $r_i < i$) or it ends at position $D[i]$ (namely $r_i = i$). The former case admits as solution the one of the previous iteration, namely $D[l_{i-1}, r_{i-1}]$, and so nothing has to be done: just set $r_i = r_{i-1}$ and $l_i = l_{i-1}$. The latter case is more involved and requires the use of some special data structures and a tricky analysis to show that the total complexity of the solution proposed is $O(n)$ in space and time, thus turning to be optimal!

We start by making a simple, yet effective, observation:

FACT 2.1

If $r_i = i$ then the segment $D[l_i, r_i]$ must be strictly longer than the segment $D[l_{i-1}, r_{i-1}]$. This means in particular that l_i occurs to the left of position $L_i = i - (r_{i-1} - l_{i-1})$.

The proof of this fact follows immediately by the observation that, if $r_i = i$, then the current step i has found a segment that improves the previously known one. Here "improved" means "longer" because the other constraint imposed by the problem formulation is boolean since it refers to a lower-bound on the segment's sum. This is the reason why we can discard all positions within the range $[L_i, i]$, in fact they originate intervals of length shorter or equal than the previous solution $D[l_{i-1}, r_{i-1}]$.

Reformulated Problem. *Given an array $D[1,n]$ of positive and negative numbers, we want to find at every step the* smallest *index $l_i \in [1, L_i)$ such that $Sum_D[l_i, i] \geq s$.*

We point out that there could be many such indexes l_i, here we wish to find the *smallest* one because we aim at determining the *longest* segment.

At this point it is useful to recall that $\text{Sum}_D[l_i, i]$ can be re-written in terms of prefix-sums of array D, namely $\text{Sum}_D[1, i] - \text{Sum}_D[1, l_i - 1] = P[i] - P[l_i - 1]$ where the array P was introduced in Section 2.4. So we need to find the smallest index $l_i \in [1, L_i)$ such that $P[i] - P[l_i - 1] \geq t$. The array P can be pre-computed in linear time and space.

It is worth to observe that the computation of l_i could be done by scanning $P[1, L_i - 1]$ and searching for the *leftmost* index x such that $P[i] - P[x] \geq t$. We could then set $l_i = x + 1$ and have been done. Unfortunately, this is inefficient because it leads to scan over and over again the same positions of P as i increases, thus leading to a quadratic-time algorithm! Since we aim for a linear-time algorithm, we need to spend constant time "on average" per step i. We used the quotes because there is no *stochastic* argument here to compute the average, we wish only to capture syntactically the idea that, since we want to spend $O(n)$ time in total, our algorithm has to take constant time *amortized* per steps. In order to achieve this performance we first need to show that we can avoid the scanning of the whole prefix $P[1, L_i - 1]$ by identifying a *subset* of *candidate positions* for x. Call $C_{i,j}$ the candidate positions for iteration i, where $j = 0, 1, \ldots$. They are defined as follows: $C_{i,0} = L_i$ (it is a dummy value), and $C_{i,j}$ is defined inductively as the *leftmost minimum* of the sub-array $P[1, C_{i,j-1} - 1]$ (i.e. the sub-array to the left of the current minimum and/or to the left of L_i). We denote by $c(i)$ the number of these candidate positions for the step i, where clearly $c(i) \leq L_i$ (equality holds when $P[1, L_i]$ is decreasing).

For an illustrative example look at Figure 2.2, where $c(i) = 3$ and the candidate positions are connected via leftward arrows.

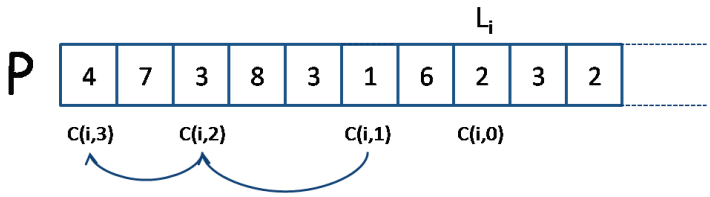

FIGURE 2.2: An illustrative example for the candidate positions $C_{i,j}$, given an array P of prefix sums. The picture is generic and reports only L_i for simplicity.

Looking at Figure 2.2 we derive three key properties whose proof is left to the reader because it immediately comes from the definition of $C_{i,j}$:

Property a. The sequence of candidate positions $C_{i,j}$ occurs within $[1, L_i)$ and moves leftward, namely $C_{i,j} < C_{i,j-1} < \ldots < C_{i,1} < C_{i,0} = L_i$.

Property b. At each iteration i, the sequence of candidate values $P[C_{i,j}]$ is increasing with $j = 1, 2, \ldots, c(i)$. More precisely, we have $P[C_{i,j}] > P[C_{i,j-1}] > \ldots > P[C_{i,1}]$ where the indices move leftward according to Property (a).

Property c. The value $P[C_{i,j}]$ is smaller than any other value on its left in P, because it is the leftmost minimum of the prefix $P[1, C_{i,j-1} - 1]$.

It is crucial now to show that the index we are searching for, namely l_i, can be derived by looking only at these candidate positions. In particular we can prove the following:

FACT 2.2

At each iteration i, the largest index j^* such that $Sum_D[C_{i,j^*} + 1, i] \geq t$ (if any) provides us with the longest segment we are searching for.

By Fact 2.1 we are interested in segments having the form $D[l_i, i]$ with $l_i < L_i$, and by properties of prefix-sums, we know that $\text{Sum}_D[C_{i,j} + 1, i]$ can be re-written as $P[i] - P[C_{i,j}]$. Given this and Property (c), we can conclude that all segments $D[z, i]$, with $z < C_{i,j}$, have a sum *smaller* than $\text{Sum}_D[C_{i,j} + 1, i]$. Consequently, if we find that $\text{Sum}_D[C_{i,j} + 1, i] < t$ for some j, then we can discard all positions z to the left of $C_{i,j} + 1$ in the search for l_i. Therefore the index j^* characterized in Fact 2.2 is the one giving correctly $l_i = C_{i,j^*} + 1$.

There are two main problems in deploying the candidate positions for the efficient computation of l_i: (1) How do we compute the $C_{i,j}$s as i increases, (2) How do we search for the index j^*. To address issue (1) we notice that the computation of $C_{i,j}$ depends only on the position of the previous $C_{i,j-1}$ and *not* on the indices i or j. So we can define an auxiliary array $LMin[1, n]$ such that $LMin[i]$ is the leftmost position of the minimum within $P[1, i - 1]$. It is not difficult to see that $C_{i,1} = LMin[L_i]$, and that according to the definition of C it is $C_{i,2} = LMin[\, LMin[L_i]\,] = LMin^2[L_i]$. In general, it is $C_{i,k} = LMin^k[L_i]$. This allows an incremental computation:

$$LMin[x] = \begin{cases} 0 & \text{if } x = 0 \\ x - 1 & \text{if } P[x-1] < P[\, LMin[x-1]\,] \\ LMin[x-1] & \text{otherwise} \end{cases}$$

The formula above has an easy explanation. We know inductively $LMin[x - 1]$ as the leftmost minimum in the array $P[1, x - 2]$: initially we set $LMin[0]$ to the dummy value 0. To compute $LMin[x]$ we need to determine the leftmost minimum in $P[1, x - 1]$: this is located either in $x - 1$ (with value $P[x - 1]$) or it is the one determined for $P[1, x - 2]$ of position $LMin[x - 1]$ (with value $P[LMin[x - 1]]$). Therefore, by comparing these two values we can compute $LMin[x]$ in constant time. Hence the computation of all candidate positions $LMin[1, n]$ takes $O(n)$ time.

We are left with the problem of determining j^* efficiently. We will not be able to compute j^* in constant time at each iteration i but we will show that, if at step i we execute $s_i > 1$ steps, then we are advancing in the construction of the longest solution. Specifically, we are extending the length of that solution by $\Theta(s_i)$ units. Given that the longest segment cannot be longer than n, the sum of these extra-costs cannot be larger than $O(n)$, and thus we are done! This is called *amortized argument* because we are, in some sense, charging the cost of the expensive iterations to the cheapest ones. The computation of j^* at iteration i requires the check of the positions $LMin^k[L_i]$ for $k = 1, 2, \ldots$ until the condition in Fact 2.2 is satisfied; in fact, we know that all the other $j > j^*$ do not satisfy Fact 2.2. This search takes j^* steps and finds a new segment whose length is *increased* by at least j^* units, given Property (a) above. This means that either an iteration i takes constant time, because the check fails immediately at $LMin[L_i]$ (so the current solution is not better than the one computed at the previous iteration $i - 1$), or the iteration takes $O(j^*)$ time but the new segment $D[L_i, r_i]$ has been extended by j^* units. Since a segment cannot be longer than the entire sequence $D[1, n]$, we can conclude that the total extra-time cannot be larger than $O(n)$.

We leave to the diligent reader to work out the details of the pseudo-code of this algorithm, the techniques underlying its elegant design and analysis should be clear enough to approach it without any difficulties.

References

[1] Kun-Mao Chao. Maximum-density segment. In Ming-Yang Kao, editor, *Encyclopedia of Algorithms*. Springer, 2008.

[2] Kun-Mao Chao. Maximum-scoring segment with length restrictions. In Ming-Yang Kao, editor, *Encyclopedia of Algorithms*. Springer, 2008.

[3] Chih-Huai Cheng, Hsiao-Fei Liu, and Kun-Mao Chao. Optimal algorithms for the average-constrained maximum-sum segment problem. *Information Processing Letters*, 109(3):171–174, 2009.

[4] Rolf Fagerberg. Cache-oblivious model. In Ming-Yang Kao, editor, *Encyclopedia of Algorithms*. Springer, 2008.

[5] Takeshi Fukuda, Yasuhiko Morimoto, Shinichi Morishita, and Takeshi Tokuyama. Mining optimized association rules for numeric attributes. *Journal of Computer System Sciences*, 58(1):1–12, 1999.