Advanced Algorithmics

Thomas Nowak

April 10, 2025

These are partial lecture notes from the undergraduate course Advanced Algorithmics, given by Serge Haddad and Thomas Nowak at ENS Paris-Saclay in the academic year 2024–2025. This document will inevitably contain some errors. If you find any, I will be grateful and happy to hear from you. You can reach me by email at thomas@thomasnowak.net.

Thomas Nowak,
Spring Term: 2024–2025,
Last Update: April 10, 2025,
ENS Paris-Saclay

Contents

April 7, 2025

Lecture 10: Algorithms and Probability I

We start our discussion of randomization in algorithms by going through a number of representative examples that can be studied either by using only elementary probability-theoretic facts or by looking solely at the expected value of certain quantities. The most important technical fact about the expected value that we need is that it is linear: $\mathbb{E}(X+Y) = \mathbb{E}X + \mathbb{E}Y$. This equality holds even if X and Y are not independent.

10.1 Expected Runtime of Quicksort

We recall the quicksort algorithm to sort the list S of distinct elements from a totally ordered universe:

1. If S has at most one element, return S.
2. Choose a pivot element $s \in S$.
3. Compare each element of S to s and construct lists S_1 and S_2 with the elements that are less than, respectively greater than, s.
4. Recursively sort S_1 and S_2.
5. Return S_1, x, S_2.

The deterministic version that chooses the first element of the list as the pivot element has the worst-case running time $\Omega(n^2)$ for lists of length n. If we make the choice in Step 2 an independent uniformly random one, this yields a randomized algorithm. To upper-bound the runtime of randomized quicksort, we first note that it is asymptotically dominated by the number of comparisons. We can then bound the expected number of comparisons:

Theorem 10.1. The expected number of comparisons of randomized quicksort on a list of length n is $2n \log n + O(n)$.

Proof. Let $x_1, \dots, x_n$ be the input values and $y_1, \dots, y_n$ the same values in increasing order. Let X be the number of comparisons, and let X_{ij} be the indicator variable for whether y_i and y_j are ever compared (for $i < j$).

Set $Y_{ij} = \{y_i, \dots, y_j\}$. We have $X_{ij} = 1$ if and only if y_i or y_j is the first pivot element selected from Y_{ij}. We thus have:

$$\begin{aligned}\mathbb{E}X &= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{E}X_{ij} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{2}{j-i+1} = \sum_{k=2}^{n} \sum_{i=1}^{n+1-k} \frac{2}{k} \\ &= \sum_{k=2}^{n} (n+1-k)\frac{2}{k} = (2n+2)\sum_{k=1}^{n} \frac{1}{k} - 4n = 2n \log n + O(n)\end{aligned}$$

Here, we used Lemma 10.1 in the last step. □

The following lemma is useful in many situations. It corresponds to a special case of the Euler–Maclaurin summation formula. Its basic message is that oftentimes it is possible to exchange an integral for a sum.

Lemma 10.1. For all $n \in \mathbb{N}$, we have $\log(n+1) \leq H_n \leq 1 + \log n$.

Proof. For the upper bound, we calculate:

$$\begin{aligned} H_n &= \sum_{i=1}^{n} \frac{1}{i} = 1 + \sum_{i=2}^{n} \frac{1}{i} \int_{i-1}^{i} 1 \, dx = 1 + \sum_{i=2}^{n} \int_{i-1}^{i} \frac{1}{i} \, dx \leq 1 + \sum_{i=2}^{n} \int_{i-1}^{i} \frac{1}{x} \, dx \\ &= 1 + \int_{1}^{n} \frac{1}{x} \, dx = 1 + \log n \end{aligned}$$

For the lower bound, we note:

$$\begin{aligned} H_n &= \sum_{i=1}^{n} \frac{1}{i} = \sum_{i=1}^{n} \frac{1}{i} \int_{i}^{i+1} 1 \, dx = \sum_{i=1}^{n} \int_{i}^{i+1} \frac{1}{i} \, dx \geq \sum_{i=1}^{n} \int_{i}^{i+1} \frac{1}{x} \, dx \\ &= \int_{1}^{n+1} \frac{1}{x} \, dx = \log(n+1) \end{aligned}$$

This concludes the proof. □

10.2 The Secretary Problem

In the secretary problem, we seek to hire a new secretary. We must decide in an online fashion: we interview candidates sequentially, and once we pass on a candidate, we cannot return to that decision. We suppose each of the n candidates has a score, which we can perfectly assess during the interview. We do not assume any *a priori* knowledge of the distribution of the scores. Our goal is to pick the candidate with the highest score. We assume all candidate scores $x_1, x_2, \ldots, x_n \in \mathbb{R}$ are distinct, and that the order of the candidates is uniformly random, *i.e.*, all permutations are equally likely.

At the i^{th} step, the algorithm has seen the scores $x_1, \ldots, x_i$ of the first i candidates. It turns out that the following class of algorithms is optimal: Reject the first m candidates, but keep track of their maximum score. Then, for each of the candidates $m+1, \ldots, n$, hire the first candidate whose score exceeds all previously seen scores.

For this class of algorithms, we now want to determine the best parameter m, *i.e.*, the one that yields the highest probability of hiring the best candidate. Denote by E_i the event that the i^{th} candidate has the highest score and that we hire them. For $i \leq m$, we have $\mathbb{P}(E_i) = 0$, since we do not hire any of the first m candidates. For $i = m+1$, we have $\mathbb{P}(E_{m+1}) = 1/n$, since we hire the highest-scoring candidate in position $m+1$ only if they happen to be there. More generally, for positions $i > m$, we have

$$\mathbb{P}(E_i) = \mathbb{P}(H_i \mid B_i) \cdot \mathbb{P}(B_i)$$

by Bayes' theorem, where H_i is the event that we hire the i^{th} candidate and B_i is the event that the best candidate is in position i. Due to the uniformity assumption, we have $\mathbb{P}(B_i) = 1/n$. Now, to calculate $\mathbb{P}(H_i \mid B_i)$, we note that we hire candidate i if and only if we did not hire anyone before them. This happens precisely when the highest score among the first $i-1$ candidates occurred within the first block of m. We thus have $\mathbb{P}(H_i \mid B_i) = m/(i-1)$. In summary:

$$\mathbb{P}(E_i) = \begin{cases} 0 & \text{if } i \leq m \\ \frac{m}{i-1} \cdot \frac{1}{n} & \text{otherwise} \end{cases}$$

Hence, letting E denote the event that we hire the best candidate, we obtain:

$$\mathbb{P}(E) = \frac{m}{n} \sum_{i=m+1}^{n} \frac{1}{i-1} = \frac{m}{n} (H_{n-1} - H_{m-1})$$

As in the proof of Lemma 10.1, we can show that $H_{n-1} - H_{m-1} \geq \log n - \log m$, so

$$\mathbb{P}(E) \geq \frac{m}{n} (\log n - \log m) \quad . \tag{1}$$

Differentiating this lower bound gives

$$\frac{d}{dm} \frac{m}{n} (\log n - \log m) = \frac{1}{n} (\log n - \log m) - \frac{1}{n} = \frac{\log n - \log m - 1}{n} \quad ,$$

which is zero if and only if $\log m = \log n - 1$, *i.e.*, $m = n/e$. Plugging this into (1) gives $\mathbb{P}(E) \geq 1/e \approx 37\%$.

We did make some approximations here: For one, we optimized a lower bound on $\mathbb{P}(E)$ rather than the probability itself. Also, we cannot choose $m = n/e$ exactly since this is not an integer, so we must round, *e.g.*, to $m = \lfloor n/e \rfloor$. In both cases, we introduce small errors, but asymptotically we remain optimal: The lower bound is tight since we also have the upper bound $\mathbb{P}(E) \leq \frac{m}{n}(\log(n-1) - \log(m-1)) \sim 1/e$ as $n \to \infty$. As for rounding, note:

$$\frac{\lfloor n/e \rfloor}{n} \log \frac{n}{\lfloor n/e \rfloor} \geq \frac{(n-e)/e}{n} \log \frac{n}{n/e} = \left(1 - \frac{e}{n}\right) \frac{1}{e} \sim \frac{1}{e}$$

as $n \to \infty$.

We also left open why this class of algorithms is optimal. Note that whenever an optimal algorithm chooses to hire candidate i, that candidate must be the best seen so far. Otherwise, we forgo a chance to hire the best candidate. Moreover, since we lack information about future scores, we always have $\mathbb{P}(B_i \mid H_i) = i/n$ for optimal algorithms. Since this increases with i, once we begin considering hiring, we must continue to do so. This implies that the choice of m depends only on n.

Thus, we conclude:

Theorem 10.2. An optimal algorithm for the secretary problem has an asymptotic success probability of $1/e$.

10.3 The Online Paging Problem

In the paging problem, we are tasked with managing a cache of size k. We receive a request sequence $\rho_1, \rho_2, \ldots, \rho_n$ of pages in a much larger (virtually infinite) main memory. At each request, if the requested page is not currently in the cache, we can choose which of the pages to evict to make space for the newly requested one. The cache is initially populated with some set of pages. Popular deterministic online algorithms for this problem are LRU (last recently used), FIFO(first-in first-out), and LFU (least frequently used). In contrast to online algorithms, offline algorithms can look at the complete request sequence—in the past as well as into the future—for each decision.

We assume that each cache miss incurs a constant cost, which we normalize to being 1. Formally, for any paging algorithm A and each request sequence $\rho_1, \ldots, \rho_n$,

we define $f_A(\rho_1, \dots, \rho_n)$ to be the number of cache misses when executing A with the request sequence $\rho_1, \dots, \rho_n$. For a randomized algorithm A, this is a random variable.

Since we can always choose a request sequence with $f_A(\rho_1, \dots, \rho_n) = n$ by requesting n different pages, it does not make much sense to look at f_A in the absolute. Rather, we will compare to the theoretical optimum:

Definition 10.1

Let O be an optimal offline algorithm and let $c \in \mathbb{R}_+$. A (randomized) algorithm A is c-competitive if there is some $b \in \mathbb{R}_+$ such that

$$\mathbb{E} f_A(\rho_1, \dots, \rho_n) \leq c f_O(\rho_1, \dots, \rho_n) + b$$

for all n and all request sequences $\rho_1, \dots, \rho_n$.

It turns out that randomization enables an exponential improvement in the competitivity ratio c. To show this, we first start with the lower bound for for deterministic algorithms:

Theorem 10.3. No deterministic online algorithm is c-competitive for $c < k$.

Proof. We construct an infinite request sequence of $k+1$ pages in the main memory, *i.e.*, $\rho_i \in \{1, \dots, k+1\}$ for all $i \in \mathbb{N}$. We do this in an inductive fashion: Initially, we populate the cache with pages $1, \dots, k$. Then, we request the unique page ρ_{i+1} that is not in the cache when executing A on the request sequence $\rho_1, \dots, \rho_i$.

For analysis purposes, we introduce the notion of a round in the request sequence. A round is a maximal subsequence in which at most k different pages are requested. The first round is $\rho_1, \dots, \rho_k$.

An optimal offline algorithm misses at most once in a round: Since at most k different pages are requested, there is one that is not requested. The optimal algorithm can thus choose to evict the non-requested page at the first cache miss. The online algorithm A, on the other hand, misses at least k times (since it misses at every request).

To conclude, we distinguish two cases. If there are infinitely many rounds, then

$$\frac{f_A(\rho_1, \dots, \rho_{n_r})}{f_O(\rho_1, \dots, \rho_{n_r})} \geq \frac{rk}{r} = k$$

where n_r is the last index of round number r. This is a contradiction to A being c-competitive with $c < k$. In the second case, there is only a finite number $R \in \mathbb{N}$ of rounds. Since $f_A(\rho_1, \dots, \rho_n) = n$ by construction and $f_O(\rho_1, \dots, \rho_n) \leq R$ for all $n \in \mathbb{N}$, we have

$$\lim_{n \to \infty} \frac{f_A(\rho_1, \dots, \rho_n)}{f_O(\rho_1, \dots, \rho_n)} \geq \lim_{n \to \infty} \frac{n}{R} = \infty \ ,$$

which is in contradiction to A being c-competitive for any c. □

When considering competitivity of randomized algorithms, multiple different definitions are possible. They differ in the amount of information that can be accessed to construct the request sequence. The procedure that constructs this sequence if often

referred to as an *adversary*. For now, we restrict ourselves to *oblivious* adversaries, who have access to the source code of the algorithm, but not to any of the random choices made during execution. Adversaries that can do that are usually called adaptive.

In the Marker algorithm, each cache location has a marker bit associated with it. It proceeds in rounds. At the start of each round, all marker bits are set to zero. When a request results in a cache hit, the marker bit of that location is set to one. When a request results in a cache miss, a random unmarked location is evicted, the requested page is placed in that location, and the location's marker bit is set to one. The current round ends when all location's marker bits are set to one.

Theorem 10.4. There exists a randomized online algorithm that is $2(1 + H_k)$-competitive against oblivious adversaries.

Proof. We use the Marker algorithm. Let $\rho_1, \dots, \rho_n$ be a request sequence.

Let R be the number of rounds of the Marker algorithm with this request sequence. Call a page *fresh* in round r if it was newly added to the cache by the Marker algorithm in round r. Let f_r be the number of fresh cache locations in round r.

If $r \geq 2$, during rounds $r-1$ and r, there are at least $k + f_r$ different page requests. Of these, at least f_r are cache misses in any algorithm, in particular the optimal one. Setting $\Delta_O(0) = 0$ and $\Delta_O(r) = f_O(\rho_1, \dots, \rho_{n_r}) - f_O(\rho_1, \dots, \rho_{n_{r-1}})$ where n_r denotes the last index of round $r \geq 1$, we have $\Delta_O(r-1) + \Delta_O(r) \geq f_r$ for all $r \geq 1$. We showed this for $r \geq 2$, but it is also true for $r = 1$ since both O and A start with the same pages in the cache. Hence:

$$\begin{aligned} 2f_O(\rho_1, \dots, \rho_n) &\geq 2\sum_{r=1}^{R} \Delta_O(r) \geq \sum_{r=0}^{R-1} \Delta_O(r) + \sum_{r=1}^{R} \Delta_O(r) \\ &= \sum_{r=1}^{R} (\Delta_O(r-1) + \Delta_O(r)) \geq \sum_{r=1}^{R} f_r \end{aligned} \tag{2}$$

We now turn to upper-bounding the expected number of cache misses of algorithm A. Let us focus on a given round r and write $\Delta_A(r) = f_A(\rho_1, \dots, \rho_{n_r}) - f_A(\rho_1, \dots, \rho_{n_{r-1}})$ for the number of cache misses in that round. There are exactly f_r cache misses due to requests for fresh pages. Let $i_1, \dots, i_{k-f_r}$ be the requested non-fresh pages, in the order of the first request to them. Let X_ℓ be the indicator variable that is 1 if and only if the first access to i_ℓ is a cache miss. The expected number of cache misses is:

$$\mathbb{E}\Delta_A(r) = f_r + \sum_{\ell=1}^{k-f_r} \mathbb{E}X_\ell = f_r + \sum_{\ell=1}^{k-f_r} \mathbb{P}(X_\ell = 1) \tag{3}$$

To bound $\mathbb{P}(X_\ell = 1)$, denote by γ the number of fresh pages in the cache before the first access to i_ℓ. The page i_ℓ might have been replaced by a previously accessed page. Compared to the start of the round, exactly γ pages are different. We know that $\ell - 1$ of the pages have to be the same, namely $i_1, \dots, i_{\ell-1}$, but

we don't know the fate of i_ℓ. We thus have:

$$\mathbb{P}(X_\ell = 1) = \frac{\gamma}{k-\ell+1} \leq \frac{f_r}{k-\ell+1} \tag{4}$$

Combining (3) and (4), the expected number of cache misses in round r is at most

$$\mathbb{E}\Delta_A(r) = f_r \left(1 + \sum_{\ell=1}^{k-f_r} \frac{1}{k-\ell+1}\right) \leq f_r(1+H_k) \ . \tag{5}$$

Now, combining (2) and (5) gives

$$\mathbb{E}f_A(\rho_1,\ldots,\rho_n) \leq (1+H_k)\sum_{r=1}^{R} f_r \leq 2(1+H_k)f_O(\rho_1,\ldots,\rho_n)$$

and concludes the proof. □

The Marker algorithm is almost optimal. It can be proved that no randomized online algorithm is c-competitive against oblivious adversaries if $c < H_k$.

April 14, 2025

Lecture 11: Algorithms and Probability II

We now turn to more involved analysis techniques than elementary probability and expected values. In terms of systems under consideration, we focus on randomized data structures.

11.1 Skip Lists

A skip list is a collection of ordered doubly linked lists $L_1, L_2, \ldots, L_r$ such that $\text{Vals}(L_r) = \emptyset$ and $\text{Vals}(L_i) \subseteq \text{Vals}(L_{i-1})$ for all $1 < i \leq r$ where $\text{Vals}(L)$ denotes the set of values in the linked list L. In addition to the horizontal links inside of its linked list L_i, each node also has a vertical link to node in L_{i-1} that contains the same value. Each linked list L_i additionally contains a start node with value $-\infty$ and an end node with value $+\infty$. The skip list encodes the set $\text{Vals}(L_1)$ of values. The linked list L_i is called the i^{th} level of the skip list.

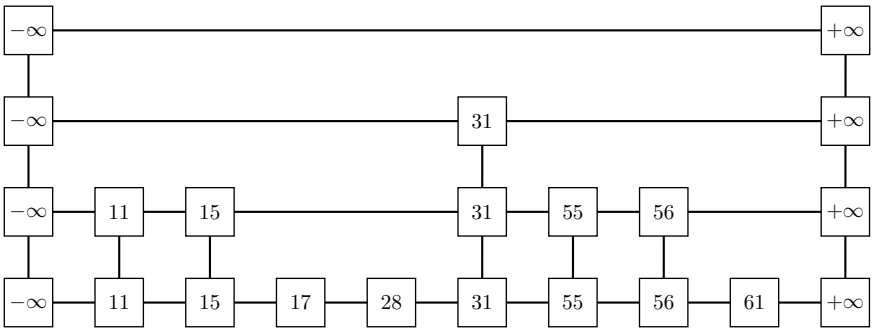

Figure 1: A skip list with $r = 4$ levels containing the values $\text{Vals}(L_1) = \{11, 15, 17, 28, 31, 55, 56, 61\}$.

For a value x in the skip list, we denote by $h(x) = \max\{i \mid x \in \text{Vals}(L_i)\}$ its height, *i.e.*, highest level that contains x. Since $\text{Vals}(L_r) = \emptyset$, we have $h(x) \leq r - 1$ for all values x of the skip list. On the other hand, we have the expression $r = 1 + \max\{h(x) \mid x \in \text{Vals}(L_1)\}$ for the number of levels.

An empty skip list has a single level. Starting from an initially empty skip list, this data structure supports three operations:

- `search`(x): returns true if x is a value in the skip list and false otherwise. The search starts at the $-\infty$ node of L_r. In each iteration, we look at the next node in the current level. If its value is equal to x, we return true. If its value is strictly smaller than x, then we advance to the next node in current level. If its value is strictly larger than x, then we drop down one level if possible. If it's not possible to drop down a level, then we return false.
- `insert`(x): inserts value x into the skip list if it's not already in the list. We first use `search`(x) to find the greatest lower bound and the least upper bound on x in each level. If x is already in the list, we do nothing more and return. Otherwise, we choose a random height $h(x)$ and insert x into the lists $L_1, \ldots, L_{h(x)}$ at the

correct location and add the vertical links. If necessary, *i.e.*, if $h(x) \geq r$, we create new empty levels before inserting x into them.

- `delete`(x): deletes value x if it's in the list. We first use `search`(x) to find look for x in each level. If x is not in any level, we do nothing more and return. Otherwise, we delete x from each level in which we found it. We then delete all empty levels except the top level.

In the description of the operations, we left one thing unspecified: how to choose the random height $h(x)$ in the insertion operation? The most often employed method is to have $h(x)$ be a geometric random variable with parameter $p = 1/2$. That is, the number of fair coin flips until we get one tails. During the analysis of the runtime of the operations, we will see why this is a good choice.

As a warm-up to analyzing the asymptotic performance of the operations, we show that the number of levels of a skip list of n values is $O(\log n)$ with high probability.

The notion "with high probability" is very common, but doesn't always have the same meaning. In its most basic meaning it says that the error probability decreases to zero with $1/n^\alpha$ for some $\alpha > 0$ as $n \to \infty$. Of course, this presupposes the parameter n to measure the size of the system under consideration in some meaningful way. The most useful version of "$X_n = O(f(n))$ with high probability" is "for all $\alpha > 0$ there exist $c > 0$ and $d > 0$ such that $\mathbb{P}\big(X_n > cf(n)\big) \leq d/n^\alpha$ for all (large enough) n". Definitions of "with high probability" that allow a free choice of the exponent α have the advantage of being composable much more easily.

Lemma 11.1. Let $\alpha > 1$. In a skip list of n values, we have $\mathbb{P}(r > \alpha \log_2 n) \leq 4/n^{\alpha-1}$ for the number r of levels.

Proof. The random variables $h(x)$ for $x \in \mathrm{Vals}(L_1)$ are i.i.d. geometric with parameter $p = 1/2$. This is true no matter the order of the preceding operations: the height chosen during the insertion of each of the currently present values is independent of all previous and later operations.

Let us write $\{x_1, x_2, \ldots, x_n\}$ for the set of values in the skip list. For the number of levels, we already established the formula $r = 1 + \max\{h(x_j) \mid 1 \leq j \leq n\}$. We thus have

$$\begin{aligned} \mathbb{P}(r > \alpha \log n) &= \mathbb{P}\big(1 + \max\{h(x_j) \mid 1 \leq j \leq n\} > \alpha \log_2 n\big) \\ &= \mathbb{P}\big(\exists 1 \leq j \leq n\colon\ h(x_j) > \alpha \log_2 n - 1\big) \\ &\leq n \cdot \mathbb{P}\big(h(x_1) > \alpha \log_2 n - 1\big) \end{aligned} \tag{6}$$

where we used the union bound $\mathbb{P}(E_1 \cup \cdots \cup E_n) \leq \mathbb{P}(E_1) + \cdots + \mathbb{P}(E_n)$.

Now, using the specific distribution of the height $h(x_1)$, we have $\mathbb{P}\big(h(x_1) > t\big) = (1-p)^t = 1/2^t$ for every nonnegative integer t. From this, we can deduce an bound that is true for all real numbers $t \geq 1$:

$$\mathbb{P}\big(h(x_1) > t - 1\big) = \mathbb{P}\big(h(x_1) > \lfloor t \rfloor - 1\big) = 1/2^{\lfloor t \rfloor - 1} \leq 1/2^{t-2}$$

Here, we used that $n > x$ if and only if $n > \lfloor x \rfloor$ for all integers n and real numbers x, and the fact that $\lfloor t \rfloor - 1 \geq t - 2$. We also see that the bound trivially remains true for the remaining values of t, *i.e.*, for $t < 1$ since then $1/2^{t-2} > 1/2^{-1} = 2$, which is bigger than any probability.

Plugging in $t = \alpha \log_2 n$, we get:

$$\mathbb{P}\big(h(x_1) > \alpha \log_2 n - 1\big) \leq 1/2^{\alpha \log_2 n - 2} = 4/n^\alpha$$

Combining this with (6) now concludes the proof. □

For the complete analysis of the time complexity of the three operations of a skip list, we will need the arguably most important inequality for the analysis of randomized algorithms, the Chernoff bound. It is an upper bound on the probability of a sum of Bernoulli random variables being far from its expected value.

Theorem 11.1 (Chernoff bound)**.** Let $X_1, X_2, \ldots, X_m$ be mutually independent 0/1 random variables and let $X = \sum_{k=1}^m X_k$. Then

$$\mathbb{P}\big(X \leq (1-\delta)\mu\big) \leq e^{-\mu\delta^2/2}$$

for all $0 < \delta < 1$ where $\mu = \mathbb{E}\, X$.

We can now prove that the search operation takes only logarithmic time with high probability. Since the time complexity of the insertion and deletion operations are dominated by the search, we get the same result for all three operations of the skip list.

Theorem 11.2. The time complexity of `search`(x) in a skip list of n values is $O(\log n)$ with high probability.

Proof. We bound the number of iterations of the search operation, which asymptotically dominates the time complexity. Independently of whether the searched value x was found, we consider the node visited in the last iteration and work our way backwards to the initial $-\infty$ node of level L_r.

The claimed bound of $m = O(\log n)$ on the number m if iterations comes from the symmetry of following vertical or horizontal links due to the choice of the fair parameter $p = 1/2$ for the heights, and the already established height bound from Lemma 11.1. Once we reach the highest level L_r, we are back at the node of the first iteration.

We are thus done if we find a constant d such that a sequence of at least $m = d \log n$ fair coin flips (iterations) has at least $\alpha \log_2 n = c \log n$ heads (vertical moves). By Theorem 11.1, the probability of this event is upper-bounded by

$$\mathbb{P}\left(X \leq (1-\delta)\frac{m}{2}\right) \leq e^{-m\delta^2/4}$$

where $(1-\delta)\frac{m}{2} = c \log n$. Solving for δ, we get $\delta = 1 - 2c/d$. Choosing $d = 4c$, we get $\delta = 1/2$ and thus:

$$\mathbb{P}\big(\leq c \log n \text{ vertical moves}\big) \leq \mathbb{P}\left(X \leq c \log n\right) \leq e^{-d \log n} = e^{-4c \log n}$$

Now, putting things together, we have:

$$\begin{aligned}\mathbb{P}\left(\geq d\log n \text{ iterations}\right) &\leq \mathbb{P}\left(\leq c\log n \text{ vertical moves}\right) + \mathbb{P}\left(r > c\log n\right)\\ &\leq \frac{1}{n^{4c}} + \frac{4}{n^{\alpha-1}}\end{aligned}$$

For any $\alpha > 1$, we have $c = \alpha/\log 2 > \alpha$ and thus:

$$\mathbb{P}\left(\geq d\log n \text{ iterations}\right) \leq \frac{1}{n^{4\alpha}} + \frac{4}{n^{\alpha}} \leq \frac{5}{n^{\alpha}}$$

This concludes the proof. □

11.2 Universal Hashing

Consider a universe U of keys of items that we want to store in a hash table. A hash function is a function $h\colon U \to V$ where $V = \{0, 1, \ldots, m-1\}$ and $|U| > m$. In the worst case, an n-element hash table can take $\Omega(n)$ time to search for an element; for example if all elements are hashed to the same value and end up in the same linked list. This is true no matter the hash function chosen. The idea of universal hashing is to chose a random hash function to circumvent the deterministic worst case. For that, we need to look at families of hash functions from which we will make the random choice.

Definition 11.2

A family $\mathcal{H}$ of hash functions $h\colon U \to V$ is *universal* if

$$\mathbb{P}\big(h(x) = h(y)\big) \leq \frac{1}{m}$$

for all $x, y \in U$ with $x \neq y$ where the hash function h is chosen uniformly at random in $\mathcal{H}$.

The use of universal families of hash functions allows to cut down on the length of the linked lists in the hash table. Here is a result on the expected number of collisions:

Theorem 11.3. In an n-element hash table S with the hash function h chosen uniformly at random from a universal family, for every element $x \in U$ of the universe, we have:

$$\mathbb{E}\,|\{y \in S \mid h(x) = h(y)\}| \leq \begin{cases} n/m & \text{if } x \notin S \\ 1 + (n-1)/m & \text{if } x \in S \end{cases}$$

Proof. Let $y_1, y_2, \ldots, y_n$ be the elements of S and let X_i be the indicator variable of the event $h(y_i) = h(x)$. Since $\mathcal{H}$ is universal, we have $\mathbb{E}\,X_i = \mathbb{P}(X_i = 1) \leq 1/m$ if $x \neq y_i$. Now, if $x \notin S$, then:

$$\mathbb{E}\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} \mathbb{P}(X_i = 1) \leq \frac{n}{m}$$

On the other hand, if $x = y_j$ for some j, then

$$\mathbb{E}\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} \mathbb{P}(X_i = 1) = 1 + \sum_{\substack{1 \le i \le n \\ i \ne j}} \mathbb{P}(X_i = 1) \le 1 + \frac{n-1}{m} \ ,$$

which concludes the proof. □

Universal families of hash functions are clearly useful, but we have not yet answered the question whether any actually exists. This question is quite easily answered in the affirmative via the probabilistic method. A harder question is whether we can find a universal family that we can easily sample from and whose hash functions are simple to compute. Fortunately, the answer to this question is also yes. There are a few constructions known, here is one of them:

Let us assume that $U = \{0, 1, \dots, u-1\}$ and choose a prime number $p \ge u$. Then, for integers $0 \le a, b < p$, we define the hash function

$$h_{a,b}(x) = \big((ax + b) \bmod p\big) \bmod m$$

and define the family $\mathcal{H} = \{h_{a,b} \mid 0 \le a, b < p\}$.

Theorem 11.4. The family $\mathcal{H}$ is universal.

Proof. Let $x, y \in U$ with $x \ne y$.

For every pair $(u, v) \in \{0, \dots, p-1\}^2$ with $u \ne v$, there is a unique pair $(a, b) \in \{1, \dots, p-1\} \times \{0, \dots, p-1\}$ with $ax + b \equiv u \mod p$ and $ay + b \equiv v \mod p$. In fact, we have the formulas $a \equiv (v-u)\cdot(y-x)^{-1} \mod p$ and $b \equiv u - ax \mod p$.

It thus suffices to upper bound the number of pairs (u, v) with $u \not\equiv v \mod p$ but $u \equiv v \mod m$. If we fix a $u \in \{0, \dots, p-1\}$, then there are at most $\lceil p/m \rceil - 1 \le (p-1)/m$ possible values for v. In total, we thus have at most $p(p-1)/m$ such pairs. Since each pair corresponds to a unique hash function $h_{a,b}$ in $\mathcal{H}$, we have

$$\mathbb{P}\big(h_{a,b}(x) = h_{a,b}(y)\big) \le \frac{p(p-1)/m}{p(p-1)} = \frac{1}{m} \ ,$$

which shows that $\mathcal{H}$ is a universal family. □

We now turn to the question whether we can completely avoid collisions. In classical hash tables, it turns out that can only be guaranteed when $m \ge n^2$, which leads to a prohibitive space complexity of $\Omega(n^2)$. It *is*, however, possible to get rid of collisions if we introduce a second level of hashing for each slot. This is a technique called *perfect hashing* and is able to guarantee a worst-case time complexity of searches of $O(1)$ while keeping an $O(n)$ space complexity. We will analyze the static variant, which does not support insertions or deletions, but dynamic variants also exist.

For the analysis of perfect hashing, we use Markov's inequality, which can be quite loose, but is often enough to prove asymptotic probability bounds. It is also the basis for more advanced bounds, including Chebyshev's inequality and the Chernoff bound.

Theorem 11.5 (Markov's inequality). Let X be a nonnegative random variable

and let $a > 0$. Then:

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}\, X}{a}$$

Lemma 11.2. In an n-element hash table S with hash function h chosen uniformly at random from a universal family, if $m \geq n^2$, then the probability of having no collisions is at least $1/2$.

Proof. Let $y_1, y_2, \ldots, y_n$ be the elements of S and let $X_{i,j}$ be the indicator variable of the event $h(y_i) = h(h_j)$. Defining X to be the number of collisions, we have:

$$\mathbb{E}\, X = \sum_{1 \leq i < j \leq n} \mathbb{P}(X_{i,j} = 1) \leq \frac{n(n-1)}{2} \frac{1}{m} \leq \frac{n^2}{2m}$$

Now, applying Markov's inequality (Theorem 11.5) with $a = n^2/m$ concludes the proof since $n^2/m \leq 1$ by assumption. □

The construction of a 2-level perfect hash table is as follows:

1. Pick a hash function $h_1 \in \mathcal{H}_m$ uniformly at random where $m = n$. Denote by ℓ_j the number of elements hashed to value j. If $\sum_{j=1}^{m} \ell_j^2 > cn$, then pick another hash function and check again.
2. For every $j \in \{0, \ldots, m-1\}$, pick a hash function $h_{2,j}$ uniformly at random from $\mathcal{H}_{m_j}$ where $m_j = \ell_j^2$. If there is a conflict for one of the hash functions $h_{2,j}$, pick another and check for conflicts again.

If this algorithm terminates, the perfect hash table allows searches in $O(1)$ time since there are no conflicts in the second level. Moreover, its space complexity is $O(n) + \sum_{j=1}^{m} O(\ell_j^2) = O(n)$.

Theorem 11.6. The construction of a perfect hash table terminates in time $O(n \log^2 n)$ with high probability.

Proof. Picking and checking a hash function in $\mathcal{H}_m$ can be done in time $O(n)$. Denoting by $X_{i,j}$ the indicator function of the event $h_1(y_i) = h_i(y_j)$, we have

$$\mathbb{E} \sum_{j=0}^{m-1} \ell_j^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{E}\, X_{i,j} \leq n + 2\frac{n(n-1)}{2} \frac{1}{m} \leq \frac{c}{2} n$$

for some constant $c > 0$. Application of Markov's inequality (Theorem 11.5) now shows $\mathbb{P}\big(\sum_{j=0}^{m-1} \ell_j^2 > cn\big) \leq 1/2$. We can thus see the picks of h_1 as i.i.d. coin flips with success probability $p \geq 1/2$. The number of coin flips until the first success is $O(\log n)$ with high probability.

Picking and checking each $h_{2,j}$ can be done in time $O(\ell_j)$. An application of the Chernoff bound (Theorem 11.1) shows $\ell_j = O(\log n)$ with high probability. Picking one set of $h_{2,j}$ thus takes time $O(n \log n)$ with high probability. As above, Lemma 11.2 shows that we need to pick the $h_{2,j}$ at most $O(\log n)$ times with high probability. This means that constructing the second level is done in

$O(n \log^2 n)$ time with high probability. □

April 28, 2025

Lecture 12: Distributed Algorithms I

The analysis of distributed systems is of high importance due to the ubiquity of systems in which communication is necessary: from the Internet all the way to systems-on-chip. One of the most difficult tasks in this analysis is that of modeling the system. In contrast to centralized computation, no single standard model exists and a small change in the model assumptions can lead to widely different properties.

12.1 Modeling: Synchronous Message Passing

One of the simplest models of distributed computation is that of synchronous message passing. In this model, processes are assumed to repeatedly execute rounds in a lock-step fashion. The steps in a round are send–receive–compute: first all processes send a set of messages, then they receive the message just sent, and then they perform a local computation. Most commonly, the local processes are assumed to be computationally unbounded. This assumption is made since local computations are rarely very heavy in the type of algorithms normally studied. An exception is the assumption of unbreakable cryptographic primitives, for which a Turing machine model would not be very useful either.

Formally, a distributed system in this model has:

- a directed or undirected graph $G = (V, E)$
- a message alphabet M with $\bot \notin M$
- for every process $p \in V$:
  - a set States_p of local states
  - a nonempty set $\text{Start}_p \subseteq \text{States}_p$ of initial states
  - a message-generating function $\text{Msgs}_p \colon \text{States}_p \times \text{Out}_p \to M \cup \{\bot\}$
  - a state-transition function $\text{Trans}_p \colon \text{States}_p \times (M \cup \{\bot\})^{\text{In}_p} \to \text{States}_p$

We construct an *execution* as follows. We initialize the local state of every process p to some initial state in Start_p. Then, in each round $r \geq 1$:

1. Apply the message-generating functions Msgs_p to the current local state of each process p to find the message sent over every link (p, q).
2. Gather all messages sent to each process q and apply the state-transition function Trans_q to find the next local state.

The main complexity measures are time and communication complexity. Time complexity is measured in the number of rounds until the desired state. Communication complexity is most often measured in total number of messages sent. Sometimes it is also measured in the total number of bits of messages sent.

A desirable property of our algorithms is that they actually halt, *i.e.*, stop executing. Formally, a halting state is one in which no messages are sent and no state transition from the halting state is possible.

Models both with and without process identifiers exist. In models without process identifiers, the processes know their neighbors only by their port numbers. If they are assumed, process identifiers take the form of unique identifiers (UIDs), which are unique in the graph and are often assumed to be encoded in $O(\log n)$ bits.

If we allow randomized algorithms, we allow the state-transition function to return a probability distribution on the set of states instead of a single state.

12.2 Breadth-First Search

The problem of constructing a distributed breadth-first search (BFS) tree can be formalized as follows: each process maintains a parent variable, which can hold the UID of a process. A unique root process should encode itself as its parent. Any other process should encode its parent in the BFS tree rooted at the root.

For this problem, we assume a connected undirected graph G. We assume the existence of process UIDs and the existence of a unique distinguished leader process p_0, but specific knowledge of the graph G (*e.g.*, number of nodes, diameter). The assumption of the existence of a leader process might seem strong, but a simple flooding algorithm can determine a leader in $O(\mathrm{diam}(G))$ rounds.

The SynchBFS algorithm has some set of processes be marked at each round. Initially, only the leader process p_0 is marked and all parent variables are null except for the leader process, which has itself set as its parent. In every round, each process that became marked in the last round sends a search message to all its neighbors. Then, every unmarked process that received a search message from one of its neighbors chooses one of these neighbors to be its parent and becomes marked.

Theorem 12.1. The SynchBFS algorithm constructs a BFS tree in $O(\mathrm{diam}(G))$ rounds and sends $O(|E|)$ messages.

Proof. We have the following invariant at the end of every round $r \geq 1$, which can be proved by induction on r: for every $1 \leq d \leq r$, every process at distance d from p_0 is marked and has its parent set to a process at distance $d-1$ from p_0.

The correctness of the constructed tree and the bound on the time complexity then follow from this invariant at round $r = \mathrm{diam}(G)$. The message complexity bound follows from the fact that every process p sends exactly $\deg(p)$ messages. The total number of messages is thus

$$\sum_{p \in V} \deg(p) = 2|E| \ ,$$

which concludes the proof. □

Once we constructed a BFS, we can broadcast messages through the tree in $O(\mathrm{diam}(G))$ rounds. We can even add a routing table to the nodes to reduce the message complexity to $O(\mathrm{diam}(G))$ for each point-to-point message. This is obvious for messages sent by the root process, but every other process can first send the to-be-sent message to the root, which will then forward it to the right process. Termination can be implemented since we can locally detect whether a process is a leaf in the BFS tree: in the round following their search message, they don't receive another search message. A termination signal can then be propagated up the BFS tree.

12.3 Maximal Independent Set

We now turn to the problem of constructing a maximal independent set (MIS). An independent set is a set of vertices that does not contain any two neighbors. We formalize a distributed MIS construction by requiring the existence of a local Boolean variable that indicates whether the process is inside the MIS or not.

We again assume a connected undirected graph G. The algorithm that we will study does not need UIDs or any other knowledge of the graph G. It does, however,

work with real-valued variables. To achieve a finite bit complexity, one can show that using $O(\log n)$ bits of the real numbers is enough. To utilize this, one needs knowledge of n, or of an upper bound on n.

In the LubyMIS algorithm, each process maintains a list of remaining neighbors, initialized to all its neighbors in the graph. It proceeds in phases of three rounds each:

1. In the first round, each active process sends a uniformly chosen random value in the interval $[0, 1]$ to all neighbors. Each active process then determines whether it is a *winner*, *i.e.*, whether it has a value strictly larger than all its neighbors.
2. In the second round, each winner notifies its neighbors of the fact that they are *losers*.
3. In the third round, each winner joins the MIS. Then all winners and all losers stop executing the algorithm and become inactive. All neighbors of losers remove the losers from their list of remaining neighbors.

For analysis purposes, we define a sequence $G_0, G_1, \ldots$ of subgraphs of G. The graph $G_\phi = (V_\phi, E_\phi)$ is the subgraph of G induced by the set of active processes at the end of phase $\phi \geq 1$. Once all processes became inactive, the graphs G_ϕ are empty. Initially, we set $G_0 = G$.

Lemma 12.1. The expected number of edges removed in phase ϕ is at least $|E_{\phi-1}|/2$.

Proof. We say that process r *single-handedly* kills edge $\{p, q\}$ from p's side if r is a neighbor of p and process r's value is maximal among those of $N(p) \cup N(q)$.

Now consider any edge $\{p, q\} \in E_{\phi-1}$. The probability of $\{p, q\}$ being single-handedly killed from p's side by one of p's neighbors is equal to:

$$\frac{|N(p)|}{|N(p) \cup N(q)|} \geq \frac{\deg(p)}{\deg(p) + \deg(q)}$$

Each edge in $E_{\phi-1}$ can be single-handedly killed at most twice; once from each side. The expected number X of edges removed in phase ϕ is:

$$\begin{aligned}
\mathbb{E}\, X &= \sum_{\{p,q\} \in E_{\phi-1}} \mathbb{P}\big(\{p, q\} \text{ removed}\big) \\
&\geq \sum_{\{p,q\} \in E_{\phi-1}} \frac{1}{2} \left(\mathbb{P}\big(\{p, q\} \text{ s.h. killed from } p\big) + \mathbb{P}\big(\{p, q\} \text{ s.h. killed from } q\big)\right) \\
&\geq \sum_{\{p,q\} \in E_{\phi-1}} \frac{1}{2} \left(\frac{\deg(p)}{\deg(p) + \deg(q)} + \frac{\deg(q)}{\deg(p) + \deg(q)}\right) = \frac{|E_{\phi-1}|}{2}
\end{aligned}$$

This concludes the proof. □

Theorem 12.2. If the LubyMIS algorithm terminates, it has constructed an MIS. Furthermore, the expected number of rounds until termination is $O(\log n)$.

Proof. Let I_ϕ be the set of processes that joined the constructed MIS in phase ϕ and let $J_\phi = \bigcup_{\psi=1}^{\phi} I_\psi$ be the constructed MIS up to the end of phase ϕ. We

have the following invariant at the end of every phase $\phi \geq 1$: The set J_ϕ is an independent set in G. Furthermore, the processes that have been removed in phase ϕ are those processes in $J_\phi \cup N(J_\phi)$ that have not already been removed in earlier phases, which means that $V_\phi = V \setminus (J_\phi \cup N(J_\phi))$.

If the algorithm terminates, independence of the constructed set now directly follows from the first part of the invariant. Maximality of the constructed independent set follows from the second part of the invariant.

We now turn to the time-complexity bound. Markov's inequality (Theorem 11.5) with $a = 3|E_{\phi-1}|/4$ and Lemma 12.1 gives:

$$\mathbb{P}\left(|E_\phi| \geq \frac{3}{4}|E_{\phi-1}|\right) \leq \frac{2}{3} \quad \Rightarrow \quad \mathbb{P}\left(|E_\phi| < \frac{3}{4}|E_{\phi-1}|\right) \geq \frac{1}{3}$$

Let us call phase ϕ a *success* if $|E_\phi| < \frac{3}{4}|E_{\phi-1}|$. Starting from $|E_0| \leq n^2$, we have $|E_\phi| < 1$, *i.e.*, $E_\phi = \emptyset$ after we have had at least

$$s > \frac{2}{\log \frac{4}{3}} \log n = c \log n$$

successes. If $E_\phi = \emptyset$, then the algorithm terminates in phase $\phi + 1$ at the latest. Now, using the Chernoff bound (Theorem 11.1) for $d \log n$ coin flips with success probability $1/3$ gives

$$\mathbb{P}\left(E_{\lfloor d \log n \rfloor} \neq \emptyset\right) \leq \mathbb{P}\left(s \leq c \log n\right) \leq e^{-\alpha \log n} = \frac{1}{n^\alpha}$$

where $\alpha = \delta^2/6$ and $1 - \delta = 3c/d$. In particular, for a given $\alpha > 0$ we can find an appropriate d. □

The time complexity of LubyMIS is asymptotically almost optimal, as there is a lower bound of $\Omega\left(\sqrt{\log n / \log \log n}\right)$, even if UIDs are allowed.

12.4 Coloring of Paths

We want to color a path graph with $\Delta + 1 = 3$ colors. Formally, we give every process a local color variable that can take the values 1, 2, or 3. Initially, the variables are initialized to null. The goal is to have neighboring nodes pick different colors. We assume no knowledge of the number n of processes, but do assume UIDs. We will study three different algorithms for this problem.

The first algorithm, which we call LocalLeader, has processes exchange their colors and UIDs. Undecided processes then decide whether they are a local leader, *i.e.*, whether their UID is larger than those of its undecided neighbors. A local leader then picks an arbitrary color that has not yet been picked by a neighbor.

Theorem 12.3. The LocalLeader algorithm constructs a 3-coloring of a path of length n in $O(n)$ rounds.

Proof. Correctness follows from a simple invariant. For the time complexity, we note that, before termination, there is at least one local leader. □

This bound is asymptotically tight: in the case of increasing UIDs along the path, the LocalLeader algorithm takes $\Omega(n)$ rounds.

The second algorithm, which we call RandomColor, has every undecided process pick a random color and send it to its neighbors. Then, if the received colors are locally consistent, it adopts its current color definitely.

Theorem 12.4. The RandomColor algorithm constructs a 3-coloring of a path of length n in $O(\log n)$ rounds with high probability.

Proof. The correctness of the algorithm again follows from a simple invariant. For the time-complexity bound, we note that every process has a probability of at least $1/3$ to pick a color in every round in which it is active. By a non-standard application of the Chernoff bound, we conclude that every process pick a color in $O(\log n)$ rounds with high probability. An application of the union bound now concludes the proof. □

The third algorithm, which we call IDReduction, starts with a large set of colors, which it then reduces over time until only three are left. The initial, large, set of colors is the set of UIDs. We can allow the UIDs to be very large positive integers; for simplicity we assume that they are bounded by a polynomial in n. Then, each process sends its current color to its right neighbor. (We do need to assume that our path is directed for now.) It then chooses a new color equal to $2i+b$ where i is the position of the first bit where the two IDs differ, and b is the bit of its own color at that position. This can reduce the number of colors to at most 6, since 7 is the smallest positive integer m for which $2\log_2 m + 1 < m$ We then use the LocalLeader algorithm to find a 3-coloring in $6 - 3 = 3$ more rounds. In fact, we didn't need globally unique IDs, but only locally unique ones.

Theorem 12.5. The IDReduction algorithm constructs a 3-coloring of a directed path of length n in $O(\log^* n)$ rounds.

Proof. For correctness, we prove the invariant of the first phase of the algorithm that the IDs form a $f^r(N)$ coloring at the end of round $r \geq 1$, where $f(m) = 2\log_2 m + 1$ and N is the largest UID, which is $\leq n^k$ for some k by assumption. For the induction step, we assume by contradiction that $2i + b = 2i' + b'$ for the new colors of two neighbors in the path. We then must have $b = b'$ and $i = i'$. In particular, i is the position of the first bit in which the old IDs differed. But since they did differ, we have $b \neq b'$, a contradiction.

The time complexity bound immediately follows from the invariant at round number $r = O(\log^* n)$. □

May 5, 2025

Lecture 13: Distributed Algorithms II

After having studied synchronous systems, we now turn our attention to asynchronous systems, in which we don't have a common time base. There are two main reasons for asynchronous models: when some processes or messages can be very slow, *e.g.*, when modeling the Internet, and when some processes or messages can be very quick and we don't want to wait for the slowest one.

13.1 Modeling: Asynchronous Message Passing

A configuration is a collection of one local state for each process and the state of the communication system. In our case, the state of the communication system is the multiset of messages that were sent but not yet received. A message is a pair (p, m) of a process p and a message value $m \in M$.

Computational steps of processes now happen in a possibly non-synchronous manner. Each step of a process includes: the reception of one or zero messages, a local state change, and the sending of a set of messages to its neighbors. Contrary to the synchronous model, local steps thus proceed in a receive–compute–send manner. A step of process p is triggered by an event $e = (p, m)$ where $m \in M \cup \{\bot\}$. If $m = \bot$, then no message is received by process p in the step. Otherwise, process p can read the message contents and have its state transition depend on it. The message (p, m) is then removed from the set of in-transit messages and the set of messages sent by p in the step is added to it. Given a configuration C and an event $e = (p, m)$ where either $m = \bot$ or (p, m) is an in-transit message, then we can apply the event e to configuration C and determine the successor configuration $e(C)$. Formally, for the description of the system, we need:

- a directed or undirected graph $G = (V, E)$
- a message alphabet M with $\bot \notin M$
- for every process $p \in V$:
  - a set States_p of local states
  - a nonempty set $\text{Start}_p \subseteq \text{States}_p$ of initial states
  - a state-transition function $\text{Trans}_p \colon \text{States}_p \times (M \cup \{\bot\}) \to \text{States}_p$
  - a message-generating function $\text{Msgs}_p \colon \text{States}_p \to 2^{\text{Out}_p \times M}$

An execution is a sequence of configurations such that every process starts in one of its initial states and each successor configuration is the result of the application of an applicable event. An execution is admissible if:

- No message is received by p more often than it was previously sent to p. (Safety)
- Every message is eventually received. (Communication liveness)
- Every process takes infinitely many steps. (Process liveness)

We would want our algorithms to be well-behaved in all admissible executions. In the synchronous model, there was a single execution for each initial configuration. This is now very different in the asynchronous model: the number of admissible executions is always uncountable if $|V| \geq 2$. Due to this non-determinism, people often find it useful to think about algorithm design as a two-player game: first the algorithm designer

presents the algorithm, then the adversary tries to find an admissible execution in which the algorithm misbehaves. People are also sometimes tempted to convert the non-determinism into a probabilistic choice. However, due to the difficulty of coming up with realistic probability distributions for message-delivery and scheduling choices, this only makes sense in very specific cases.

Message complexity is measured the same way as in synchronous systems: by counting the total number of messages sent. It is not immediately obvious how to define time complexity in an asynchronous system, however. The approach that is usually taken is to add real times in $\mathbb{R}_+$ to each configuration in each admissible execution in a nondecreasing fashion, with the additional constraints that the initial configuration is at time 0 and that the real-time difference between process activations, as well as a message send and its reception, is at most 1. The time complexity is then defined as the supremum of all real times of the desired event when real-time tagging according to the above rules.

13.2 Breadth-First Search

We return to the first problem that we studied for synchronous message passing—the construction of a BFS tree—and see whether and how we can solve it in asynchronous systems. As in Section 12.2, we assume an undirected connected graph G and the existence of a leader node p_0.

We can't use the SynchBFS algorithm directly in an asynchronous system because it relied on the fact that, after d rounds, all processes at distance at most d from p_0 received a search message. There is no way to make this time-based property work in an asynchronous system. We will thus need to pack some distance information into the messages that we send. One possibility is to send our current distance from p_0 to our neighbors and have them update their parent variable to us if we can offer them a shorter path than they currently have, akin to the Bellman–Ford algorithm.

This leads us to the definition of the AsynchBFS algorithm: Every process maintains a parent and a distance variable. Initially, all processes except p_0 have parent equal null and distance equal $+\infty$. The leader process p_0 initializes parent to itself and distance to 0. In the first step of p_0, it sends a message with content '0' to all its neighbors. Upon reception of a message, each process compares the received value m to its current distance d. If $m + 1 < d$, then it updates its parent variable to the process from which it received the value m and its distance to $m + 1$. It then sends this new distance to all neighbors.

Theorem 13.1. The AsynchBFS algorithm constructs a BFS tree in $O(\mathrm{diam}(G))$ time and sends $O(n \cdot |E|)$ messages.

Proof. Denote by dist_p the value of the distance variable at process p. For correctness of the constructed tree, we can show the following invariants:

1. If $\mathrm{dist}_p < +\infty$, then dist_p is the length of some path from p_0 to p.
2. Every message sent by process p is the length of some path from p_0 to p.
3. For every edge $\{p, q\} \in E$, either $\mathrm{dist}_q \leq \mathrm{dist}_p + 1$ or process p has sent a message with content dist_p to q.

The time bound follows from invariant (3) since it implies that, after time d_p,

we have $\text{dist}_p = d_p$ where d is the distance of p from p_0. The message-complexity bound follows from the fact that every process sends at most n different values, since there are at most n possible distances. □

We proved that the AsynchBFS algorithm stabilizes to a BFS tree, but we have not discussed how to make processes decide and terminate in finite time. This can be achieved by adding acknowledgment messages and propagating a termination signal to the leader process once the exploration of all neighbors is done. The bookkeeping involved in this is a bit tedious, but it can be done in the same asymptotic complexity.

13.3 Modeling: Process Faults

One of the fundamental reasons for distribution is fault-tolerance. That is, we want to have our system function even if some of its components fail. Faults come in all kinds of varieties. For now, we will focus on crash faults, *i.e.*, having process stop taking steps during executions. These faults are difficult to tolerate in asynchronous systems, since it can be impossible to distinguish a slow from a crashed process.

To every execution, we associate a set F of faulty processes with $|F| \leq f$ such that:

- No message is received by p more often than it was previously sent to p. (Safety)
- Every message sent to a non-faulty process is eventually received. (Communication liveness)
- Every non-faulty process takes infinitely many steps. (Process liveness)

We assume that F is minimal with this property. That is, we do not flag a process faulty if it takes infinitely many steps and receives all messages sent to it.

13.4 Impossibility of Consensus in Asynchronous Systems with Process Faults

It turns out that a large class of problems are unsolvable in asynchronous systems with process faults, even for complete graphs G, which we assume in the following. A particularly striking example is the consensus problem. In the consensus problem, every process starts with an initial value $v_p \in \{0, 1\}$. We want all processes to decide on a common value. For this, we equip every process with a write-once decision variable. In every execution of a consensus algorithm, we require the following three properties:

- No two decision values are different. (Agreement)
- If all initial values are the same, then this value is the only possible decision value. (Validity)
- All non-faulty processes decide. (Termination)

We start with a general lemma about asynchronous systems. For its formulation, we extend the application of an event to sequences of events, which we call schedules. If a schedule σ is applicable to a configuration C, then we write $\sigma(C)$ for the resulting configuration if σ is finite and for the resulting execution postfix if σ is infinite.

Lemma 13.1. If σ_1 and σ_2 are applicable to configuration C and the sets of processes taking steps in σ_1 and σ_2 are disjoint, then:

1. σ_2 is applicable to $\sigma_1(C)$

2. σ_1 is applicable to $\sigma_2(C)$
3. $\sigma_2(\sigma_1(C)) = \sigma_1(\sigma_2(C))$

For the analysis of executions of consensus algorithms, we introduce the notion of valency of a configuration. A configuration C is 0-valent if all executions that contain C decide 0, and 1-valent if all execution that contain C decide 1. In this case, we say that C is univalent. A configuration that is not univalent is bivalent.

Our goal is to construct, for every algorithm, an admissible execution in which processes can't decide. For that, we construct an execution whose configurations are all bivalent. We start with the initial configuration:

Lemma 13.2. There is a bivalent initial configuration.

Proof. Without loss of generality, assume $V = \{1, 2, \ldots, n\}$. We consider the initial configurations I_k in which all processes $p \leq k$ have initial value 1 and all processes $p > k$ have initial value 0. By validity, I_0 is 0-valent and I_n is 1-valent.

Assume by contradiction that all initial configurations are either 0-valent or 1-valent. Then there must be some k such that I_{k-1} is 0-valent and I_k is 1-valent. Take any admissible infinite schedule σ that is applicable to I_{k-1} in which process k doesn't take any steps. Then, since I_{k-1} is 0-valent, the decision value in execution $\sigma(I_{k-1})$ is 0.

Now, since the only difference between I_{k-1} and I_k is the local state of process k and process k doesn't participate in schedule σ, the same schedule is also applicable to I_k. Furthermore, the only difference in the executions $\sigma(I_{k-1})$ and $\sigma(I_k)$ is the local state of process k. Hence, the decision values must be the same in both executions. But then $\sigma(I_k)$ decides 0, which is in contradiction to I_k being 1-valent. □

For the inductive construction of the bivalent execution, we need the following technical lemma:

Lemma 13.3. Let C be a bivalent configuration and let $e = (p, m)$ be an event applicable to C. If $\mathcal{C}$ is the set of configurations reachable from C without applying e and $\mathcal{D} = \{e(E) \mid E \in \mathcal{C} \text{ and } e \text{ is applicable to } E\}$, then $\mathcal{D}$ contains a bivalent configuration.

Proof. Assume by contradiction that there are no bivalent configurations in $\mathcal{D}$.

The event e is applicable to all configurations in $\mathcal{C}$. This is obvious if $m = \bot$. If $m \neq \bot$, then the message m is still in transit to process p in configuration $E \in \mathcal{C}$, and can thus be received by process p.

We next argue that there is a 0-valent and a 1-valent configuration in $\mathcal{D}$. Let E_v be a v-valent configuration reachable from C for $v \in \{0, 1\}$. Both E_0 and E_1 exist since C is bivalent. If $E_v \in \mathcal{C}$, then set $F_v = e(E_v) \in \mathcal{D}$. If $E_v \notin \mathcal{C}$, then E_v has a predecessor configuration for which e was applied, so $F_v \in \mathcal{D}$.

Now, call two configuration neighbors if one is a direct successor of the other. Since every $D \in \mathcal{D}$ is of the form $D = e(C)$ with $C \in \mathcal{C}$, $\mathcal{C}$ is connected with respect to the neighbor relation, and C is bivalent, there are two neighbors $C_0, C_1 \in \mathcal{C}$ such that $D_0 = e(C_0)$ is 0-valent and $D_1 = e(C_1)$ is 1-valent. Without loss of generality, let $C_1 = e'(C_0)$ where $e' = (p', m')$. We distinguish two cases:

1. If $p' \neq p$, then $D_1 = e(D_0)$ by Lemma 13.1. This is a contradiction to the fact that D_0 and D_1 have opposite valencies.
2. If $p' = p$, then let σ be any finite schedule in which p doesn't take any steps and for which all non-p processes have decided in configuration $A = \sigma(C_0)$. By Lemma 13.1, the schedule σ is also applicable to the configurations C_1, D_0, and D_1. Moreover, $e(A) = e(\sigma(C_0)) = \sigma(e(C_0)) = \sigma(D_0)$ and $e(e'(A)) = e(e'(\sigma(C_0))) = \sigma(e(e'(C_0))) = \sigma(e(C_1)) = \sigma(D_1)$. Since D_0 and D_1 have opposite valencies, the configuration A is a bivalent, a contradiction to the fact that some process has already decided in A. □

We can now conclude:

Theorem 13.2. There is no consensus algorithm in asynchronous message passing with $f \geq 1$ crash faults.

Proof. Assume by contradiction that there is a consensus algorithm.

By Lemma 13.2, there is a bivalent initial configuration C_0. From this initial configuration, we construct an execution in phase. The configuration at the end of each phase will be bivalent by construction. We then show that the constructed execution is admissible, deriving the desired contradiction.

For constructing the phases, we maintain a queue of processes. Initially, all processes are in the queue in arbitrary order. A phase ends when the process at the head of the queue receives its oldest message that was in transit at the start of the phase. If no such message exists, the phase ends when it takes its first step in the phase. The process is then moved to the end of the queue. Let $e = (p, m)$ be the event that ends the current phase. By Lemma 13.3, we can end the current phase in a bivalent configuration.

Since every phase contains at least one step, this leads to an infinite execution of the algorithm. By construction, no process can ever decide, since there are bivalent configurations arbitrarily late in the execution. Furthermore, every message sent is eventually received by construction of the phases. This is a contradiction to the fact that the algorithm decides in every admissible execution. □

13.5 Asynchronous Rounds

The proof of Theorem 13.2 is kind of tricky in that we needed to make sure that the constructed execution was admissible. The hard part about that is the fact that validity of a liveness condition is not a local problem of the configurations; we can only decide it for infinite executions. In general, reasoning about liveness properties is hard, which is why there has been a push towards models that capture important aspects, like asynchrony, but which do not have liveness conditions, only safety conditions.

For asynchronous message passing with crash faults, this often comes in the form of asynchronous rounds. These are rounds constructed in an asynchronous system. Every process sends its round-r message to all other processes and proceeds to the next round $r + 1$ when it has received $n - f$ messages (including its own). Waiting for more than $n - f$ messages might block the process, since there could be f process that do not send a round-r message; so this choice of parameter is optimal.

It is not clear why one would want to construct asynchronous rounds; whether they are useful. If one looks at algorithms in asynchronous models, however, it is

exceedingly rare to find one that does not proceed in rounds.

With the above construction like this, we get a Heard-Of set $\mathrm{HO}_p(r)$ for every round r and every process p. They satisfy the following two properties:

1. $p \in \mathrm{HO}_p(r)$
2. $|\mathrm{HO}_p(r)| \geq n - f$

Note, however, that no stability from one round to next exists: processes whose messages are slow in one round can be fast in the next. The Heard-Of sets of a given round describe a Heard-Of graph HO, which is the directed graph on V such that $\mathrm{In}_p = \mathrm{HO}_p$.

Since the Heard-Of model can be implemented in asynchronous message passing, the impossibility result of Theorem 13.2 also implies impossibility in the above Heard-Of model. There is, however, a much easier direct proof of this fact:

Theorem 13.3. There is no consensus algorithm in the f-receive-omission Heard-Of model if $f \geq 1$.

Proof. The existence of a bivalent initial configuration follows just like in the proof of Lemma 13.2, by silencing one process.

Let us write $\mathrm{HO} \sim \mathrm{HO}'$ if they only differ in the Heard-Of set of a single process. In this case, it is impossible to have the resulting configurations be of opposite valencies: If $\mathrm{HO}(C)$ is 0-valent and $\mathrm{HO}'(C)$ is 1-valent, then we can choose a schedule σ in which the single process that has a possibly different local state in $\mathrm{HO}(C)$ and $\mathrm{HO}'(C)$ is forever silenced. So all other processes decide the same value in both executions, so both 0 and 1, a contradiction.

Since the set of Heard-Of graphs is connected with respect to the relation $\sim$, it is impossible to have a bivalent configurations whose successors are all univalent. There hence is an infinite executions in which all configurations are bivalent, a contradiction. □

May 19, 2025

Lecture 14: Distributed Algorithms III

We have seen that consensus is impossible in asynchronous systems with process faults. We now turn to the question of where to go from here. Is it just too hard to achieve consensus? Is the model too unrealistic to capture practical systems? There is, of course a non-trivial modeling question that has no definitive answer as of yet. But it also turns out that consensus *can* be achieved if we relax the problem definition just a bit. In what follows, we assume an upper bound of $f < n/2$ on the number of crashed processes in each execution. If $f \geq n/2$ crashes are possible, then the system can be disconnected and many reasonable forms of consensus become impossible.

14.1 Approximate Consensus

We first study what happens if we relax the Agreement condition by just a bit:

- For any two decision values y_p and y_q, we have $|y_p - y_q| \leq \varepsilon$. ($\varepsilon$-Agreement)

It turns out that not only this approximate consensus is solvable in asynchronous systems, it is solvable with a very easy class of algorithms: An averaging algorithm is one that updates its value to some average of the $n - f$ received values in each round. This guarantees that the new value is inside the convex hull of the set of received values.

Let $\alpha \in [0, 1]$. We call an averaging algorithm α-safe if the new value doesn't go too close to the boundary of the convex hull of received values: the new value y_p satisfies $m + \alpha\Delta \leq y_p \leq M - \alpha\Delta$ where m and M are the minimal and maximal received values, and $\Delta = M - m$. Many averaging algorithms are α-safe:

Lemma 14.1. Every averaging algorithm with weights $\geq \alpha$ is α-safe.

Proof. Let $x_1, x_2, \ldots, x_k$ be the values received by process p in round r. Assume that these values are ordered non-decreasingly, *i.e.*, $x_1 = m$ is the minimal and $x_k = M$ is the maximal value. Then:

$$y_p = \sum_{i=1}^{k} a_i \cdot x_i \leq \alpha m + (1 - \alpha)M = M - \alpha\Delta$$

We show $y_p \geq m + \alpha\Delta$ in the same way. □

Denote by Δ_r the maximum distance between values at the end of round r. Initially we have $\Delta_0 \leq 1$ and we achieved ε-Agreement when $\Delta_r \leq \varepsilon$. It would be convenient if we had a contraction ratio $\beta < 1$ by which Δ_r shrinks from round to round: $\Delta_r \leq \beta\Delta_{r-1}$ for all rounds $r \geq 1$. This is not too much to hope for, as we will see in the next lemma. It uses a fundamental fact about asynchronous rounds with $f < n/2$, the non-split property: for any pair of non-crashed processes and any round, there exists one process that both hear from in that round. The common process that both hear from can change from round to round and it need not be the same for all pairs of processes.

Lemma 14.2. Every α-safe averaging algorithm has a contraction ratio of $1 - \alpha$.

Proof. Denote by $I_p = [m_p, M_p]$ the interval spanned by the values received by process p in the current round r. Let p and q be two processes and y_p and y_q their values at the end of the round. Without loss of generality let $m_p \leq m_q$. Then, because of the non-split property, we have $m_q \leq M_p$.

For process p, we calculate

$$y_p \geq m_p + \alpha(M_p - m_p) \geq (1-\alpha)m_p + \alpha m_q \geq (1-\alpha)m + \alpha m_q$$

and

$$y_q \leq M_q - \alpha(M_q - m_q) \leq (1-\alpha)M_q + \alpha m_q \leq (1-\alpha)M + \alpha m_q \ ,$$

where m and M are the minimal and maximal values of non-crashed processes at the start of the round, from which follows that $y_q - y_p \leq (1-\alpha)(M-m) = (1-\alpha)\Delta_{r-1}$. Similarly, for process q, we get

$$y_p \leq M_p - \alpha(M_p - m_p) \leq (1-\alpha)M_p + \alpha m_p \leq (1-\alpha)M + \alpha m_q$$

and

$$y_q \geq m_q + \alpha(M_q - m_q) \geq (1-\alpha)m_q + \alpha M_q \geq (1-\alpha)m + \alpha m_q \ ,$$

from which follows that $y_p - y_q \leq (1-\alpha)(M-m) = (1-\alpha)\Delta_{r-1}$. We thus have $|y_p - y_p| \leq (1-\alpha)\Delta_{r-1}$ for all p and q, which implies that $\Delta_r \leq (1-\alpha)\Delta_{r-1}$ and concludes the proof. □

We can now conclude that all α-safe averaging algorithms solve approximate consensus:

Theorem 14.1. Every α-safe algorithm achieves ε-agreement in $\left\lceil \log_{1/(1-\alpha)} \frac{\Delta}{\varepsilon} \right\rceil$ rounds.

Proof. By Lemma 14.2, we have:

$$\Delta_r \leq (1-\alpha)^r \Delta$$

Thus, whenever $r \geq \log_{1/(1-\alpha)} \frac{\Delta}{\varepsilon}$, we have

$$\Delta_r \leq e^{\log_{1/(1-\alpha)} \frac{\Delta}{\varepsilon} \cdot \log(1-\alpha)} \Delta = e^{-\log \frac{\Delta}{\varepsilon}} \Delta = \varepsilon \ ,$$

which concludes the proof. □

We talked about classes of averaging algorithms, which all solve approximate consensus, but which specific one should we choose? A natural choice is the algorithm that chooses the unweighted average of all received values. By Lemma 14.1, this algorithm is $(1 - 1/n)$-safe. Another choice, which turns out to be time-optimal in the worst case, is the algorithm that maximizes the safety parameter α: The Midpoint algorithm chooses the midpoint of the convex hull of received values and has the maximum possible safety parameter $\alpha = 1/2$.

14.2 Randomized Consensus

Allowing randomization also allows to circumvent consensus impossibility. Multiple randomized relaxations of consensus are possible. The most useful one still requires Agreement and Validity to hold for all executions, but requires only almost sure Termination, *i.e.*, with probability 1.

The BenOr algorithm is one of the simplest randomized consensus algorithms. It proceeds in phases of two rounds each:

1. Send v_p to all other processes. Then, if all received values are equal to v, then set $w_p = v$, otherwise set $w_p = \bot$.
2. Send w_p to all other processes. Then, if all received values are equal to $v \in \{0,1\}$, then set $v_p = v$ and decide v. Otherwise, if one received value is $v \neq \bot$, then set $v_p = v$. Otherwise, set v_p to a random value in $\{0,1\}$.

One of the key insights into why this algorithm works is the fact that no two different values can remain after the first round:

Lemma 14.3. In any phase, if $w_p \neq \bot$ and $w_q \neq \bot$, then $w_p = w_q$.

Proof. Assume $w_p \neq w_q$. Then there exist sets P and Q of processes with $|P| \geq n - f$ and $|Q| \geq n - f$ with $v_p = w_p$ for all $p \in P$ and $v_q = w_q$ for all $q \in Q$ at the start of the phase. But this is impossible because then the total number of processes is at least $|P \cup Q| = |P| + |Q| \geq 2(n - f) > 2\frac{n}{2} = n$, a contradiction. □

The rest of the correctness proof is now in reach:

Theorem 14.2. The BenOr algorithm solves randomized consensus in expected time $O(2^n)$.

Proof. We show the three consensus properties: Agreement, Validity, and almost sure Termination.

We start with Validity. If $v_p = v$ for all non-crashed processes at the start of the first phase, then $w_p = v$ at the end of the first round and thus all non-crashed processes decide v at the end of the second round.

To show Agreement, assume that process p decides value v in phase ϕ, and assume that ϕ is the first phase in which any process decides. No other processes can decide a different value in phase ϕ because of Lemma 14.3. Furthermore, in the second round of phase ϕ, every non-crashed process q receives at least $n - 2f \geq 1$ times the value $w_p = v$. But this means that $v_q = v$ at the end of phase ϕ. All non-crashed processes hence start phase $\phi + 1$ with the same value $v_q = v$, which leads them to decide v in the second round of phase $\phi + 1$ by Validity and the memoryless nature of the phases.

We now show almost sure Termination and the $O(2^n)$ upper bound on the expected number of phases until decision by all correct processes. At the end of any phase in which there is one non-crashed processes that has not yet decided, by Lemma 14.3, all non-crashed processes start the next phase with the same value v_p if all random choices are equal to the unique non-$\bot$ value w_p. The probability of this happening is at least $1 - 1/2^n$. The number of phases until

decision by all correct processes is thus upper-bounded by a geometric random variable with success probability $1 - 1/2^n$. This concludes the proof. □

14.3 Byzantine Processes

The case of process crashes is challenging, but one can and does consider worse than crash faults. A Byzantine process can behave arbitrarily. This includes crashing, but also sending wrong messages, sending different wrong messages to different processes, following the algorithm initially and deviating later, etc. It turns out that both approximate consensus and randomized consensus can still be solved. This of course requires adapting the Agreement and Validity conditions to only cover correct, non-Byzantine, processes. It also requires us to lower the fault-tolerance threshold to $f < n/3$. Many problems are known to be unsolvable when $f \geq n/3$.

We need to employ some tricks to get there, of course. These come in the form of reliable broadcast and the witness technique, which combine into the following guarantees in every asynchronous round: every correct process receives at least $n - f$ messages from other processes and $n - f$ of them are common to *all* correct processes. Note that receiving more than $n - f$ messages can actually be a disadvantage here, because they can be from Byzantine processes.

For approximate consensus, we will need to throw away the f smallest and the f largest received values. Otherwise, a Byzantine process can make us exit the convex hull of values of correct processes. The only critical point is the non-split property of the resulting intervals. But, after trimming, there are at least $n - f - 2f = n - 3f \geq 1$ values in common at all processes. The intervals thus intersect.

For randomized consensus, we will need to introduce some thresholds to guard the rules for setting w_p and v_p. For setting $w_p = v$ in the first round of a phase, we place the condition that all but f received values are equal to v. For deciding in the second round of a phase, a value needs to make up all but f of the received values. For adopting a value, it needs to have been received at least $f + 1$ times.

The analog of Lemma 14.3 is true: Writing V for the number of times that value v was received, we have $V = V_c + V_r$ where V_c is the number of times that v appears in the set of $k \geq n - f$ common values received by all correct processes. We have $V_r \leq m - k$ where m is the total number of received values. But then $V_c = V - V_r \geq (m - f) - (m - k) = k - f = k/2 + (k - 2f)/2 \geq k/2 + (n - 3f)/2 > k/2$, which shows that there can be only one such value v.

A value that has been decided is adopted by all other correct processes in the same phase: For the deciding process, we have shown $V_c \geq k - f$ in the above calculation. Now, for the number V' of times that any other correct process has received value v, we have $V' \geq V_c \geq k - f \geq n - 2f > f$. That process hence adopts value v in the same phase.