Lecture outline: strings and permutations

We introduce the formal definitions of strings and permutations and various important related concepts such as prefix, suffix, substring, ...

We review the basic counting results that should already be known and that involve three kinds of numbers: exponential numbers, factorial numbers and binomial numbers (and the binomial theorem).

We see how strings can be used to model/encode other kinds of combinatorial objects such as lattice paths or integer compositions.

Strings

Definition. A **string** S, of size n, over an alphabet $\mathcal{A}$ is a linearly ordered list of n elements (**letters**) taken from $\mathcal{A}$, possibly with repetition.

Notation. For a string S of length n, we denote by $S[1]$ its first letter, $S[2]$ its second letter, ..., $S[n]$ its last letter.

Remark. Strings are sometimes called sequences or words. We will always use strings. The size of a string is also called its **length**.

Remark. The important property of strings is the total/linear order on the atoms (symbols from the alphabet): in a string, there is a first symbol, a second symbol, ..., a last symbol.

Definition. Let S be a string of size n.

For any $1 \leq i \leq j \leq n$, the string $S[i]\ S[i+1] \ldots S[j-1]\ S[j]$ is called a **substring** of S. It is denoted by $S[i, j]$

If $i = 1$, then $S[1, j]$ is a **prefix** of S.

If $j = n$, then $S[i, n]$ is a **suffix** of S.

A substring that is neither a prefix nor a suffix is a **proper substring**.

Example. $S = \underbrace{\overset{1}{0}\ \overset{2}{1}\ \overset{3}{1}\ \overset{4}{0}}_{\text{prefix}}\ \overset{5}{0}\ \underbrace{\overset{6}{0}\ \overset{7}{1}\ \overset{8}{0}\ \overset{9}{1}}\ 0\ \underbrace{1\ 1\ 0}_{\uparrow\ \text{suffix}}$

$S[6,9] = 0101$

$S[3] = 1$

We now review the first, basic counting result.

Theorem (review). If $\mathcal{A}$ is an alphabet of k symbols, there are

$$k^n$$

strings of size n over $\mathcal{A}$, for $n \geq 0$.

Notation. The set of all strings over $\mathcal{A}$ is often denoted by $\mathcal{A}^*$. The set of all strings of size n is denoted by $\mathcal{A}^*_n$.

Remark. The result above implies there is one string of size 0. This is the **empty string**, that contains no symbol, and is often denoted by ϵ.

Proof. If we want to generate a arbitrary ~~random~~ string S of length n over $\mathcal{A}$, we need to chose one symbol for each of the n positions of S.

There are k choices for each $S[i]$, $i \in [1, n]$ as the choice made for a given position does not influence the choices that can be made for the other positions.

Examples. $\mathcal{A} = \{0, 1\}$ (binary strings)

0110111110101

$\mathcal{A} = \{a, b\}$ (binary strings again, the alphabet is of size 2)

ababbba

$\mathcal{A} = \{A, C, G, T\}$ (DNA strings)

$\mathcal{A} = \{0, 1, 2, \ldots, 9\}$ (how many PIN)

Factorial numbers, Binomial numbers and the Binomial Theorem

Factorial numbers (review). For $n \geq 1$,

$$n! = n(n-1)(n-2)\cdots 2 \cdot 1$$

For $n = 0$, we define $0! = 1$

Remark. $n!$ is the number of ways to arrange n pair-wise distinct elements (permutations as we will soon see).

Example.

$A = \{a, b, c\}$

perm of A :

$$\left.\begin{array}{l} abc \;,\; acb \\ bac \;,\; bca \\ cab \;,\; cba \end{array}\right\} 6 = 3!$$

Binomial numbers (review). For $n \geq 0,\ n \geq k \geq 0$,

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

If $k < 0$ or $k > n$,

$$\binom{n}{k} = 0$$

Remark. $\binom{n}{k}$ is the number of ways to select a subsets of k elements out of a set of n pair-wisely distinct elements.

Remark. It should be obvious, both from the formula and from the combinatorial interpretation that

$$\binom{n}{k} = \binom{n}{n-k}$$

Example.

$A : \{a, b, c, d, e\}$

choose subset of size 2

ab, ac, ad, ae
bc, bd, be
cd, ce
de

$10 = \binom{5}{2}$

We use the binomial numbers to count strings with some prescribed properties.

Theorem. Let $\mathcal{A}$ and $\mathcal{B}$ be two alphabets with no common letters, of respective sizes a and b.
Let $\mathcal{C} = \mathcal{A} \cup \mathcal{B}$ be the alphabet obtained by the union of $\mathcal{A}$ and $\mathcal{B}$.
The number of strings of size n over $\mathcal{C}$ with exactly k symbols from $\mathcal{A}$ and $n-k$ symbols form $\mathcal{B}$ is given by

$$\binom{n}{k} a^k b^{n-k}$$

Proof. To construct a string of size n over $\mathcal{C}$, one needs to

- chose which positions will receive symbols from $\mathcal{A}$, and there are $\binom{n}{k}$ choices,
- for each of these k chosen positions we need to choose one of the k symbols of $\mathcal{A}$,
- for each of the remaining $n-k$ chosen positions we need to choose one of the ℓ symbols of $\mathcal{B}$.

Example. $\mathcal{A} = \{0, 1\}$, $\mathcal{B} = \{a, b, c\}$, $n = 4$ and $k = 3$

$C = \{0, 1, a, b, c\}$

want strings of size 4 with 3 elts from A
1 elt from B

thm says: the number of strings of this type is $\binom{4}{3} \cdot 2^3 \cdot 3^1 = 96$

8 ways to fill in

a _ _ _	_ a _ _	_ _ a _	_ _ _ a
b _ _ _	_ b _ _	_ _ b _	_ _ _ b
c _ _ _	_ c _ _	_ _ c _	_ _ _ c

Using the previous counting result, we can understand (almost prove) combinatorially a classical result you probably know and that will be useful later.

Binomial Theorem. If x and y are two variables and n is a positive integer:

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

Proof idea. If x and y are assumed to be positive integers, then $(x+y)^n$ is the number of strings of size n over an alphabet $\mathcal{C}$ obtained by the union of two completely distinct alphabets $\mathcal{A}$ and $\mathcal{B}$ of respective sizes x and y.

For any such string, denote by k its number of symbols from $\mathcal{A}$ (so it has $n-k$ symbols from $\mathcal{B}$). k ranges from 0 (no symbol from $\mathcal{A}$) to n (no symbol from $\mathcal{B}$).

In other words, combinatorially, this identity just states the obvious fact that the set of all strings of size n over $\mathcal{C}$ is the union, for $k=0$ to n, of the sets of all strings of size n over $\mathcal{C}$ that contains exactly k symbols from $\mathcal{A}$.

This proves, combinatorially, the result in the case of x and y being integers, which will be enough for us here. The general proof is available on pages 21-22 of your textbook (**please, read it**).

$$(x+y)^n = \underbrace{(x+y)(x+y)(x+y)\ldots(x+y)}_{n}$$

$$= \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

Correspondence with sequences over $\{x,y\}$ of length n

Example. For binary strings, this gives

$$2^n = (1+1)^n \cancel{2^n} = \sum_{k=0}^{n} \binom{n}{k} = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n}$$

2^0 1

2^1 1 1

2^2 1 2 1 $\binom{2}{0}$ $\binom{2}{1}$ $\binom{2}{2}$

2^3 1 3 3 1

2^4 1 4 6 4 1

1 5 10 10 5 1

Example. What is $\binom{n}{0} - \binom{n}{1} + \binom{n}{2} + \cdots + (-1)^n\binom{n}{n} = \sum_{k=0}^{n}(-1)^k\binom{n}{k}$?

$0 = (1-1)^n$ //

Permutations

Definition. A permutation P over an alphabet $\mathcal{A}$ is a string over $\mathcal{A}$ where every symbol of $\mathcal{A}$ occurs exactly once.
Equivalently, it is a total/linear order over the symbols of $\mathcal{A}$.

Theorem. The number of permutations over an alphabet $\mathcal{A}$ of size n is $n!$

Proof. We follow the same proof principle than for counting sequences, counting the number of symbols we can chose for any position when generating a random permutation.

We have n choices for the symbol in first position. But now this symbol can not be used again in another position. So we have $n-1$ choices for the symbol in second position. And so on, for the last position, we have only one symbol left we can chose. This gives

$$n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot 1$$

possible permutations over $\mathcal{A}$.

Multinomial numbers and strings with prescribed content

The **content** of a string S over an alphabet $\mathcal{A}$ is the information about how many times each symbol of $\mathcal{A}$ appears in the string S. Formally, this is a function $c : \mathcal{A} \to \{0, 1, 2, 3, \dots\}$.

However, it is usually simpler to represent it as follows. Without loss of generality, one can assume that $\mathcal{A} = \{1, 2, \dots, k\}$ or that its symbols are totally ordered.

If $\mathcal{A}$ has k symbols, the content of S is a sequence of k non-negative integers, in general denoted by $(n_1, n_2, \dots, n_k)$, where n_i is the number of occurrences of the symbol i (equivalently or the i^{th} symbol of $\mathcal{A}$) in S.

It then becomes natural to ask the following question: how many strings over $\mathcal{A}$ have content $(n_1, n_2, \dots, n_k)$?

There are two cases we already saw. First, permutations are strings with content $(1, 1, \dots, 1)$ (so the answer is $k!$). Second, from slide 6, we can also answer this question if $k = 2$ (binary strings).

Ex of content $A = \{a, b, c, d\}$

cabbdacbddaa

$\text{Content} = (\underset{\#a}{4}, \underset{\#b}{3}, \underset{\#c}{2}, \underset{\#d}{3})$

letter	#
a	4
b	3
c	2
d	3

content

(a function from $\{a,b,c,d\} \to \mathbb{Z}^+$

For $k > 2$ we need to generalize factorial and binomial numbers and introduce **multinomial numbers**.

Definition. Let $(n_1, n_2, \dots, n_k)$ be a sequence of k non-negative numbers. The number of strings over an alphabet $\mathcal{A}$ of size k with content $(n_1, n_2, \dots, n_k)$ and size $n = n_1 + n_2 + \cdots + n_k$ is

$$\binom{n}{n_1, n_2, \dots, n_k} = \frac{n!}{n_1!\, n_2! \cdots n_k!}$$

note $\binom{n}{k} = \binom{n}{k, n-k}$

Proof. The principle is simple, we only need to express the multinomial number as a product of binomial numbers:

$$\frac{n!}{n_1!n_2!\cdots n_k!} = \binom{n}{n_1}\binom{n-n_1}{n_2,\dots,n_k}$$

$$\frac{n!}{n_1!n_2!\cdots n_k!} = \boxed{\binom{n}{n_1} \times \binom{n-n_1}{n_2} \times \cdots \times \binom{n-n_1-n_2-\cdots-n_{k-1}}{n_k}}$$

= # strings over alphabet size k with content $(n_1, n_2, \dots n_k)$ and size n

possible choices for position of 1st letter

for 2nd letter

This just means that to construct a string with content $(n_1, n_2, \dots, n_k)$ we need to choose the n_1 positions that will receive the first symbol of $\mathcal{A}$ (assume that $\mathcal{A} = \{1, 2, \dots, k\}$), then, over the $n - n_1$ remaining positions, we need to construct a string over the smaller alphabet $\mathcal{A} - \{1\}$ and with content $(n_2, \dots, n_k)$, which we do by picking n_2 positions for the second symbol of $\mathcal{A}$, and repeating this process until we are done.

Exercises

We have almost all we need to count strings and permutations. Below are a few examples.

Permutations over a subalphabet. Assume we have an alphabet of n letters, and we want to count the number of strings of size k over this alphabet where no symbol occurs twice or more. These are thus permutations over a subset of size k of the alphabet. Prove that this number is

$$P(n,k) = \frac{n!}{(n-k)!}$$

# of ways to choose a k-elt subset of n-elt set

$$\binom{n}{k} k! = \frac{n!}{(n-k)!\,k!} \cdot k! = \frac{n!}{(n-k)!}$$

perm of set of size k

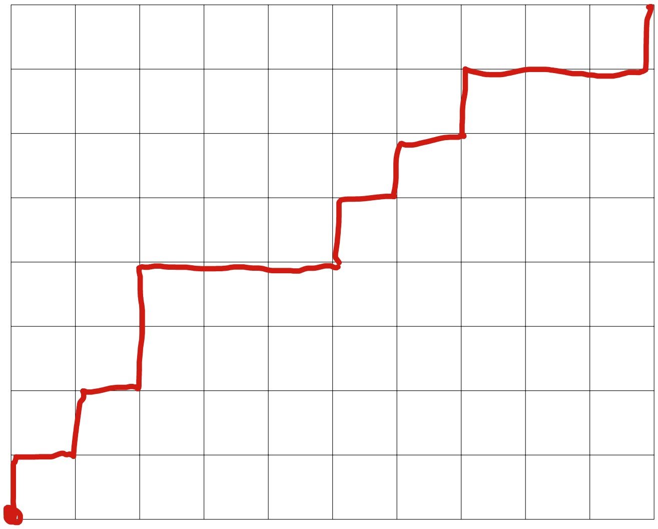
Fixed content prefix. Let $n \geq m \geq k$ be three non-negative integers. What is the number of binary strings of length n over the alphabet $\mathcal{A} = \{0, 1\}$ whose prefix of size m contains exactly k 0s ?

$$= \binom{m}{k} 2^{n-m}$$

ways of choosing elts after prefix

ways of choosing prefix of size m content k

Lattice paths. Many models in theoretical physics involve lattice paths, that are paths in the square lattice with a prescribed set of allowed steps. Here we consider paths with North and East steps.

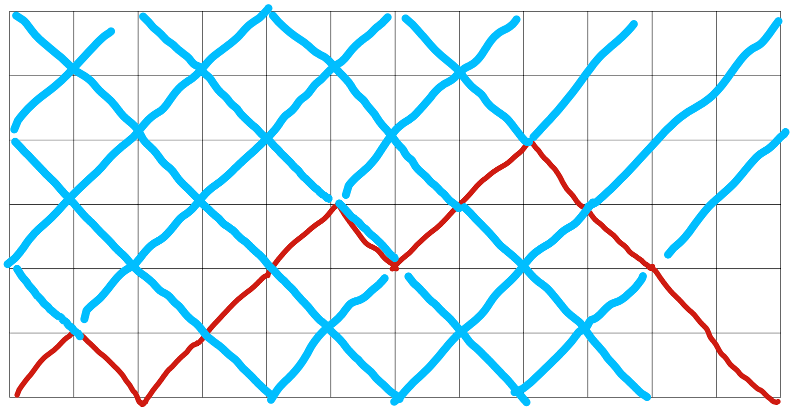


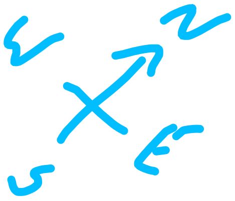
NENENNEEENENENEEEN

We want to know how many lattice paths there are from the point of coordinates $(0, 0)$ to the point of coordinates (a, b), where a and b are two non-negative integers.

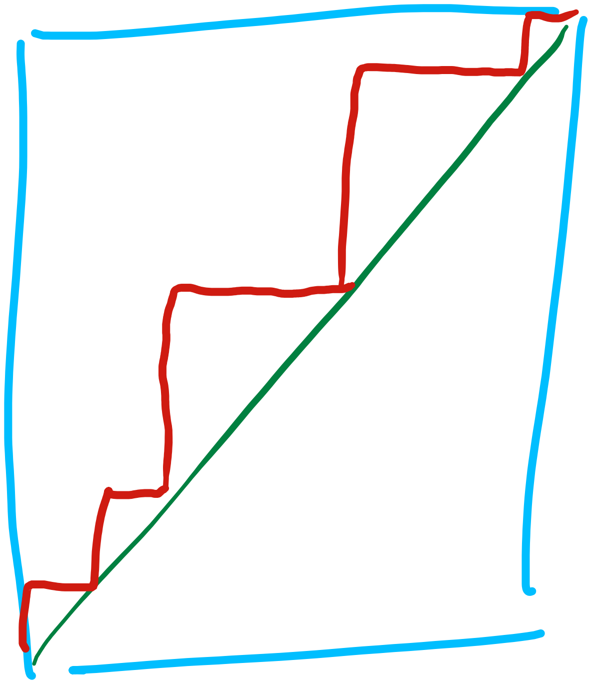
$$= \#\text{ strings with } \begin{matrix} a \text{ E's} \\ b \text{ N's} \end{matrix} = \binom{a+b}{a}$$

Lattice paths again. We now consider lattice paths with North-East and South-East steps.





We want to know how many lattice paths there are from the point of coordinates $(0, 0)$ to the point of coordinates $(0, b)$, where b is a non-negative integer.



want to count lattice paths N, E
start at (0,0)
end at (n,n)
stay above diagonal

$$= \frac{1}{n+1}\binom{2n}{n}$$

see Section 1.5

lattice path argument
for Catalan numbers

Looking for significance of a string property. Suppose you are a biologist, you sequence a (very short) gene whose DNA sequence has length 21, and you observe that every time the nucleotide (symbol) A appears, it is followed by the nucleotide C. An immediate question is to wonder if that is something we can expect from a random sequence or if this might be exceptional enough that the biological function of this observation is investigated. To answer this we need to find the number of strings of size 21 over the alphabet $\{A, C, G, T\}$ in which an A is always followed by a C.

Counting with repetitions (Section 1.4). We want to prove, using strings, that the number of non-negative integer solutions of the equation $x_1+x_2+\cdots+x_k = n$ is

$n=3$
$k=1$

$\binom{3}{1}=3$ ✗

$\binom{3}{0}=1$ ✓

$$\binom{n+k-1}{k-1} = \binom{n+k-1}{n}$$

$x_1 + x_2 + x_3 + x_4 = 10$ $\quad x_1 .. x_4 \geq 0$

$5 + 2 + 1 + 2 = 10$

$$\underbrace{11111}_{5}0\underbrace{11}_{2}0\underbrace{1}_{1}0\underbrace{11}_{2}$$

binary strings

10 1's

3 0's

$\#sol = \binom{10+3}{3}$

in genl bin str

n 1's

k-1 0's