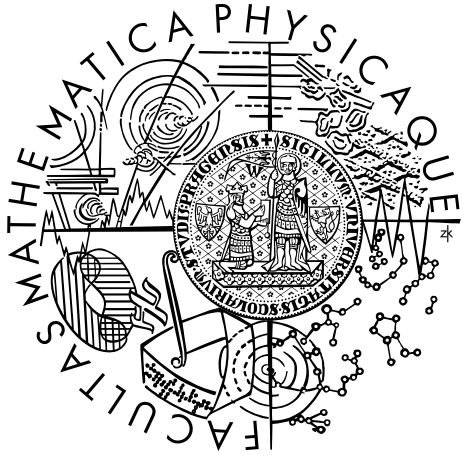
Charles University in Prague
Faculty of Mathematics and Physics

BACHELOR THESIS



Stanislav Kučera

Generating random pattern-avoiding matrices

Computer Science Institute of Charles University

Supervisor of the bachelor thesis: RNDr. Vít Jelínek, Ph.D.
Study programme: Computer Science
Study branch: General Computer Science

Prague 2016

I declare that I carried out this bachelor thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ..... signature of the author

Title: Generating random pattern-avoiding matrices

Author: Stanislav Kučera

Department: Computer Science Institute of Charles University

Supervisor: RNDr. Vít Jelínek, Ph.D., department

Abstract: Abstract.

Keywords: binary matrix pattern-avoiding Markov chain Monte Carlo

Dedication.

Contents

Preface

Theses at the faculty of mathematics and physics usually fit into one of three categories:

1. Theoretical thesis
2. Experimental thesis
3. Implementation thesis

My thesis does not fit entirely into only one category and it does not try to. The project consists of several similarly important parts which are:

- Design of the algorithm for generating a special binary matrix
- Making it run fast on inputs which are usual for researchers
- Implementing the algorithm to provide practical tool

None of these points would make sense alone but together the thesis may become very useful for scientists as it is a common practice to test hypothesis on random data.

Introduction

The area of pattern avoidance is heavily studied for permutations and it also becomes studied more for its generalization - binary matrices.

We denote by $M \in \{0,1\}^{n \times m}$ a binary matrix of size n by m, calling n the number of rows of M - the height of the matrix M and m the number of columns - its width. A line of a matrix is one of its rows or columns and for matrix M, we denote $L(M)$ the ordered set of all lines of M. Its order is given by the natural indexing of rows and columns.

We say a binary matrix M contains a binary matrix P, which we call a "pattern", as a submatrix, if there is a mapping $f : L(P) \to L(M)$, such that

- $l \in L(P)$ is a row of P iff $f(l) \in L(M)$ is a row of M
- $\forall l, l' \in L(P) : l < l' \Rightarrow f(l) < f(l')$ (preserves the order)
- $\forall l, l' \in L(P)$: if lines l and l' intersect and there is a one-entry at the intersection, then there is a one-entry at the intersection of $f(l)$ and $f(l')$.

otherwise, it avoids the pattern P.

$$\begin{array}{c} \\ 0 \\ 1 \end{array}\!\!\begin{array}{c} \begin{array}{cc} 2 & 3 \end{array} \\ \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \end{array} \qquad \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array}\!\!\begin{array}{c} \begin{array}{ccc} 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \end{array} \qquad \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array}\!\!\begin{array}{c} \begin{array}{ccc} 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \end{array}$$

Let the pattern P be the left matrix from the picture. The middle matrix contains the pattern P, because a mapping $\{(0,0),(1,2),(2,3),(3,4)\}$ satisfies all the conditions. On the other hand the right matrix avoids the pattern as there is no such mapping.

There are other possible definitions of containing but from now on we will always consider a pattern to be contained as a submatrix. We will always denote M the binary matrix for which we test the containing and by P the pattern that is being tested. Moreover, we denote h the height (the number of rows) of P and w its width.

Generating random matrix

Our main goal is to show an iterated algorithm, which, with high enough number of iterations, for a given forbidden pattern and number n generates a uniformly random binary matrix $M \in \{0,1\}^{n \times n}$ avoiding the pattern.

To create such an algorithm we generalize an algorithm generating fixed size permutations without a forbidden pattern introduced by [Madras-Liu]. To achieve randomness it uses Markov chain Monte Carlo method, which is explained in the next section.

The process starts with a binary matrix of the correct size avoiding the pattern (zero matrix is always sufficient) and in each step it changes uniformly at random one bit of the matrix and tests whether the matrix still avoids the pattern after the change. If it does the step is over, if it does not, we revert the change (flip the bit back) and end the step.

It can be proven that if the Markov chain is chosen carefully, this leads to a uniformly random matrix avoiding the pattern, after a sufficient number of steps. Sadly no-one knows what is "sufficient" in this manner of speaking.

Testing avoidance

The hardest task to do in a step is to test whether the new matrix avoids the pattern. We show a very fast algorithm which only works for a special class of binary matrices as well as a slightly less performing algorithm for a general pattern, which again comes as a generalization of an algorithm for permutations from the article by Madras and Liu.

1. Markov chain Monte Carlo

We shall consider discrete-time Markov chains $X_0, X_l, X_2, \ldots,$ where each X_i takes values in a finite or countably infinite state space S. For $k = 0, 1, 2, \ldots,$ the k-step transition probabilities are:

$$p_{i,j}^{(k)} = Pr(X_{t+k} = j | X_t = i) \qquad (i, j \in S)$$

A Markov chain is **irreducible** if the chain can eventually get from each state to every other state, that is, for every $i, j \in S$ there exists a $k \geq 0$ (depending on i and j) such that $p_{i,j}^{(k)} > 0$.

An irreducible chain has period D if D is the greatest common divisor of $\{k \geq 1 | p_{i,j}^{(k)} > 0\}$ for some $i \in S$ (equivalently, for all $i \in S$). A chain is called **aperiodic** if its period is 1. In particular, if an irreducible chain has $P_{i,i} > 0$ for some i, then it is aperiodic.

A Markov chain is said to be **symmetric** if $P_{i,j} = P_{j,i}$ for every pair of states i and j.

[Madras] [Madras-Liu]

1.1 Markov chain for pattern-avoiding binary matrices

2. An algorithm for testing pattern-avoidance of a general pattern

We begin with a very basic algorithm that we improve a lot to get a fast algorithm for testing avoidance of a general pattern.

2.1 Sketch of a brute force algorithm

Let $L = \{l_1, l_2, \cdots, l_{w+h}\}$ be the lines (rows and columns) of the pattern P. Partial mapping of lines of P is a function f from $L' \subseteq L$ to lines of the big matrix M satisfying two conditions:

- Both $l' \in L'$ and $f(l')$ are rows or they are both columns.
- If $l' \in L'$ and $l'' \in L'$ are both rows or columns and $l' < l''$, then $f(l') < f(l'')$. This means partial mapping keeps the order of the lines.
- If $l' \in L'$ is a row of P and $l'' \in L'$ is a column of P and there is a one-entry at the intersection of l' and l'', then there is a one-entry at the intersection of $f(l')$ and $f(l'')$.

The basic algorithm we use goes as follows. First it maps l_1 to all possible lines of M, creating partial mappings of $\{l_1\} \subseteq L$. For $k = 2, \cdots, w + h$ it takes each partial mapping from previous iteration and extends it by adding line l_k to the partial mapping in all possible ways. If we manage to map all the lines of P, then M does not avoid it and if at some point there are no partial mappings to extend it means M avoids P.

Note that we do not use even the slightest heuristics like not to map a line having k one-entries to a line with fewer one-entries. The fact we made a useless mapping will be discovered at the time we try to map the k-th orthogonal line (or maybe earlier). This leads to possibly very large number of partial mappings and our next goal will be to reduce this number significantly.

2.2 Equivalent mappings

There are a lot of possible heuristics we can check before adding a line to a partial mapping, but the most fundamental operation we do to decrease the number of found mapping is saying that two mapping are equivalent.

The idea behind it is very basic. There is no need to remember two different mappings if they can be both extended exactly the same way as our function is only supposed to check whether a pattern can be mapped to a big matrix not to find all such mappings.

A **level** of a partial mapping is the number of lines mapped by the mapping.

We call a line l of a pattern **important** in a partial mapping if one of the conditions is met:

- An adjacent line of the pattern has not been mapped yet.
- There is a one-entry on the line l at the intersection with line l' that has not been mapped yet.

. Otherwise the line is **unimportant** in the mapping.

At the beginning, when no line is mapped, all lines are important. After some lines get mapped a line can become unimportant in the partial mapping as all lines that bound in are in the mapping as well. If a line is unimportant in a partial mapping of some level, it will stay unimportant in all extensions of the mapping we can find.

We say two partial mappings of the same level are **equivalent** if all important lines in the mapping of that level are mapped to the same lines of the big matrix in both mappings.

Picture and use...

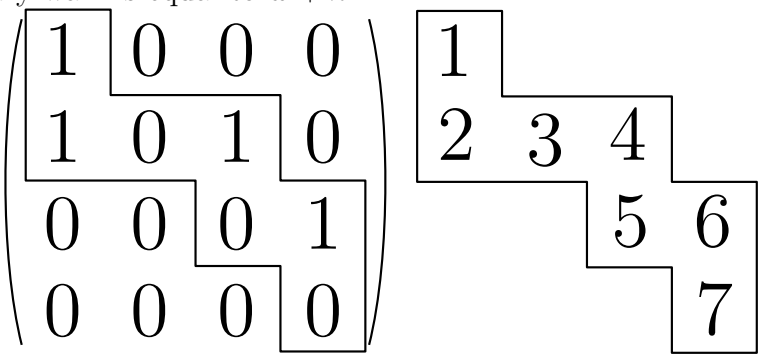
3. An algorithm for testing pattern-avoidance of a special pattern

In the previous chapter, we have seen an algorithm for a general forbidden pattern which, using some heuristics, runs pretty fast. In this chapter, we introduce a special kind of a pattern, satisfying additional conditions, for which we can produce much faster algorithm.

3.1 Walking pattern

We call the specific pattern a walking pattern. The additional condition we want the pattern to satisfy is that there is a walk from one corner to the opposite one and all the one-entries of the pattern are contained on the walk.

The **walk** in a matrix is a sequence of some of its entries beginning in the top left corner and ending in the bottom right one. If an entry at the position $[i, j]$ is in the sequence, the next one is either $[i + 1, j]$ or $[i, j + 1]$. Therefore the length of an arbitrary walk is equal to $w + h - 1$.



In the picture you can see a matrix that is a walking pattern as all the one-entries are included in a walk. Not all entries of a walk need to be one-entries though.

It can be shown a walking pattern with a walk is exactly a matrix avoiding a forbidden pattern

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

3.2 Dynamic program

Now that we know what the walking pattern is, we show an algorithm deciding whether a pattern P is contained in a big matrix M or not.

The pattern P is a walking pattern so there is a walk containing all the one-entries of the pattern. We choose one such walk arbitrarily and index its entries $w_1, w_2, \ldots, w_{m+n-1}$ starting from the beginning of the walk. For each entry of the walk we remember whether its value is one or zero and whether the walk continues

from the entry vertically, which we call a **vertical entry** or horizontally, calling it a **horizontal entry**.

For an element e of M at the position $[i, j]$, the matrix $M_{\leq e}$ is a $(i+1) \times (j+1)$ submatrix of M consisting of rows with the index smaller than or equal to i and columns with the index smaller than or equal to j. The element e lies in its bottom right corner. Similarly $M_{\geq e}$ is a $(n-i) \times (n-j)$ submatrix of M consisting of rows with the index greater than or equal to i and columns with index greater then or equal to j. The element e is its first element.

To determine whether P is contained in M we find out for each element e of M what is the longest part of the pattern that can be found in $M_{\leq e}$. If there is an element for which we manage to find the last entry of the pattern, the pattern is contained in the matrix, otherwise it is avoided.

For each element e of M at the position $[i, j]$ we remember two numbers. The number $c_v(e)$ says what is the longest part of the walk in $M_{\leq e}$ with the last entry in j-th column and being a vertical entry. The number $c_h(e)$, symmetrically, says what is the longest part of the walk in $M_{\leq e}$ with the last entry in i-th row and and being a horizontal entry.

An observation we make is that if we have a fixed element e of M and any other element e' above e in the same column then if $c_v(e')$ is equal to some k, then $c_v(e)$ is at least k. This means that for e we can find the maximum part of the pattern ending in the column of e and continuing vertically by looking only to elements in that column above e and since this is true for all of them, it is sufficient to only check the value of the element right above e (at the position $[i-1, j]$). Similarly the argument goes for the value of c_h in horizontal way.

The algorithm iterates through diagonals. A diagonal in this matter of speaking is a subset of elements of M, such that all elements have the same sum of their coordinates. For example, zero diagonal only consists of an element $[0, 0]$, the first diagonal contains elements $[0, 1]$ and $[1, 0]$, and so on.

For simplicity, in the pseudo-code below we do not deal with elements outside M (like $-1, 0$) explicitly. Instead for those elements we just assume the values of c_v and c_h are always equal to zero.

3.2.1 The algorithm

1. For $d = 0, \cdots, w + h - 1$
2. For e element of d-th diagonal at the position $[i, j]$
3. $e_v := [i-1, j]$
4. $e_h := [i, j-1]$
5. $c_v(e) := c_v(e_v)$
6. $c_h(e) := c_h(e_h)$
7. If $w_{c_v(e)+1}$ can be mapped to e
8. If $c_v(e) + 1 = w + h + 1$
9. Terminate - M contains P as a submatrix

10. If $w_{c_v(e)+1}$ is a vertical entry

11. $c_v(e) := c_v(e) + 1$

12. Else

13. $c_h(e) := max\{c_h(e), c_v(e) + 1\}$

14. If $w_{c_h(e)+1}$ can be mapped to e

15. If $c_h(e) + 1 = w + h + 1$

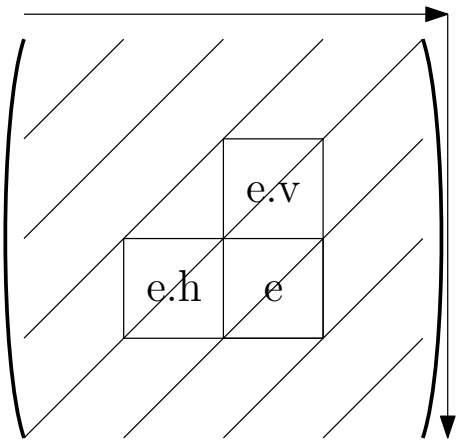
16. Terminate - M contains P as a submatrix

17. If $w_{c_h(e)+1}$ is a vertical entry

18. $c_v(e) := max\{c_v(e), c_h(e) + 1\}$

19. Else

20. $c_h(e) := max\{c_h(e), c_h(e) + 1\}$



3.2.2 Inner structure

To run properly the algorithm needs two structures. The first one is a description of the walk, which is just an array of the values of its entries as well as the information whether the entry is vertical or horizontal. The second structure is a matrix of the values c_v and c_h as described above.

3.2.3 Correctness

We need to show that the values of c_v and c_h are always correct for the recomputed elements after at the end of the function. We proceed by induction.

For the first element it is definitely true since there can be only the first entry of the pattern mapped and we check just that.

When we compute an element e of a computed diagonal d, by induction assumption all the diagonals $d' < d$ are correctly computed. In particular, the values are correct in the diagonal $d - 1$. To compute the correct values of e, we use the values of two element on the diagonal $d-1$: e_v, which is right above e and

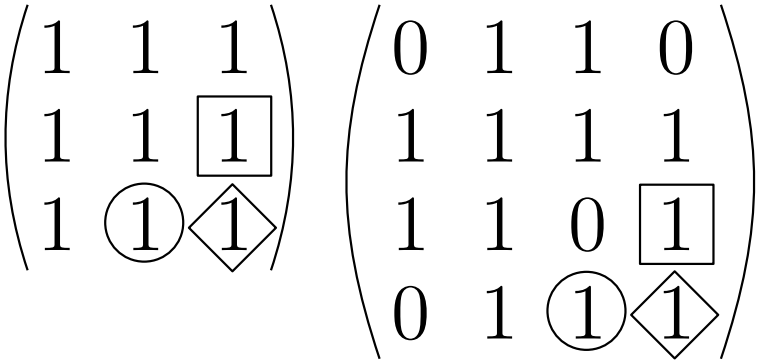
e_h, which is the first element to the left of e. If e_v or e_h are outside the matrix then from that direction we cannot expect to find anything more than just the first entry of the pattern and that is what we check for.

Let v be the true length of the longest part of P in $M_{\leq e}$ continuing vertically in the same column as e. Now if e itself is not an entry of that part of the pattern, it is a different element e' in the same column. But then the value of $c_v(e')$ is correctly computed by the inductive hypothesis and it is copied to all element underneath. Especially e_v gets the value and the algorithm copies the value from it to e. On the other hand if e is an entry of the part of the pattern we work with, it is the last entry. The entry right before the last one needs to be mapped to the same row or column; therefore, either e_v or e_h contain the part of the pattern shorter by one and the algorithm extends it to a correct value.

3.2.4 Generalization

The same algorithm, just rotated by 90 degrees, can be also used for a pattern where all one-entries are contained on a walk from top right corner to the bottom left one. Indeed the program uses it and if given a walking pattern it determines by itself which walk it is.

On the other hand a direct generalization for a general pattern does not work. While we can index all entries of the pattern, when trying to map a certain w_k to an element it is not sufficient to just check whether w_l is above and w'_l to the left from the element.



In the picture, let the matrix on the left side be the pattern P and let M be the other matrix. The entry in the square can be mapped to the element in the square and the same holds for entries in the circle but it is not a sufficient condition for the entry in the triangle to be mapped to the element in the triangle.

4. Improvements to basic algorithms

4.1 General pattern

4.1.1 Improving memory consumption

As the algorithm for testing avoidance of a general pattern was described in [chapter2], it creates all possible partial mappings and checks whether at least one can be extended to a full mapping. Note that to compute all the partial mappings of some level l, it only uses mappings of level $l-1$; therefore, it is enough to only store partial mappings of two levels in memory.

In [chapter2] we also introduced the idea of (un)important lines for a partial mapping of level l and equivalence based on not using unimportant lines at all - as they are fully bounded by other already mapped lines. When a line becomes unimportant it stays unimportant till the end of the run; as a result, we can forget where we mapped those lines to save memory. This is not as big of a deal as the previous observation was but note there are cases of patterns in which each line becomes unimportant just two levels after it gets added.

4.1.2 Not mapping empty lines

An empty line is a row or a column that does not contain any one-entries. Such a line can be mapped to any line and if the algorithm leaves space for it (which it does), we do not need to map it at all.

4.1.3 Using the last changed position

As the MCMC process works, it always changes one element of the big matrix and asks whether it still avoids the pattern. If it does not and we know that before the change it did, we are sure the changed element $[r, c]$ is a part of the pattern. It is hard to use this fact in the algorithm. It just maps one line after another and we do not know at the beginning to which line the changed position lines should be mapped.

What we can do is to enforce that neither the r-th line nor the c-th one get skipped. We will only look at the restriction for rows. The restrictions for columns are symmetrical. There are three situations we want to avoid:

- The first row of P is mapped under the r-th row. This prevents any other row to be mapped to r-th one and we don't want that.
- The last row of P is mapped above the r-th row. This again prevents any other row to be mapped to r-th one.
- Two adjacent rows $l, l+1$ of P are mapped to $L < L'$ respectively and $L < r < L'$ which leaves no other row to be mapped to r.

4.1.4 Line order

An important thing, if we want the algorithm to run fast, is to choose a good line order. A line which is unimportant in level l in a line order may easily be important till the nearly last level in a different order.

We choose line order to hopefully enforce two things:

- Make as many unimportant lines as possible. This really allows the equivalence based improvements to kick in. The more lines are unimportant the more mappings become equivalent and the faster it is to iterate through all of them.

- Recognize hopeless partial mappings as soon as possible. A partial mapping gets extended if the line does not break the rule that there is a one-entry where it needs to be. If we map all the rows first, the rule will get broken only after we start to map columns and we probably want to find out sooner.

In the program a user can either choose their own custom order or one of four algorithms with different main purposes:

- AUTO - this one tries the other three line orders and chooses the one which shows the best performance over some iterations on a matrix. While this may sound like a good thing to use, it is only so if an initial matrix is chosen and it takes a lot of time since a lot of iterations need to be made in order to make a good sample. I would recommend not to use AUTO order at all and instead to try all the line orders by hand with a number of iterations depending on the pattern and a good initial matrix; for instance, generated with a smaller number of iterations on the same pattern and with any line order.

- DESC - the lines are ordered in descending order depending on the number of one-entries. This follows the idea to start with the lines that are the hardest to map. Note that this algorithm does poorly if there are a lot of lines with the same number of one-entries (for example an identity matrix).

- MAX - it orders the lines so that the maximum number of important lines throughout the levels is as small as possible. This focuses straightforwardly to having many unimportant lines, which the program does not remember.

- SUM - it orders the lines so that the sum of the numbers of the important lines is the smallest possible throughout all levels. The purpose is the same as in the MAX order and quite often it is the case both approaches produce the same order.

- TWO - it orders the lines so that the maximum number of important lines in two consecutive levels throughout all the levels is as small as possible. This again focuses to having many unimportant lines, which the program does not remember. The constant two is chosen due to the fact general pattern always stores two levels of partial mapping at a time.

4.1.5 Mapping approaches

The one thing the approaches we will introduce have in common is that they try to recognize those partial mappings that have no chance to be extended to a full mapping as early as possible.

While the algorithm introduced in [chapter2] finds out the partial mapping is invalid only at the time it maps two lines having a one-entry at their intersection to two lines having a zero-entry at the intersection, different approaches try to reveal the fact we would end up in the situation earlier by checking more conditions.

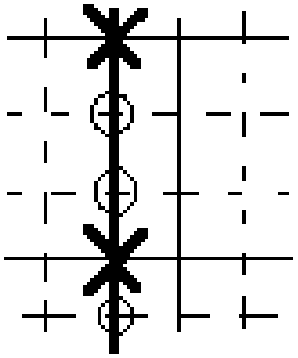
Imaging this as the pattern, where $0, \cdots, 8$ are indices of the lines:

	5	6	7	8
0	1	1	1	1
1	0	1	1	0
2	1	1	0	0
3	0	1	0	1
4	1	1	1	0

We are in a situation when only lines 0, 3 and 7 are mapped and line 6 is currently being mapped. Here are some mapping approaches:

Enough one-entries

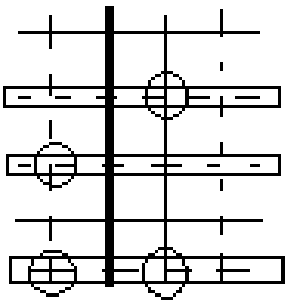
In the situation above, we not only want to check there are one-entries at the intersections of line 6 with lines 0 and 3, but we also check if there are enough one-entries in lines between where lines 0 and 3 are mapped so that there is a hope we can map lines 1 and 2 there and if there is a one-entry below the line where line 3 is mapped so we can map line 4 there later.



Recursive mapping

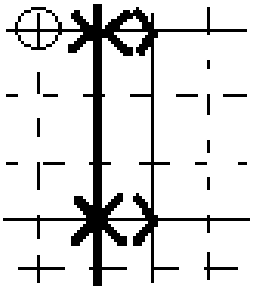
While we were only testing whether there are enough one-entries in between already mapped lines in the previous approach, this time, we also check whether those one-entries can be used for the lines that are intended to be mapped there. For example, when we check there is a one-entry to be used for line 1 later, we also check the line 1 can be mapped to that row, which in this situation means

to also check there is a one-entry at the intersection with the line to which the line 7 is mapped.



Orthogonal bounds

When adding line 6, we check whether there are enough one-entries on the already mapped lines orthogonal to line 6, in between line 6 and the closest mapped lines next to line 6. The same idea as in "Enough one-entries", but checking different lines.



Usage

These restrictions on the added lines are not a fixed part of the program. A user can decide which approaches they want to use in the configuration file. This is due to the fact that there is no right path to choose.

In the testing that was done for a fixed pattern, we found out it is useful to use all the mentioned restrictions when generating a matrix of size 100×100, as it turned out to be much faster than without the restrictions. On the other hand, in the same test for a generated matrix of size 500×500, it was much better not to use any of those restrictions.

4.1.6 Using the whole structure in the next iteration

It may seem like a good idea to remember all the partial mappings, to propagate them to the next iteration of the MCMC process and alter them depending upon

the change.

This really can be done. If the change is from zero-entry to one-entry, for each partial mapping we already have we want to try to extend it by the line that just changed and if we manage to do that we then try to extend it to a full mapping in all possible ways if it is a new mapping or do nothing if it is equivalent with a partial mapping of higher level. This can be easily done by means already used in the standard algorithm and may lead to a better performing one.

However, if the element gets changed from one-entry to zero-entry we need to go through the partial mappings and delete those that used the currently changed one-entry. This gets a bit messy as we can no longer forget unimportant lines and moreover for each partial mapping we need to remember how many partial mappings of the previous level can be extended to that one, to delete that mapping from the list if there are no longer any mappings extensible to that one.

This can all be done, but the whole thing comes with three huge inconveniences:

- Memory consumption - there can be a LOT of partial mappings and we need to remember them all. Of course we can still use the equivalence but we need to remember mappings of all levels.

- The change from one-entry to zero-entry is no longer for free. If this change is done, we already know the pattern is not contained in M, but we still need to do a lot of work to change the structure in order to use it in the next iteration.

- Reverting - if the change is unsuccessful (the pattern is contained) we need to revert the change which means to completely revert all changes we did to the list of partial mappings. This can be either done by making a backup copy of the whole structure and override the structure if needed, which again is very costly as the structure is huge, or we can remember what partial mappings are new and we go through all partial mappings and remove those new ones. This again means to iterate through the big structure one more time.

After realizing these issues it no longer looks useful to me to implement this version of the algorithm.

4.2 Walking pattern

While the brute force implementation of an avoid algorithm for a general pattern was improved heavily, the algorithm for a walking pattern is very fast in its nature and cannot be much better. Or can it be?

4.2.1 Using the last changed position

As the MCMC process works it always changes one element of the big matrix and asks whether it still avoids the pattern. If it does not and we know that before the change it did, we are sure the changed element is a part of the pattern. Knowing that and using the same inductive proof as we did in the proof of correctness of

the avoid algorithm (see [chapter2]) it is sufficient to only recompute the part of the inner structure under the changed element and check if the last entry of the pattern can be found there.

Not only that. We also know, using the fact the structure was completely correct before the change, that if the values of both c_v and c_h of an element did not change, the element won't cause the element underneath it to change and we no longer have to recompute the other parts of the structure.

To use both these facts we replace the cycle through the diagonals by a simple queue, starting at the position of the last changed element and putting more positions in if the values of c_v or c_h are different than they were before. The function ends either when the pattern was discovered or when the queue becomes empty.

4.2.2 Lazy avoid

Lazy avoid is a variant of avoid function used when the MCMC parallelism is chosen. While all the other types of patterns have a trivial implementation of revert function, when using the walking pattern the inner structure needs to be modified even when reverting. The MCMC parallelism turned out to work much better if the revert calls are handled by the main thread (more in [chapter4]) and it requires the function to run as fast as possible so the other threads are not blocked by the call for too long. That is a reason why functions lazy revert and lazy avoid were created.

The avoid function expects the inner structure of the walking pattern (see [chapter3]) to be in a valid state and that requires some effort. To make lazy revert the fastest possible, we postpone the work until the next call of lazy avoid, meaning that lazy avoid then needs to do more things at once. It is no longer sufficient to only compute the submatrix under the position changed last as we did above, but it needs to also compute changes in the positions changed in those lazy revert calls that are postponed.

We discuss several approaches, starting with the easiest one and ending with the one that is fast and used in the final implementation.

Recompute the whole structure every time

The easiest way how to implement lazy avoid would be to always recompute the whole inner structure. In that case we do not worry which positions are correct and which are not, because every time we find the pattern, we recomputed all the entries that form it, so we know it really is there. On the other hand, if we manage to recompute the whole structure without finding the last entry of the pattern, it just is not there.

The issue is efficiency. If the whole structure was correct and there was a change of the last entry of the matrix it is sufficient to only recompute that one entry. Instead we recompute a possibly very big structure. This results in a very bad performance negating the advantage of parallel computation.

Recompute only a part of the structure diagonal by diagonal

A simple improvement would be to remember the changes done in previous calls of lazy revert and together with the change done in lazy avoid call only recompute the part of the structure that has possibly altered.

This gets a bit tricky when lazy avoid call actually discovers the pattern because we cannot be sure the rest of the structure is in a correct order. It is still possible to remember some horizontal, vertical and diagonal bounds and use them to restrict the recomputed part of the matrix. The improvement is not that significant though and we can do better.

Queue of positions to recompute

A different approach is closer to the one used in a standard avoid function. Instead of going through diagonal one after another, we have a queue of entries-to-recompute. It is no longer sufficient to have a standard queue since in different calls of lazy revert/avoid we can possibly change an entry of different priority (the higher the more important) so we need to have some kind of a priority queue. That is exactly what I tried.

Using std::priority_queue the function had no more problems with recomputing the entries that were not influenced by the changes and used all the benefits mentioned in the previous section. But the container does not come for free and in the end I found out the price I payed for the operations on the priority queue made the whole implementation comparably slow as in the previous attempt.

Two leveled queue of positions to recompute

The final solution comes with the same idea, but a different storage type. As the priority depends upon a diagonal (two entries on the same diagonal can be recomputed in any order) we only remember a priority queue of diagonals and an array of diagonals saying whether a diagonal is already a member of the priority queue. As far as the entries are concerned for every diagonal we have a std::vector of entries-to-recompute as well as an array saying whether an entry is already a member of the vector. So finally it is the case that the storage used is not only good theoretically but as the numbers say, also practically. [reference to a table of measurements or something]

4.3 Parallel computing

4.3.1 MCMC parallelism

The first approach makes the MCMC generator parallel and it can be used with all types of patterns. While the serial MCMC generator changes one element in the generated matrix and checks whether it still avoids forbidden patterns, the parallel one counts on the fact at some point it is unlikely a change does not create a mapping of the pattern. So as it expects almost every change of one element to form a pattern, it tries to deal with those changes as fast as possible.

When the parallel version of MCMC generator is chosen and it is assigned n threads, it creates $n - 1$ copies of the generated matrix and assigns one thread,

called worker, to each of them. The last thread, which we call the main thread, makes one change of a bit in each matrix and makes the corresponding worker check the avoidance.

The job of a worker is, when woken up, only to check if its copy of the matrix still avoids the pattern when one bit is changed. On the other hand all synchronization is left to the main thread. We still want the generator to satisfy the conditions we have for the Markov chain (more in [chapter1]) in order to generate a random matrix. To achieve that we have an ID for each task it assigns and it always waits for the task with the lowest ID to end and depending on its result it continues:

- success: If the task with the lowest ID is successful, we can propagate the changed bit into the generated matrix. Also, since all other threads expected this one to fail, they compute a change that will never be propagated to the generated matrix and therefore is useless. What we do with that is we stop all the workers that compute a task with higher ID, revert the changes they made on their matrices and we synchronize the change computed by the successful thread. After all of this is done, all the matrices are the same.

- fail: If the task with the lowest ID is not successful, there is no change to propagate to the generated matrix. This still makes a valid iteration of the MCMC process though and since all the other threads expected this one to fail, they compute correct changes.

In both cases a new task is then assigned to the worker.

Speculative computing

It may easily happen that a task not having the lowest ID ends first. In that case we could just wait until it has the lowest ID and process it later. This would be pretty costly though. Instead we process the task immediately, just as described for the task with the lowest ID, but we don't propagate the changes to the generated matrix and we do not stop the workers processing a task with lower ID. This gets a bit tricky when the task succeeds. Not only we revert all the changes computed by tasks with higher ID and synchronize the changes computed by the worker (we would do that even if waiting for the task to have the lowest ID), but it might happen a task with even lower ID succeeds as well. This leads to a need of reverting the synchronizations we made. Luckily this is the only precarious situation we may encounter and it is not that hard to deal with it.

The way we deal with these inconveniences is described in [chapter5] and should be clear from the code itself.

Reverting and synchronizing in the main thread

The speculative computing discussed above is not the only improvement we can make. It turns out that it is costly to wake a thread so it computes a trivial function, sets a few atomic variables and falls asleep again. This happens a lot. Every time a call of avoid succeeds it makes other workers revert their changes and synchronize the successful change, which are both trivial functions.

To workaround this problem we make a theoretically bad decision which has very nice practical results. All the reverts and synchronizations are computed by the main thread instead of by an appropriate worker. There is no problem with concurrency because the worker is always asleep when a revert or sync calls are to be assigned and using the fact those calls are really trivial, it does not make the rest of threads wait for the main thread for too long while it computes changes.

5. Technical documentation

In this chapter, we cover those parts of the algorithm that may be hard to understand just from the code. This only means functions that are technically hard - functions with unexpected dependencies, side effects and so on. Algorithmic difficult tasks are explained in [chapter4].

5.1 General pattern

The general pattern class contains a lot of function. Most of them are easy to follow and they all should be commented enough in the code. The only part which deserves more attention is the constructor.

5.1.1 Construction

In the constructor of a general pattern, there are a few function that are easy in nature but as they somehow use each other it is hard not to lose track of their dependencies and results. In order to make this part of the code, which is a very important part indeed, more understandable, we go through the constructor and explain all that is happening in the order it is happening in.

Storing the pattern

The first thing, which is done right after initialization of variable, is storing the pattern. Instead of storing the pattern in a Matrix<bool>, I decided to use to store lines into a number, where in the binary coding a one-entry in the position i means there is a one-entry in the line at the intersection with i-th orthogonal line. This comes handy when computing lines orders. At the same time we also find those lines that are empty (more in [chapter4]) and remember them, because we do not have to map them at all.

Choosing the line order

After that we need to choose the right line order (again more in [chapter4]). To compute MAX or SUM order we just use a brute force algorithm that checks sequences of line adding and for each it computes how many lines are unimportant. Then it just chooses the order which is the best in chosen metric.

What to remember

In the next step, we find what do we need to remember in each level of partial mappings with respect to chosen order. As mention earlier, for MAX or SUM order it is already computed when finding the order, but for other variants of orders it is not, so we just compute it every time. What to remember is based on the equivalence introduced in [chapter2] and the decision not to remember unimportant lines (which we explained in [chapter4]).

Parallel bound indices

Now comes the hardest to follow part - precomputing the indices for searching for parallel bounds. The idea behind is simple. When we are adding a new line and we already have a partial mapping, it restricts to where we can add the line. For example, if there are three rows in the pattern and the rows 1 and 3 are mapped, the second one need to be mapped in between those two. The question is, where are those two lines mapped to? First we add in a chosen order and second we do not remember all lines, as some are unimportant. What do we want is to have a instant access to the index of the line, which bounds added line, in the partial mapping so we do not need to compute the index over and over again. That is exactly what gets computed when the function "find_parralel_bound_indices" is called. The series of other function calls follows just because we compute the indices for all added lines in the order in which they are going to be added.

Extending order

The last function, "find_extending_order" just specifies how the next partial mapping will look like a from where from the previous mapping the values will be copied. Again, unimportant lines play their role here and it may easily be the case from a partial mapping storing k lines, after mapping one more line, we end up with a partial mapping only storing $k - 1$ lines, because two lines become unimportant by adding the line.

5.2 Parallel computing

5.2.1 MCMC parallelism

While the idea behind MCMC parallelism is described in [chapter4.3] and the code is heavily commented, the work done by the main thread may still be hard to understand.

Let I be the ID the process is currently waiting for, that is, the lowest ID of a task that is being tested by a worker. In a structure called "queue" (which is std::vector<std::deque>) each worker has a queue of tasks related to it. In the queue, there are tasks that are either being computed or have been computed. The history of tasks is needed to allow reverting changes that should have not happen when the main thread encounters a different successful task with lower ID. There is no need to have a complete history of all tasks computed. There are only those tasks, that have higher ID than I or have lower ID, but those are going to be removed from the "queue" as soon as possible. The name "queue" is not random, it describes the order in which the tasks are being stored - the tasks with lower ID have been inserted earlier and therefore they are at the bottom.

Now that we know the most important structure let's see how the main thread works with that and what are the situations.

- pop_front: The main thread deletes the first tasks (the one with the lowest ID) if one of two things happen:
  - The ID of the task being deleted is equal to I. That means the change computed by the task is being propagated to the generated matrix and

there is no need to remember the task anymore. This also increases I, not necessarily by one.

  - The ID of the task being deleted is less than I. This situation is due to synchronization. The worker was supposed to synchronize a task computed by a different worker that did not have the lowest ID at the time. Therefore the task needs to be in the list of tasks so we can revert it if needed. If there is no need to revert it and the lowest ID gets greater or equal to the ID of the task, we can just delete it from the "queue".

- pop_back: There is only one reason to delete tasks from the end of the "queue" and that is reverting. Imagine there is a task with id J at the end of the "queue". Now a different worker computes a task with lower ID and finds out the change is successful. This means the task J won't propagate to the generated matrix and there in no use for it. If it is still being computed, we cannot do much about it, so we just tell the worker to stop computing and deal with it later. If the task is finished, we need to revert it, but only in case the task was successful, because if it was not, it had already been reverted by the worker. So we revert the task if needed and we can just delete it from "queue" as it will never be used.

- emplace_back: The main thread only inserts new tasks to the end of the "queue" and there are two reasons to insert:

  - Worker is assigned a completely new task to check the avoidance. In this situation the task is given a new, globally highest ID and we add the task at the end of the list.

  - The second reason to insert into "queue" are, again, synchronizations. The situation is the same as it was in the case, when we pop_back - after we revert all the tasks in the list, we need to synchronize changes that forced reverting and if their ID is not lower or equal to I, we need to add them to the list so they can be reverted if needed.

5.3 Library interface

6. User documentation

In the last chapter of the thesis we show how to make the program generate random matrices or to test whether a certain matrix avoids a given forbidden pattern. In the first section we find out how to set a configuration file. After that we go through necessary input and output files and see how they are formatted.

6.1 Configuration file

In order to modify what the program computes, we use a configuration file. There is only command line options to be used and that is a path to the configuration file. If no path is inserted the configuration file is expected to be located in the same directory as the executable file is and its name is "config.txt".

The file is a standard text file which can be modified by any text editor and is structured into four sections:

- input
- pattern
- output
- statistics

The order of the sections is not fixed and there can be additional empty lines for better readability. In each section there is a list of values that can be set either to arbitrary value or to a specific one. There is at most one command of format "option=value" per line and there might be additional white spaces surrounding the "=" sign.

If an option is set more than once, the latter value is always used. If, on the other hand, an option is not set at all, the default value is used. If there is a line encountered that sets a wrong option, for instance when the user mistypes a valid option, the line is skipped and the user gets a warning in the standard error output.

Let us provide a list of all options for each section together with their default values.

6.1.1 Input

In the first section of the configuration file we set the generating process.

- size: The size of the generated matrix. Results in $M \in \{0,1\}^{size \times size}$.

  Possible value: $s \in \mathbb{N}$
  Default value: 100

- iterations: The number of iterations of the MCMC process.

  Possible value: $i \in \mathbb{N}$
  Default value: 10,000

- random_seed: The random seed for the MCMC process.
  Possible value: $s \in \mathbb{N}$
  "random" - chooses a random seed
  Default value: "random"

- init_matrix: A $size \times size$ matrix the MCMC process starts with.
  Possible value: *input file path*
  "zero" - a matrix containing no one-entries
  Default value: "zero"

- parallel_mode: choose whether to compute in parallel or serial.
  Possible value: "serial"
  "mcmc" - more iterations of the MCMC generator are tested in parall
  "map" - more partial mappings are being extended in parallel
  Default value: "serial"

- threads_count: The number of threads that are going to be used if a parallel mode is chosen.
  Possible value: $t \in \mathbb{N}$
  -1 - chosen according to the number of cores
  Default value: 1

6.1.2 Pattern

It is time to set the options that matter the most - matrix patterns. As you generate a matrix which avoids more than just one pattern, the section [pattern] can be used multiple times, specifying one pattern for each occurrence.

- pattern_file: A path to a input matrix file - the pattern.
  Possible value: matrix file path
  Default value: "input.txt"

- pattern_type: The type of the pattern. Determines the method used for testing avoidance.
  Possible value: "general"
  "walking" - see [chapter3]
  "slow" - brute force algorithm for a general pattern
  Default value: "general"

The next options are only useful if the general pattern type is chosen. It specifies how the mappings are stored as well as what the map function tests.

- map_one_entries: If set to "yes", the map function not only checks that already mapped lines do cross the currently added one in a one-entry if needed but also tests whether there is enough one-entries in between those already mapped lines. The result is that the map function takes a little longer but it recognizes hopeless mappings earlier.
  Possible value: "yes"
  "no"
  Default value: "yes"

- map_recursion: If set to “yes” and the map_one_entries is also set to “yes”, the map function not only tests there are enough one-entries in between already mapped lines on the currently added one, but for those non-mapped lines crossing the current one in a one-entry, it also tests if the partial mapping has one-entries everywhere they are supposed to be. Although this comes with a big time consumption it can filter a big portion of partial mapping that would later become unfeasible.

  Possible value: “yes”
  “no”
  Default value: “yes”

- map_orthogonal_bounds: If set to “yes”, the map function also tests the orthogonal bounds of added line - see [chapter2].

  Possible value: “yes”
  “no”
  Default value: “no”

- map_container: A container in which the partial mappings are stored.

  Possible value: “set” - std::set (red-black tree)
  “hash” - std::unordered_set (hash table)
  “vector” - std::vector (dynamic array)
  Default value: “hash”

- line_order: Choose the order in which the lines are being added to the partial mapping. See [chapter2]

  Possible value: “max”
  “two”
  “sum”
  “desc”
  “auto”
  “order file path”
  Default value: “max”

6.1.3 Output

There is no reason to generate a matrix noone will ever see. In this section we make the matrix show. As the matrix can be output to console, a text file or a bmp file an option in the section can be set more than once and all the settings make difference.

- matrix_output: The generated matrix can be output as a bmp file in which one-entries are black pixels and zero-entries white. To do that, the file path has to have a pattern “path.bmp”. If a different path is given the file is stored as a matrix text file. It can also be output into a console if “console” is set. In that case it has the text format.

  Possible value: “console”
  matrix bmp file path
  matrix text file path
  “no”
  Default value: “no”

- performance_stats: If the serial computation or the map parallelism is chosen, the program can output a statistics like the percentage of avoid call success, how long did one call take on average and what was the average size of structures. Note that this is not very useful as a matrix statistics, see the next section for that. If more patterns are set, the statistics may get misleading as it also counts the cases when the first pattern is contained in the matrix and the other patterns are not tested at all.

  Possible value: "console"
  performance file path
  "no"
  Default value: "no"

- performance_csv_stats: The same information as above but formatted to a csv file so the data can be more easily worked with.

  Possible value: "console"
  csv file path
  "no"
  Default value: "no"

- time_to_console: Prints how long the computation took into a console.

  Possible value: "yes"
  "no"
  Default value: "no"

- patterns_to_console: Prints all the used patterns into the console.

  Possible value: "yes"
  "no"
  Default value: "no"

6.1.4 Statistics

The last section handles the options important for scientists. While generating a random matrix is a great result, on its way the program can also create some statistics, namely make a histogram of occurrences of one-entries in a generated matrix as the MCMC iterates as well as store the matrix with the highest amount of one-entries. As the process usually does not start with a random matrix, the user can decide to only compute the statistics after a certain number of iterations has been done and to only check a small portion of iterations, every 10th for instance, as a single iteration may not make any difference and counting the histogram takes time.

- histogram_frequency: Sets how often the histogram gets refreshed.

  Possible value: $f \in \mathbb{N}$
  0 - the histogram is not computed at all
  Default value: 0

- histogram_initial: Sets the initial iteration of the MCMC process when the histogram gets refreshed.

  Possible value: $i \in \mathbb{N}$
  Default value: 1,000

- histogram_final: Sets the last iteration of the MCMC process when the histogram gets refreshed.

  Possible value: $f \in \mathbb{N}$
  -1 - the histogram is computed till the end
  Default value: -1

- histogram_file: Sets where to output the histogram computed during the MCMC process.

  Possible value: matrix bmp file path
  matrix text file path
  "console"
  "no"
  Default value: "no"

- max_ones_matrix_file: Sets where to output the matrix that had the most one-entries among all matrices iterated through during the MCMC process.

  Possible value: matrix bmp file path
  matrix text file path
  "console"
  "no"
  Default value: "no"

6.2 File input

There are only two types of input files expected by the program. Either you want to read a matrix file, which can be a pattern or an initial matrix, or an order file that determines an order in which the lines are going to be mapped if the general pattern is chosen.

6.2.1 Matrix file

A matrix file is a standard text file having the format as follows:

- 2 natural numbers specifying the number of rows and columns in this order.
- a sequence of zeros and ones of length rows×columns specifying the matrix from the top left corner one row after another.

Example:

```
2 3
1 0 1
1 1 0
```

6.2.2 Order file

If you want to choose the order in which the lines are going to be mapped when a general pattern is chosen, it is your responsibility to check that all lines that need to be mapped are mapped. It is for example possible to only map three lines even if the pattern consists of six lines just because there is for example no need

to map empty lines at all. Therefore the program does not check the validity of the order and just uses it.

Now that the user has been warned, the format of the custom order file is simple. It consist of the indices of the lines of the pattern numbered starting with 0 and starting from the top row and ending with the right column.

One possible order for the matrix given as an example in [6.2.1] is this file:

2 1 0 3 4

First mapping the left column, the second and first row after that and finishing the mapping with the middle column and the right one.

6.3 File output

Let us now find out what the output files look like.

6.3.1 Matrix text file

The matrix text file has the same format as the input one. It consists of:

- 2 natural numbers specifying the number of rows and columns in this order.
- a sequence of zeros and ones of length rows×columns specifying the matrix from the top left corner one row after another.

The matrix is binary except for the one produced as a histogram, which can have higher natural numbers and contains the number of samples as the last number. If you then divide all the entries by the last number, you get a percentage of the entry being a one-entry.

6.3.2 Matrix bmp file

For an $n \times n$ matrix the standard bmp file contains $n \times n$ pixel of black color meaning a one-entry and a white color for a zero-entry. If the histogram is output as a bmp file, the pixels are greyscaled and the darker a pixel is the more often the entry was a one-entry during the MCMC process.

Conclusion

List of Figures

List of Tables

List of Abbreviations

Attachments