Sudoku as a Constraint Satisfaction Problem

Formal Definition

Sudoku can be defined as a Constraint Satisfaction Problem with:

- X – The set of variables
- D – The set of domains of each variable
- C – The set of constraints for each variable

For the case of Sudoku, these are as follows:

- X – The set of 81 variables named '1', '2', '3'……upto '81'. The variable named X_i represents the X_ith cell of the Sudoku
- D – where each D_i is defined as
  - D_i = {1, 2, 3, 4, 5, 6, 7, 8, 9} for an empty cell
  - D_i = {assignedValue} – The domain of an assigned cell only has a single value, which is the assigned value
- C – The set of 27 *AllDiff* constraints, which can be informally defined as:
  - One for each row - AllDiff(1,2,3,4,5,6,7,8, 9) for 1st row, AllDiff(10, 11, 12, 13, 14, 15, 16, 17, 18) for 2nd row etc upto 9th row
  - One for each column – AllDiff(1,10,19,28,37,46,55,64,73) for 1st column, AllDiff(2,11,20,29,38,47,56,65,74) for 2nd column etc upto 9th column
  - One for each box (3X3 square) – AllDiff(1,2,3,10,11,12,19,20,21) for 1st box, AllDiff(4,5,6,13,14,15,22,23,24) for 2nd box etc upto 9th box

  Where *AllDiff*(X1, X2, X3, X4,X5,X6,X7,X8,X9) is True if all the values of all the variables are different, otherwise False

With this definition, a **state** of the Sudoku CSP is any assignment (partial or complete) of the 81 variables. And a **solution** of the the Sudoku CSP is any complete assignment of all the 81 variables satisfying all the 27 *AllDiff* constraints

Results

Version A: Standard Backtracking Search

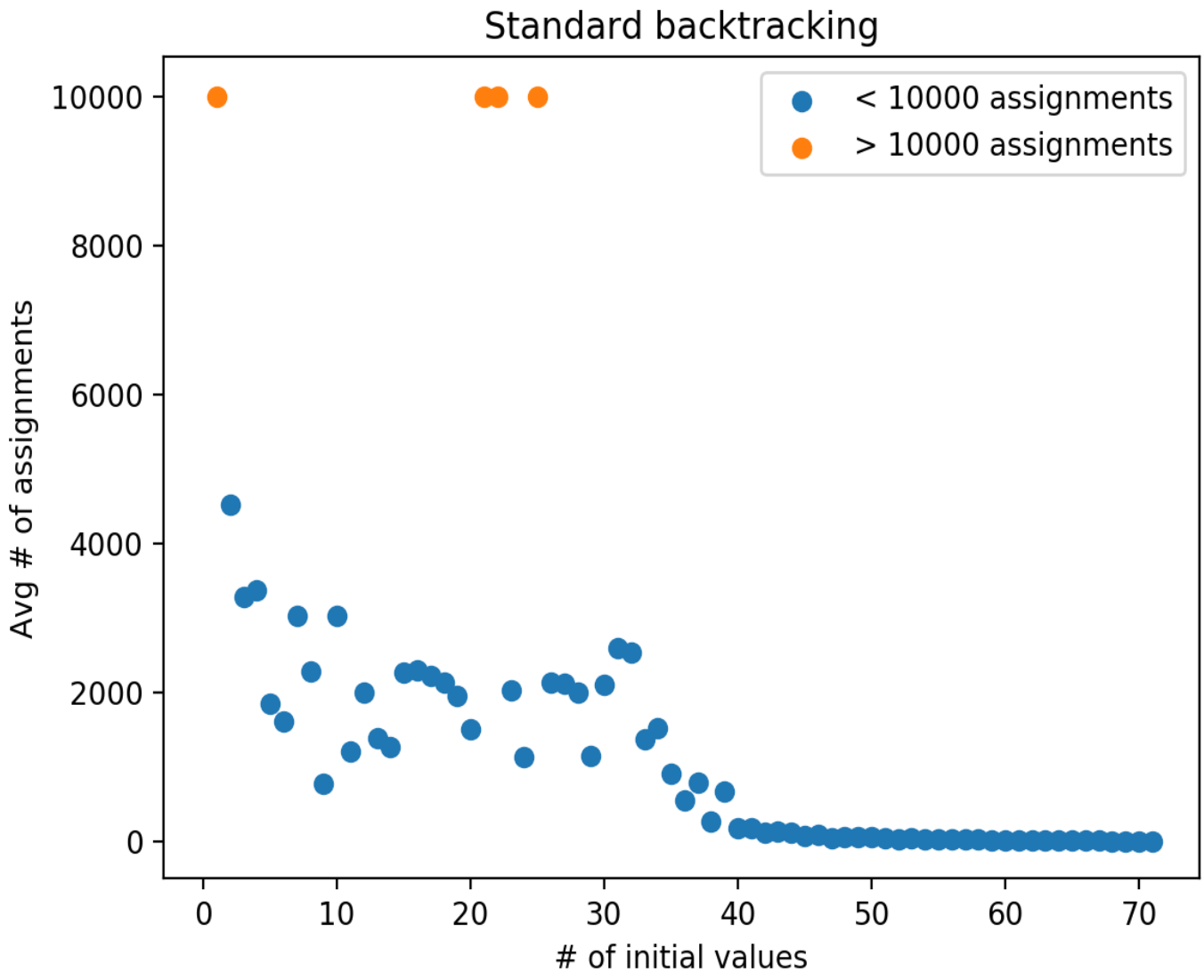

Figure 1

- Out of 710 instances, 90 instances crossed 10, 000 assignments
  - For 4 data points (where each data point Is the average of 10 sudoku problems), 10, 000 steps was crossed majority of the times. This is represented by the 4 orange dots

Version B: Standard Backtracking Search with Forward Checking

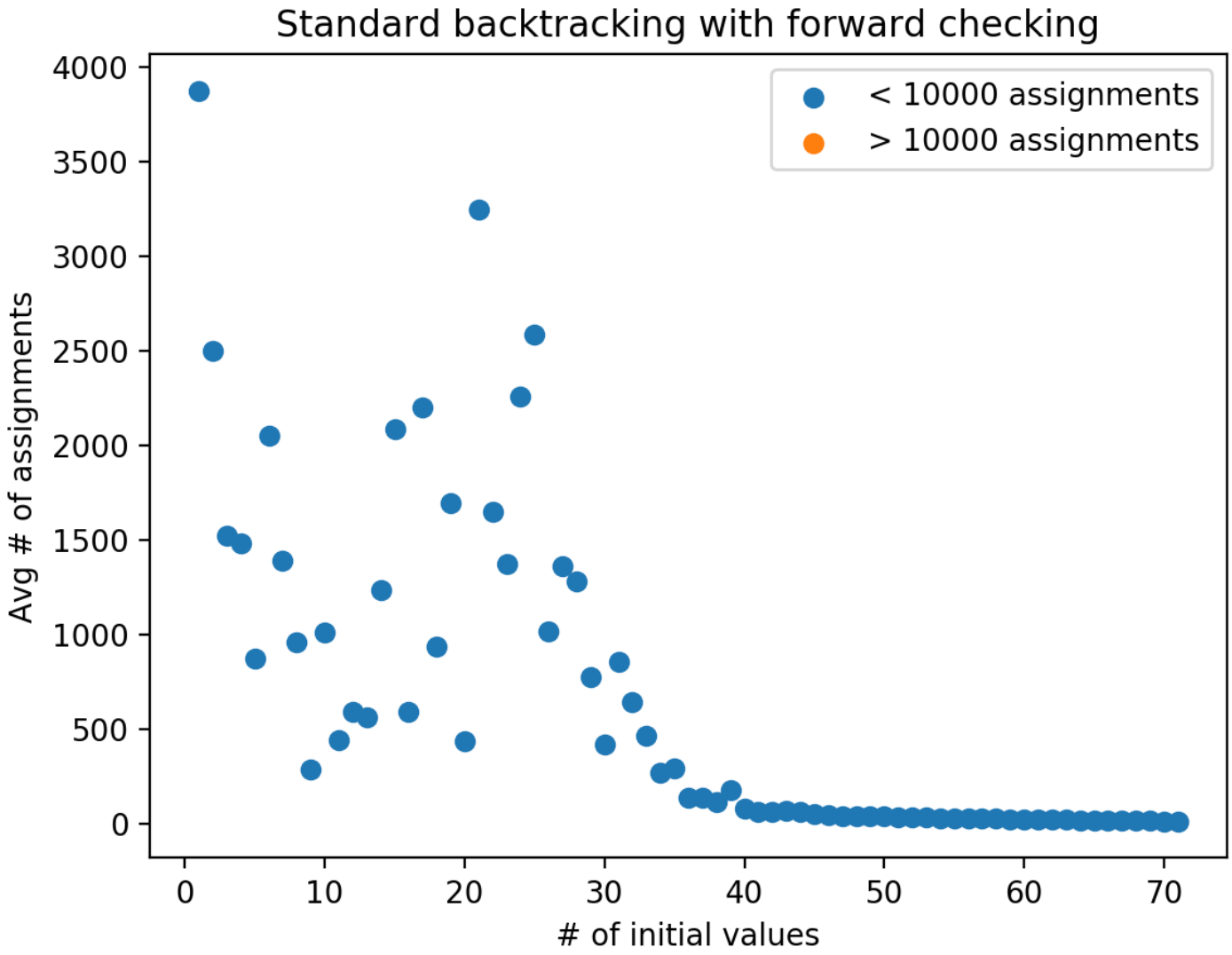

Figure 2

- Out of 710 instances, 43 instances crossed 10, 000 assignments

Version C: Standard Backtracking Search with Forward Checking and Heuristics (MRV, MCV and LCV)

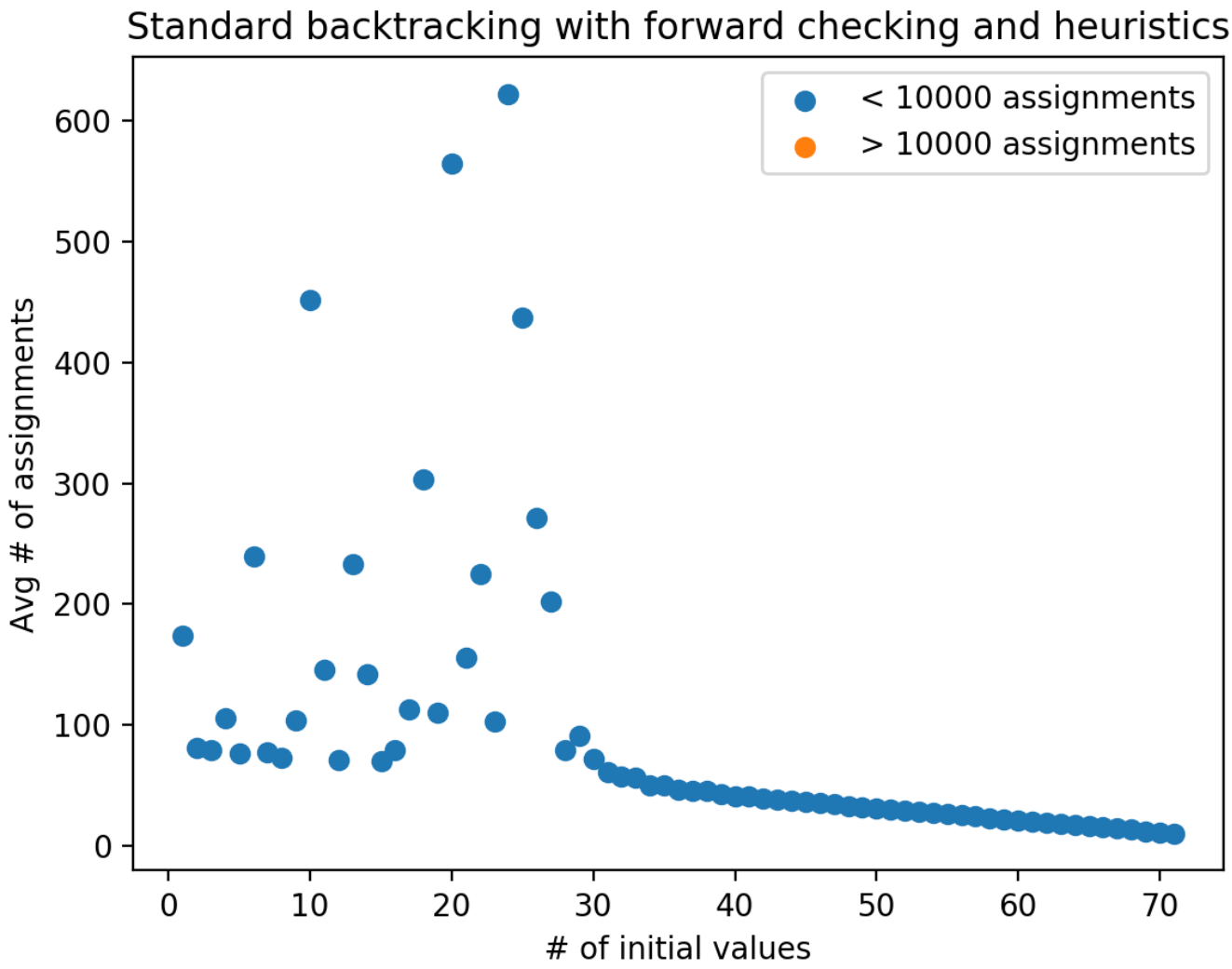

Figure 3

- Out of 710 instances, 14 instances crossed 10, 000 assignments

General Observations across all the 3 versions:

The entire state space, including solution Sudokus and incomplete Sudokus, is incredibly huge. As calculated by Bertram Felgenhauer and Frazer Jarvis in 2005, the number of solution sudokus alone is 6,670,903,752,021,072,936,960.

- The number of times, the algorithm crosses 10, 000 steps is decreasing as we go from version A to B to C
- For any given version (A or B or C), the Average # of assignments (y-axis value) is on a decreasing trend in general. This means that greater the number of initial values, the easier is the Sudoku and the algorithm solves it faster. But, there are cases where Average # of assignment increases, even though # of initial value increases, i.e y2 > y1 when x2 > x1. This is because, the count of initial values alone does not help in solving the Sudoku. How the initial values are distributed (whether they are spread across the Sudoku, whether they are same numbers or different numbers) plays a significant role.

  For example, consider the below 2 Sudokus both with 17 initial values:

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

Figure 4: Version C solved it in 196 assignments

		2		9		3		
8		5						
1								
	9			6			4	
							5	8
								1
	7					2		
3			5					
			1					

Figure 5: Version C took > 10, 000 assignments

- In terms of performance, Version C > Version B > Version A as shown by the plots and the below table
  - 2 exceptions to this are marked in red in the below table, where Version A performed better than Version B

# of initial Values	Version A (Avg # of assignments)	Version B (Avg # of assignments)	Version C (Avg # of assignments)
1	>10000	3872	174
2	4523	2499	80
3	3298	1522	79
4	3376	1482	105
5	1861	870	76
6	1621	2048	239
7	3044	1390	76
8	2284	959	73
9	782	286	103
10	3041	1007	452
11	1213	437	145
12	2000	588	71
13	1395	561	233
14	1271	1235	142
15	2273	2086	69
16	2304	590	78
17	2232	2196	112
18	2143	933	303
19	1967	1693	110
20	1518	432	565
21	>10000	3245	155
22	>10000	1649	225
23	2036	1369	102
24	1137	2255	622
25	>10000	2582	437
26	2146	1014	271
27	2124	1357	201
28	2011	1281	79
29	1156	772	91
30	2112	418	71
31	2607	851	60
32	2547	641	57
33	1378	462	56
34	1531	270	50
35	915	289	49
36	563	136	45

37	793	136	45
38	275	113	45
39	683	177	42
40	186	76	41
41	192	60	41
42	119	58	39
43	139	63	38
44	122	59	37
45	78	48	36
46	94	44	35
47	55	38	34
48	69	37	33
49	64	37	32
50	61	37	31
51	52	34	30
52	41	31	29
53	53	31	28
54	34	28	27
55	40	27	26
56	33	26	25
57	30	25	24
58	31	24	23
59	25	22	22
60	24	21	21
61	25	20	20
62	22	19	19
63	19	18	18
64	19	17	17
65	17	16	16
66	16	15	15
67	15	14	14
68	13	13	13
69	12	12	12
70	11	11	11
71	10	10	10

- When the # of initial values approaches closer to 81, the performance of all 3 versions become same as the Sudoku gets easier to solve
    - Forward Checking has no values to eliminate from the domains of the neighboring cells(same row, same column and same box)
    - Ordering of the variables and their values don't work anymore because there are fewer remaining variables with fewer possible values

Assumptions for the plots:

- For each data point on x-axis, the y-axis value is the average # of assignments over the 10 instances of the Sudokus with same # of initial values
- For problems where # of assignments crossed 10,000, followed the below given approach:
    - The orange dot indicates if it crossed 10, 000 or not, instead of showing the actual # of assignments
    - Out of the 10 instances of the Sudoku problem, only if majority of them (> 5) crossed 10, 000 steps then it is classified as an orange dot
    - If only a minority of them (< 5) crossed 10,000 steps, then the average was taken over the rest of the values (which took < 10, 000 steps) and that Average # of assignments was plotted

Implementation Details

How to Run:

- Requires Python 3, Numpy, Matplotlib
- Download the code into your folder. Important files are:
    - variable.py – class for representing a CSP variable and the operations on the variable
    - sudoku_grid.py – class that loads the Sudoku from input file. Contains methods to play around with the Sudoku like editing the Sudoku, printing the Sudoku
    - sudoku_csp.py – the main class that performs much of the CSP related operations on the variables, domains and constraint checking
    - backtrack.py – contains the search algorithm. Implemented a generic algorithm from which version A, B and C can be called upon using Boolean switch

- sudoku_runner.py – for running the algorithm on all the 710 instances
- result_plotter.py – for obtaining the plots and the output data

1. To run on a specific Sudoku:
    a. Store the Sudoku in a file and note down the absolute file path
    b. Open the backtrack.py file and edit the *inputFilePath* variable with the path of your Sudoku file and click Save
    c. To Run version A
        i. Set values of the variables *forwardCheck*, *mrvHeuristic*, *maxDegreeHeuristic* and *lcvHeuristic* to all *False*
    d. To Run version B
        i. Set value of *forwardCheck* to *True* and value of variables *mrvHeuristic*, *maxDegreeHeuristic* and *lcvHeuristic* to all False
    e. To Run version C
        i. Set values of the variables *forwardCheck*, *mrvHeuristic*, *maxDegreeHeuristic* and *lcvHeuristic* to all *True*
    f. Run backtrack.py
    g. The solution and the # of assignments is printed on the terminal. The algorithm gives up if # of assignments crosses 10,000

2. To run on all the 710 instances of the Sudoku
    a. Open sudoku_runner.py and edit the variable *inputRootDirectory* with the absolute path of the root folder where your input sudokus are stored and Save
    b. Run sudoku_runner.py. This will take a long time and the progress will be displayed on terminal
    c. An output-data folder will be created, which contains the solutions and # of assignments in appropriate files
3. To plot the results for a version
    a. Make sure you have the output-data generated by running sudoku_runner.py
    b. Open result_plotter.py and edit the variable *version*. Set as ‘v1’ for version A, ‘v2’ for version B and ‘v3’ for version C.
    c. Run result_plotter.py

Assumptions for the implementation:

- The initial configuration of the Question Sudoku is considered while setting the domains for all the variables in the beginning. That is, the domain of an unassigned cell is not always {1, 2, 3, 4, 5, 6, 7, 8, 9}. Instead it is
    - {1, 2, 3, 4, 5, 6, 7, 8, 9} – {the values in neighbors which are assigned in the initial configuration of the Question Sudoku}.
- A list of *Variable* objects is used to represent the variables of the CSP. Each variable has a *name* attribute and a *value* attribute.
    - name – represent the name/identifier of the variable. For Sudoku it can take the strings '1', '2', …., '81' each representing a cell
    - value – the current value of the Sudoku cell
- '0' is an allowed value representing an unassigned cell