Sudoku Players' Forums

FAQ Search Memberlist Usergroups
Profile You have no new messages Log out [denis_berthier]

THE REAL DISTRIBUTION OF MINIMAL PUZZLES

Goto page Previous 1, 2, 3 ... 41, 42, 43

newtopic postreply **Sudoku Players' Forums Forum Index -> General/puzzle**

View previous topic :: View next topic

Author **Message**

denis_berthier Posted: Tue Oct 27, 2009 1:00 pm Post subject: quote edit

Joined: 19 Jun 2007
Posts: 1137
Location: Paris, France

NEW RESULTS FOR THE DISTRIBUTION OF CLUES AND THE MEAN NUMBERS OF MINIMALS

Since my last results on the distribution of clues, here: http://www.sudoku.com/boards/viewtopic.php?t=14615&postdays=0&postorder=asc&start=554, I've let the controlled-bias generator run and I now have a sample of 3,884,587 minimal puzzles generated with gsf's collection piped into 3 versions of suexg-cb:
- with optim46 and U4,
- with optim48 and U4,
- Paul's version.
I checked that the 3 versions give consistent results (which was to be expected as they differ only by optimisations of the same algorithm) and I amalgamated them.

As of now, I have done 183 full scans of gsf's collection.
There was some (very small: 0.01%) discrepancy in the number of tries, whether one counted them directly in the algorithm (as I did when I needed this datum, i.e. only for the estimated number of minimals) or one multiplied the number of gsf scans with the number of gsf grids. This discrepancy was so small that it would have been harmless even if it had not been random.
But, for some time, I've been wondering about the cause. I've found.
UNIX piping is far from perfect: there is some (very small) random data leak. In the present case, a very small percentage of the gsf grids are lost before reaching suexg-cb. This has no impact on the following results, as one can consider this data leak as a (real, not pseudo) random sampling of the input, with high probability of acceptance.
The leak seems to be different on different machines. It is very low on my Mac and a little higher on other Unix machines I've been able to use occasionally. The positive aspect of this difference is, I could check that the results don't depend on the leak.

I've found my second 32 and my first 20.

Here are the new, more precise estimates for the controlled-bias and the real distributions of clues:

Code:

```
#clues  #instances    %              unbiased %    standard deviation of
unbiased %
        in cb sample in cb sample   (estimated)    (estimated)
20      1            2.574e-05      1.01e-07       1.001e-07
21      106          0.0027         3.10e-05       3.02e-06
22      4365         0.1124         0.003487       5.28e-05
23      72234        1.860          0.1480         0.00055
24      461557       11.88          2.2859         0.0034
25      1188634      30.60          13.422         0.012
26      1313976      33.83          31.957         0.028
27      659918       16.99          32.694         0.040
28      161868       4.167          15.466         0.038
29      20489        0.5274         3.5778         0.025
30      1390         0.03578        0.4207         0.011
31      47           0.001201       0.0234         0.0034
32      2            5.149e-05      0.00156        0.0011
```

```
controlled-bias mean = 25.6665      controlled-bias standard-deviation =
1.116
controlled-bias skewness = 0.086    controlled-bias kurtosis = 0.0255

real mean= 26.577                   real standard-deviation= 1.117
```

and for the mean numbers of minimals:

Code:

```
#clues    mean #minimals       relative error
          per complete grid
20        4.69e+6              100.0%
21        1.445e+9             9.7%
22        1.623e+11            1.5%
23        6.888e+12            0.37%
24        1.0637e+14           0.15%
25        6.2454e+14          0.092%
26        1.4870e+15          0.087%
27        1.5213e+15          0.12%
28        7.1965e+14          0.25%
29        1.6648e+14          0.70%
30        1.9576e+13          2.68%
31        1.089e+12           14.58%
32        7.24e+10            70.7%
```

Back to top profile pm www

Red Ed

Joined: 06 Jun 2005
Posts: 1022

Posted: Tue Oct 27, 2009 6:22 pm Post subject: quote

... and for comparison, for 30s, I have 1.969458e+13 w. relative error 2.38%
(EDIT 5th Nov: 1.960626e+13 w. relative error 1.16%)

As usual we are in blissful agreement ...

Back to top profile pm

denis_berthier

Joined: 19 Jun 2007
Posts: 1137
Location: Paris, France

Posted: Sat Nov 07, 2009 9:10 am Post subject: quote edit

FINAL RESULTS FOR THE REAL DISTRIBUTION OF CLUES AND THE MEAN NUMBERS OF MINIMALS

The controlled-bias generator has continued running in the background. As of now, it has accomplished 279 full scans of gsf's collection (with the remarks in my previous post still valid) and (having done a total of 1,526,116,703,532 tries), it has produced a sample of 5,926,343 minimal puzzles generated with gsf's collection piped into 3 versions of suexg-cb:
- with optim46 and U4,
- with optim48 and U4,
- Paul's version.
I have checked again that the 3 versions give consistent results (which was to be expected as they differ only by optimisations of the same algorithm) and I amalgamated them.

I've stopped the cb-generator and these will be my final results relative to the distribution of clues. The precision I get is now much beyond what I needed for my complexity estimates.
You can find these estimates in the "Rating" thread, mainly in the following 3 posts:
http://www.sudoku.com/boards/viewtopic.php?t=5995&postdays=0&postorder=asc&start=425
http://www.sudoku.com/boards/viewtopic.php?t=5995&postdays=0&postorder=asc&start=429
http://www.sudoku.com/boards/viewtopic.php?t=5995&postdays=0&postorder=asc&start=433
or on the classification page of my website:
http://www.carva.org/denis.berthier/HLS/Classification.

Here are my final estimates for the controlled-bias and the real distributions of clues:

Code:

```
#clues  #instances     %               unbiased %     standard deviation of
unbiased %
        in cb sample   in cb sample    (estimated)    (estimated)
20      2              3.7e-05         1.32e-07       0.93e-07
21      164            0.0027          3.14e-05       0.25e-05
22      6,651          0.1124          0.00348        0.00043
23      110,103        1.858           0.148          0.00045
24      704,089        11.88           2.285          0.0027
25      1,814,413      30.62           13.425         0.010
26      2,002,349      33.79           31.909         0.023
27      1,007,700      17.00           32.712         0.033
28      247,259        4.172           15.480         0.031
29      31,449         0.531           3.598          0.020
30      2,088          0.0352          0.414          0.009
31      74             0.00125         0.0241         0.0028
32      2              3.37e-05        0.00102        0.0007

controlled-bias mean = 25.667      controlled-bias standard-deviation =
1.116
controlled-bias skewness = 0.087   controlled-bias kurtosis = 0.024

real mean= 26.577                  real standard-deviation= 1.116
```

and for the mean numbers of minimals per complete grid:

Code:

```
#clues    mean #minimals       relative error
          per complete grid
20        6.152e+6             70.7%
21        1.4654e+9            7.8%
22        1.6208e+11           1.13%
23        6.8827e+12           0.30%
24        1.0637e+14           0.12%
25        6.2495e+14           0.074%
26        1.4855e+15           0.071%
27        1.5228e+15           0.10%
28        7.2063e+14           0.20%
29        1.6751e+14           0.56%
30        1.9277e+13           2.2%
31        1.1240e+12           11.6%
32        4.7465e+10           70.7%

all       4.6553e+15           0.065%
```

which
- multiplied by the number of complete grids (6,670,903,752,021,072,936,960) gives **3.1055e+37 minimal puzzles**
- multiplied by the number of non isomorphic grids (5,472,730,538) gives "only" **2.5477e+25 non equivalent minimal puzzles** (still with 0.065% relative error)

Last edited by denis_berthier on Fri Dec 18, 2009 8:42 am; edited 4 times in total

Back to top profile pm www

JPF

Joined: 07 Dec 2005
Posts: 3066
Location: Paris, France

Posted: Sun Nov 08, 2009 9:48 pm Post subject: quote

> **denis_berthier wrote:**
>
> and for the mean numbers of minimals per complete grid:
>
> **Code:**
>
> ```
> #clues    mean #minimals       relative error
> ```

> | | per complete grid | |
> |---|---|---|
> | 20 | 6.152e+6 | 70.7% |
> | 21 | 1.4654e+9 | 7.8% |
> | 22 | 1.6208e+11 | 1.13% |
> | 23 | 6.8827e+12 | 0.30% |
> | 24 | 1.0637e+14 | 0.12% |
> | 25 | 6.2495e+14 | 0.074% |
> | 26 | 1.4855e+15 | 0.071% |
> | 27 | 1.5228e+15 | 0.10% |
> | 28 | 7.2063e+14 | 0.20% |
> | 29 | 1.6751e+14 | 0.56% |
> | 30 | 1.9277e+13 | 2.2% |
> | 31 | 1.1240e+12 | 11.6% |
> | 32 | 4.7465e+10 | 70.7% |
> | | | |
> | all | 4.6553e+16 | 0.22% |
>
> which
> - multiplied by the number of complete grids (6,670,903,752,021,072,936,960) gives **3.1055e+38 minimal puzzles**
> - multiplied by the number of non isomorphic grids (5,472,730,538) gives "only" **2.5477e+26 non equivalent minimal puzzles** (still with 0.22% relative error)[/i]

I think the mean #minimals per grid is 4.655e+15 and the number of minimal puzzles should be modified accordingly.

JPF

Back to top profile pm

Red Ed

Joined: 06 Jun 2005
Posts: 1022

Posted: Sun Nov 08, 2009 10:01 pm Post subject: quote

While edits are on the cards, can we have the total number of trials too? Without that, the mean #minimals figures cannot be verified.

Back to top profile pm

denis_berthier

Joined: 19 Jun 2007
Posts: 1137
Location: Paris, France

Posted: Mon Nov 09, 2009 5:26 am Post subject: quote edit

JPF wrote:

> I think the mean #minimals per grid is 4.655e+15 and the number of minimal puzzles should be modified accordingly.

Additions !!!

Back to top profile pm www

denis_berthier

Joined: 19 Jun 2007
Posts: 1137
Location: Paris, France

Posted: Sat Nov 14, 2009 10:33 am Post subject: quote edit

A (PERHAPS NOT SO) NAIVE UNBIASED GENERATOR OF UNCORRELATED N-CLUE MINIMAL PUZZLES, N FIXED

An easy application of the previous results.

For each n, what's the mean proportion of n-clue minimals among n-clue subgrids of a complete grid?

Equivalent question:
Choose an integer n. Randomly choose a complete grid (e.g. from gsf's stream), randomly delete 81-n clues (without doing any test), thus obtaining an n-clue grid P (which may have 0, 1 or more solutions). How many times in the mean should you repeat this procedure before you get a minimal puzzle?
(The number of tries is the inverse of the above proportion.)

The following table gives the answer as a function of n.

Code:

```
n       #tries(n)
20      7.6306e+11
21      9.3056e+09
22      2.2946e+08
23      1.3861e+07
24      2.1675e+06
25      8.4111e+05
26      7.6216e+05
27      1.5145e+06
28      6.1721e+06
29      4.8527e+07
30      7.3090e+08
31      2.0623e+10
32      7.6306e+11
```

Method: for each n, divide the number of n-clue subgrids - i.e. 81! / n! / (81-n)! - by the mean number of n-clue minimals given in my previous post.

This method can't give the distribution-of-clues of minimals * (which isn't a problem, as we now already know it), but, as there is only one test per try, it can provide a relatively fast generator of n-clue minimals when #tries(n) isn't too large.
Unfortunately, #tries(n) increases fast when n goes above 31 or below 21.

(*) Indeed, it could provide it indirectly, if we keep the mean numbers of tries for each n.

Back to top profile pm www

denis_berthier Posted: Sat Jan 09, 2010 12:00 pm Post subject: quote edit

Joined: 19 Jun 2007
Posts: 1137
Location: Paris, France

Non correlation of the controlled-bias puzzles

Immediately after defining the controlled-bias generator (http://www.sudoku.com/boards/viewtopic.php?t=14615&start=134), I gave formulæ explaining how to use the puzzles it produces (in section 3).
In all this thread, I have always implicitly assumed the source of complete grids was uncorrelated (a very natural assumption). This obviously entails that the controlled-bias puzzles obtained from this source are uncorrelated.
This hypothesis is not necessary everywhere; it is useless for the computation of the mean of a random variable X (in section 3), but it is necessary for the computation of its variance and standard deviation (in section 3 also).

As for the gsf's collection used for all the final computations, it is essentially uncorrelated, i.e. uncorrelated modulo isomorphisms (*), which is as good as being uncorrelated, provided that variable X itself is invariant under isomorphisms.
It is the case for X = the NRCZT rating.
It is not strcictly the case for X = the SER rating, but I don't think this changes a lot about the SER classification results.

* From an essentially uncorrelated collection (of complete grids or puzzles), one can obtain an uncorrelated collection, just by applying a random isomorphism to each element (of course, all these isomorphisms must be uncorrelated).

Back to top profile pm www

Red Ed Posted: Sat Jan 09, 2010 2:01 pm Post subject: quote

Joined: 06 Jun 2005
Posts: 1022

As explained on the original minimal puzzles thread, non-correlation of puzzles is **not necessary** when estimating the relative error of the (experimental) mean number of minimals. Link: <here>. There's a computational cost associated with this "bonus" feature of non-correlation which makes *suexg-cb* less effective than subsets/supersets for minimal count estimation when n<26ish or n>29ish.

Back to top profile pm

denis_berthier Posted: Sat Jan 09, 2010 3:53 pm Post subject: quote edit

Joined: 19 Jun 2007
Posts: 1137
Location: Paris, France

> **Red Ed wrote:**
> As explained on the original minimal puzzles thread, non-correlation of puzzles is **not necessary** when estimating the relative error of the (experimental) mean number of minimals. Link: <here>. There's a computational cost associated with this "bonus" feature of non-correlation which makes *suexg-cb* less effective than subsets/supersets for minimal count estimation when n<26ish or n>29ish.

Less effective remains to be proven, as you say in the other thread that you haven't tried it. But that's secondary.
The question of interest for me is that non correlation **is** necessary for estimating other variables than #clues.

Back to top profile pm www

Red Ed

Posted: Sat Jan 09, 2010 4:17 pm Post subject: quote

Joined: 06 Jun 2005
Posts: 1022

I don't think there's any doubt that your algorithm is less effective outside of the central part of the distribution; but, if you like, you can point me to the latest implementation and I will perform a test to demonstrate it.

Good spot re gsf's collection needing to be randomly morphed before the SE Rating can be fully trusted; I'd forgotten that SE was unstable like that. What a nuisance (though hopefully, as you indicate, not a particularly significant one).

Back to top profile pm

denis_berthier

Posted: Sat Jan 09, 2010 6:02 pm Post subject: quote edit

Joined: 19 Jun 2007
Posts: 1137
Location: Paris, France

> **Red Ed wrote:**
> I don't think there's any doubt that your algorithm is less effective outside of the central part of the distribution

How can you compare two algorithms that don't do the same thing (uncorrelated vs correlated)?

Back to top profile pm www

Red Ed

Posted: Sat Jan 09, 2010 6:04 pm Post subject: quote

Joined: 06 Jun 2005
Posts: 1022

I'm just talking about the job of estimating the total number of *n*-clue minimals; nothing more.
Thus relative error as a function of time is a good comparative measure, as you suggested earlier.

Back to top profile pm

denis_berthier

Posted: Sat Jan 09, 2010 6:13 pm Post subject: quote edit

Joined: 19 Jun 2007
Posts: 1137
Location: Paris, France

> **Red Ed wrote:**
> I'm just talking about the job of estimating the total number of *n*-clue minimals; nothing more.
> Thus relative error as a function of time is a good comparative measure, as you suggested earlier.

dialogue de sourds !
I'm mainly interested in using the controlled-bias puzzles for estimating any random variable, especially (SER or NRCZT) complexities, not only number of clues.

Back to top profile pm www

Red Ed

Posted: Sat Jan 09, 2010 6:39 pm Post subject: quote

Joined: 06 Jun 2005
Posts: 1022

> *dialogue de sourds !*
Reminds me of an excellent QOTSA album.

Back to top profile pm

Display posts from previous: All Posts Oldest First Go

newtopic postreply **Sudoku Players' Forums Forum Index -> General/puzzle**

All times are GMT + 1 Hour
Goto page Previous 1, 2, 3 ... 41, 42, 43

Page 43 of 43

Stop watching this topic

Jump to: General/puzzle Go