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 ... , 10, 11, 12 Next

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

View previous topic :: View next topic

Author **Message**

coloin

Joined: 06 May 2005
Posts: 1050
Location: Devon UK

Posted: Sat Jul 18, 2009 11:45 am Post subject: quote

Denis is pretty good at "executive summarys" !

However I would like to understand the probability theory of our bias, and its correction factor. The bias wasnt at all obvious to most initially, and the correction factors seemed to go up from 25.2 to 26.5 without explaination ! I have been trying to think of it in terms of a large sack with a large number of different sized balls.

The small factor [?] of puzzle loss when an essential clue in a subgrid was removed how this makes it more likely that some puzzles will be produced preferentially hasnt really been explained either.....

Also its not totally clear to me how the number of minimal puzzles in a subgid of a certain size - reflects the total number of puzzles of a certain size. There are large variations in the number of minimal puzzles in a subgrid - especially more puzzles in subgrids with no "essential" clues.

Allan has automated and analysed the manufacture of large puzzles it would appear, a while back I was interested in this, Max number of clues 2. The fact that there are common clues between puzzles and Dwalk = 2 is effectively a {-2+2} - this was the way **ravel** and **havard** generated the 38s and 39s. Although I believe they used a {-2+3} or a {+3-2} to achieve the addition of a clue in a large minimal puzzle. [no mean feat]

C

Back to top profile pm

denis_berthier

Joined: 19 Jun 2007
Posts: 740

Posted: Sat Jul 18, 2009 12:25 pm Post subject: quote edit

> **Red Ed wrote:**
>
> All this stuff about controlled-bias generators and correction factors is

Posts: 749
Location: Paris, France

> implementation detail: it doesn't make sense to challenge me to say where I introduced the same concepts ... because my implementation was different.

Obviously, you don't know the difference between conceptual and implementation levels. Before being implemented, "controlled-bias generator" and "correction factors" had to be defined in a way indepependent of any implementation.
I could also ask: where did you define any formal model of what you're doing?

As for the maths, I think choose(81, c) is nobody's property. My way of using it is clearly simpler than yours.

I wonder: are you simply angry that I stepped on your hunting grounds without asking permission?

Red Ed wrote:

> you are the only person generating unbiased samples of c-clue puzzles.

Obviously again, you've missed the point. I'm not only generating unbiased samples of c-clue puzzles. What I'm generating is samples of any c-clue size, with known bias between different clue sizes.

Back to top profile pm www

denis_berthier

Joined: 19 Jun 2007
Posts: 749
Location: Paris, France

Posted: Sat Jul 18, 2009 12:56 pm Post subject: quote edit

coloin wrote:

> I would like to understand the probability theory of our bias, and its correction factor.

Are we speaking of the same thing? The correct theory is here:
http://www.sudoku.com/boards/viewtopic.php?t=14615&start=134
It is about controlled-bias generators, not about the standard top-down or bottom-up generators. Contrary to what I thought at the begining of this thread (because my understanding of top-down generators was not correct), you can't apply my correction factors to the usual generators.

coloin wrote:

> The bias wasnt at all obvious to most initially, and the correction factors seemed to go up from 25.2 to 26.5 without explaination !

What kind of additional explanation do you need? I've given the explicit formula allowing to pass from a biased mean (25.4) to an unbiased one (26.5), using a sample from a controlled-bias generator. Mathematically, once the formal model of the generator has been clearly stated, this is just a weighted average: the correction factor cf(n) is the (multiplicative) weight you have to give to minimal puzzles with n clues.

coloin wrote:

> how this makes it more likely that some puzzles will be produced preferentially hasnt really been explained either.....

Wrong: the Pn+1/Pn = (n+1)/(81-n) formula shows that 2 different puzzles with different numbers of clues don't have the same probability of being reached by the controlled-bias generator. If the generator wasn't biased, this shouldn't be the case.
Moreover, the same formula, re-written as Pn = Pn+1 * (81-n)/(n+1) and applied for n < 40 (which is the case for all our minimal puzzles), shows that Pn > Pn+1: i.e. puzzles with fewer clues are more likely to be reached.
P41 = P40
P39 = 42/40 * P40
P38 = 43/39 * P39 ...

It also shows that, should there exist minimal puzzles with more than 40 clues, then the effect on them would be reversed (still for the controlled-bias generator - I don't know for the usual top-down generators).

How does this apply to ordinary top-down generators: as, when they reach a multi-solution puzzle, what they do is backtracking and trying once more to go deeper, instead of just giving up with the current complete grid (as the controlled-bias generator does), it is obvious that they tend to go deeper than the controlled-bias generator and are thus still more biased towards minimals with fewer clues.

Back to top

profile pm www

ronk

Joined: 02 Nov 2005
Posts: 2398
Location: Southeastern USA

Posted: Sat Jul 18, 2009 1:04 pm Post subject: quote

Nice work **Allan**. If I read it 100 more times, maybe I'll understand some of it. 🙂 For this snippet that I do understand ...

Allan Barker wrote:

> **Dwalk.** The average distance traveled to find a minimal. Although I call this Dwalk, this is not a random walk, it's a convergence distance. Example
>
> **Code:**
>
> ```
> .-----------.          .-----------.
> | A . | . . |          | A . | d*. |
> | . B | . F |          | . . | . F |
> |-----------|   --->   |-----------|   Dwalk = 2
> | C D | . . |          | C . | . b*|
> | . . | E . |          | . . | E . |
> .-----------.          .-----------.
> ```
>
> I don't yet have a good handle on how these distances relate to "grid space" but they are very large numbers, as they should be. I'm trying to work that out now.

... I would like to see it considered as Dwalk = 4. Its definition would then be consistent with "Hamming distance" or "distance" term used elsewhere in this forum. It can take on odd values and be measured with **gsf**'s software. [edit: I don't recall if a changed clue at one location is counted as 1 or 2.]

Back to top profile pm

Red Ed

Joined: 06 Jun 2005
Posts: 611

Posted: Sat Jul 18, 2009 1:54 pm Post subject: quote

Denis, calm down.

denis_berthier wrote:

> I wonder: are you simply angry that I stepped on your hunting grounds without asking permission?

No, dear boy, far from it. A little baffled that you refuse to set your work in the context of that which preceded it, that's all.

Back to top profile pm

Red Ed

Joined: 06 Jun 2005
Posts: 611

Posted: Sat Jul 18, 2009 2:15 pm Post subject: quote

A quick side-by-side comparison of the number-of-clues probability distribution estimation methods.

Canonical experiment
Pick a random grid and, within it, a random s-clue subgrid. Find all c-clue minimal puzzles within that subgrid; call that M(t) where t is the trial number.

Theorem
1/t (M(1)+...+M(t)) * choose(81,c)/choose(81,s) is an unbiased estimate of the average number of c-clue minimal puzzles per grid.

Implementation
Denis removes clues at random until a minimal or an improper puzzle is reached. In effect, this performs the canonical experiment, with s=c, for all c at the same time.
Red Ed picks a c in advance, picks s>c to achieve better variance than would be possible with s=c, then performs the canonical experiment.

Pros and Cons
Denis can perform all canonical experiments in parallel, but produces estimators which, since s=c for him, have relatively high variance. At present he throws away t, the number of trials, so can report only the probability distribution, not the estimated actual number of minimal puzzles. Denis' method can also be used to generate puzzles which, at each clue level c, are unbiased.
Red Ed must perform canonical experiments for each* c that he's interested in, but can choose to focus CPU on just the most difficult ones (e.g. 28/29/30) and can tune s>c to give lower variance. Also, he keeps track of t, so can report the number of minimals of each size, not just the probability distribution. Ed can only produce puzzles which, at each clue level c, are unbiased if he sets s=c.
*EDIT: actually the first version of my algorithm did c,c+1,c+2,...,s in parallel, so not so different from Denis.

Shared problems

Neither of us has made much progress on 29- and 30-clue minimals. Perhaps a new idea is needed.

Historical note

Ed got there first 😁

Back to top profile pm

eleven

Joined: 10 Feb 2008
Posts: 364

Posted: Sat Jul 18, 2009 3:00 pm Post subject: quote

coloin wrote:

> However I would like to understand the probability theory of our bias, and its correction factor.

Coloin, from secondary school level to secondary school level 🙂

The controlled bias generator goes down dropping one random clue after the other (always keeping the starting grid as a solution), until it reaches a puzzle, which is either minimal or has multiple solutions.
Now, if you need X tries to find a 24 (then you had found X-1 multisolution and minimal puzzles with less equal 24 clues, which means X-1 24's with multiple and no solutions), you can estimate, that out of all possible 24's in this grid 1/X are minimal. There are choose(81,24) possible 24's to use Red Ed's notation. Thus you can estimate the numbers of all minimal puzzles (you found) with n clues with (found n-clues)*choose(81,n). This is the simple thing, Red Ed told me above.
Now we have (choose(81,25) possible 25's and choose(81,24) 24's. As we know from school, choose(n,k)=[n*(n-1)*...(n-k+1)]/[1*2*...*k], therefore (choose(81,25)/choose(81,24))=57/25. So to get the real number of 25's relative to the 24's, you have to multiply it with 57/25. (all 25's/all 24's = found 25's/found 24's * 57/25)
Thats all you need for Denis' method.

The rest are words and statistics.

Back to top profile pm

denis_berthier

Joined: 19 Jun 2007
Posts: 749
Location: Paris, France

Posted: Sat Jul 18, 2009 3:16 pm Post subject: quote edit

Red Ed wrote:

> Denis, calm down.

Where did you see I'm not calm?

Red Ed wrote:

> A little baffled that you refuse to set your work in the context of that which preceded it, that's all.

I don't. You did a lot before I became interested in puzzle generation, exactly one week ago, after there were persistent rumours of a possible bias in the existing generators and my results in the "rating" thread displayed some small trend for mean greater complexity with higher number of clues - with a very

small correlation coefficient.
In one week, I specified a new type of generator, not unbiased but with known bias, and I developed a formal model for it. It explains in a straightforward way why few-clue puzzles are more likely to be obtained. Eleven implemented it easily by modifying suexg. By using my correction factors, both of us easily reached the conclusion that the mean number of clues is 26.55 (which you had also found previously by your method).
You keep claiming that you've done all this before and better and more and bigger than me. Alright. The fact is, the only thing I need from the generator is puzzles. Contrary to what's suggested by eleven's explanation above, I need access to no internal data of the program (such as his X). And I need no prior knowledge of the number of minimal puzzles with n clues.
So, it may be the case that your approach is more general. But you can't deny that mine is conceptually simpler.

Your "canonical experiment" is yours, not mine. It is more general than what I do. Once again, the fact is, I just have to apply correction factors to the statistics based on puzzles produced by the controlled-bias generator.

Red Ed wrote:

> **Historical note** Ed got there first 😀

Did I ever deny this? This is your hunting grounds. Don't worry, this was only a short incursion and I'll soon be withdrawing from this area: I've now done almost all that I needed to do for my main purposes (classification according to complexity).
And, above all, I'll soon leave for summer holidays 😀

Back to top profile pm www

eleven

Joined: 10 Feb 2008
Posts: 364

Posted: Sat Jul 18, 2009 3:57 pm Post subject: quote

denis_berthier wrote:

> Contrary to what's suggested by eleven's explanation above, I need access to no internal data of the program (such as his X).

This is a misunderstanding. As i showed it, the X falls away, when you devide the "all n" numbers and its obvious you dont need it (if you are not interested in the "all" numbers, but only the distribution).

However, i hope i made clear, that there is much unnecessary debate about a nice idea. I would have wished to get more relevant information to understand it instead of, what we became to read above.

Back to top profile pm

Allan Barker

Joined: 21 Feb 2008
Posts: 294

Posted: Sat Jul 18, 2009 3:59 pm Post subject: quote

denis_berthier wrote:

> [
> Very interesting data and original approach of generation. If you think

Location: Bangkok

> again of the interpretation of distances, let's know. There's probably some interesting theory behind it.

I look at it as an 81 dimension binary space where distances are a bit different. Once I get some numbers to work right I'll post a bit more. I should also look through some of the older threads, a lot was done way back when.

denis_berthier wrote:

> **Allan Barker wrote:**
>
> > If anyone is interested, I can put the first batch of 30s on my website.
>
> I'm interested. Large samples of 30s, 31s, 32s ... are not so frequent.

http://sudokuone.com/xsudo1/puzzle30.txt. These are not compressed, a bit over 20000.
[EDIT: Corrected the file name to puzzle30.txt]

denis_berthier wrote:

> I think your procedure is sufficiently random to allow unbiased computations of the SER and NRCZT for each clue size.

Or maybe for "reasonably accurate" computations, subject to Red Ed's caveats posted later, which are right of course.

denis_berthier wrote:

> If one wants to compute global unbiased distributions of SER and/or NRCZT (the reason why I opened this thread), one needs both the distribution for each number of clues (which are currently missing for large clue sizes) and unbiased distribution of clues (an estimation of which I hope to get with the controlled-bias algorithm).

Yes and no. Given 10000 or 1000000 samples of clue size C that were (assuming) made in a unbiased manner, how can knowledge of the distribution affect their rating? The distribution is only needed for some inter clue size computations or comparisons, it should not affect the intra clue size results. No?

denis_berthier wrote:

> Not the proper thread for this (the "rating" thread would be more appropriate), but can you say more about BITS?

Yes, I'll put something there in a bit.

Red Ed wrote:

> Your step (2) introduces bias for the same reason that the original top-down generator did; but it's easy to overcome, e.g. loop back to step 0 if you get multiple solutions (or, equivalently, generate (grid,mask) pairs until the c-clue subgrid under the mask has a single solution).

Yes, fully agreeded, and I'll try that. It might slow things down but most of the enhancement is the Monte-Carlo part. But then, how biased is that? In principle, I would think not too much, however any process that depends on a path probably has the potential for some kind of bias.

> **Red Ed wrote:**
>
> Step (4) just outputs a single minimal puzzle, the first one found, right?

Right.

> **Red Ed wrote:**
>
> In that case it can't be used as an unbiased source of puzzles because you'd end up producing on average the same number of c-clue puzzles for every solution grid (and we know that different solution grids have different numbers of puzzles).

Hum, yes 1 per grid, so I get your point, the probability that particular solution grids would be represented in the output is now completly washed out. Along the same lines, the generator can also work this way.

0. Choose C = desired clue size.
1. Generate a random 81 candidate grid.
2. Randomly remove clues until clues=C, maintaining single sol. at each step.
4. Maintaining the number of clues = C but allowing the grid to change, Monte Carlo search for minimal puzzle.

I have no idea how biased that is (among puzzles of size C), but choosing an item at random and deciding if it is closer or not might not be too bad.

The term "randomly biased" comes to mind.

Last edited by Allan Barker on Mon Jul 20, 2009 11:20 am; edited 1 time in total

Back to top profile pm www

Red Ed

Posted: Sat Jul 18, 2009 4:10 pm Post subject: quote

Joined: 06 Jun 2005
Posts: 611

> **denis_berthier wrote:**
>
> And, above all, I'll soon leave for summer holidays 😀

A timely reminder that there are better things to do than sudoku! 😀

Back to top profile pm

denis_berthier

Posted: Sat Jul 18, 2009 4:22 pm Post subject: quote edit

Joined: 19 Jun 2007
Posts: 749
Location: Paris, France

> **eleven wrote:**
>
> > **denis_berthier wrote:**
> >
> > Contrary to what's suggested by eleven's explanation above, I need access to no internal data of the program (such as his X).
>
> This is a misunderstanding. As i showed it, the X falls away, when you

> devide the "all n" numbers and its obvious you dont need it (if you are not interested in the "all" numbers, but only the distribution).

I didn't say you were wrong. I just said that, in my approach, taking X into account, in any way, was not necessary.

eleven wrote:

> I would have wished to get more relevant information to understand it

Well then. Why don't you ask about what remains unclear? But please refer to the right place (http://www.sudoku.com/boards/viewtopic.php?t=14615&start=134).
If some points remain obscure, your questions may help me improve the way I've written them.

Back to top profile pm www

denis_berthier

Joined: 19 Jun 2007
Posts: 749
Location: Paris, France

Posted: Sat Jul 18, 2009 4:41 pm Post subject: quote edit

Allan Barker wrote:

> **denis_berthier wrote:**
>
> > **Allan Barker wrote:**
> >
> > > If anyone is interested, I can put the first batch of 30s on my website.
> >
> > I'm interested. Large samples of 30s, 31s, 32s ... are not so frequent.
>
> **http://sudokuone.com/xsudo1/puzzle33.txt**. These are not compressed, a bit over 20000.

Thanks, I don't know when I can study them (all my computing resources are currently centered on the controlled-bias generator), but I'll do it.
BTW, the url is: http://sudokuone.com/xsudo1/puzzle30.txt

Allan Barker wrote:

> **denis_berthier wrote:**
>
> > I think your procedure is sufficiently random to allow unbiased computations of the SER and NRCZT for each clue size.
>
> Or maybe for "reasonably accurate" computations, subject to Red Ed's caveats posted later, which are right of course.

I don't think that Red Ed meant there may be a bias in the SER or NRCZT, once the clue size has been fixed.

> **Allan Barker wrote:**
>
> > **denis_berthier wrote:**
> >
> > If one wants to compute global unbiased distributions of SER and/or NRCZT (the reason why I opened this thread), one needs both the distribution for each number of clues (which are currently missing for large clue sizes) and unbiased distribution of clues (an estimation of which I hope to get with the controlled-bias algorithm).
>
> The distribution is only needed for some inter clue size computations or comparisons, it should not affect the intra clue size results. No?

That's what I said. We can use your samples of c-clue puzzles to make computations of SER and NRCZT restricted to c-clue size.
When it comes to statistics on all the minimal puzzles, we have to combine the results obtained for each c. That's where the estimation of relative frequencies of clue sizes comes into play.

Back to top

profile pm www

Allan Barker

Joined: 21 Feb 2008
Posts: 294
Location: Bangkok

Posted: Sat Jul 18, 2009 5:10 pm   Post subject:

quote

> **ronk wrote:**
>
> Nice work **Allan**. If I read it 100 more times, maybe I'll understand some of it. 🙂 For this snippet that I do understand ...
>
> > **Allan Barker wrote:**
> >
> > **Dwalk.** The average distance traveled to find a minimal. Although I call this Dwalk, this is not a random walk, it's a convergence distance. Example
> >
> > **Code:**
> >
> > ```
> > .------------.        .------------.
> > | A . | . . |         | A . | d*. |
> > | . B | . F |         | . . | . F |
> > |------------|  --->  |------------|   Dwalk =
> > 2
> > | C D | . . |         | C . | . b*|
> > | . . | E . |         | . . | E . |
> > .------------.        .------------.
> > ```
> >
> > I don't yet have a good handle on how these distances relate to "grid space" but they are very large numbers, as they should be. I'm trying to work that out now.
>
> ... I would like to see it considered as Dwalk = 4. Its definition would then be consistent with "Hamming distance" or "distance" term used elsewhere in this forum. It can take on odd values and be measured with **gsf**'s software. [edit: I don't recall if a changed clue at one location is counted as 1 or 2.]

Yes, the true distance would be the number of address bits changed in the 81 dimensional binary space, or 4, which is why I said some of these distances are very large. I've also now been able to relate the path distance to some other numbers like the ratio of puzzles to minimals, which look suprisingly good considering I would expect them to be awful.

Coloin wrote:

> The fact that there are common clues between puzzles and Dwalk = 2 is effectively a {-2+2} - this was the way ravel and havard generated the 38s and 39s. Although I believe they used a {-2+3} or a {+3-2} to achieve the addition of a clue in a large minimal puzzle. [no mean feat]

Definitely similar. The one difference is that the Monte-Carlo is a guided approach based on distance, rather than a random one.

39 ?????? Were they using specially selected grids or random ones?

.

Back to top profile pm www

eleven

Joined: 10 Feb 2008
Posts: 364

Posted: Sat Jul 18, 2009 6:32 pm Post subject: quote

denis_berthier wrote:

> If some points remain obscure, your questions may help me improve the way I've written them.

My problem just was, that i need at least an hour to get out there, what could have been explained to me in my 10 lines.
But thats a problem of notation, and i am aware of this problem from many discussions about sudoku solutions. Its a matter of taste and nothing, what i want to discuss any more.

Allan Barker wrote:

> 4. Maintaining C clues, Monte Carlo search for minimal puzzle.
> 5. Measure **Dwalk**, the number of clues that have changed. (D means distance)

What, if you make n Monte Carlo searches within a fixed Dwalk ?

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

Goto page Previous 1, 2, 3 ... , 10, 11, 12 Next

Page 11 of 12

Stop watching this topic

Jump to: General/puzzle Go