THE REAL DISTRIBUTION OF MINIMAL PUZZLES

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

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

View previous topic :: View next topic

Author	Message
denis_berthier	Posted: Sat Jul 11, 2009 9:05 am Post subject: THE REAL DISTRIBUTION OF MINIMAL PUZZLES quote edit

Joined: 19 Jun 2007
Posts: 745
Location: Paris, France

THE REAL DISTRIBUTION OF MINIMAL PUZZLES

1) Introduction

Yesterday evening, I read that camera lens makers have always been confronted to a series of problems: chromatic aberration, purple fringing, barrel or pincushion distortion, vignetting, ... The classical approach, and the only one feasible with classical cameras, was to improve the lenses - which led to large, complex and expensive lenses. But, with the advent of digital camears, slight problems of the above types (and others) can be accepted in the lens, provided that we can measure them precisely and correct them by the software. This is heresy for traditional lens makers such as Leica, but some camera makers have adopted this solution.

This morning, I woke up with the following idea: we are unable to build unbiased collections of minimal puzzles. What if we take the collections as they are produced by the current generators and apply a correction to the results? Indeed, I was very surprised that the analysis of bias in top-down generators is straightforward and I could program the corrections needed in less than one hour.

2) A forest of puzzles

Consider first the following 80-floor structure (a forest of trees with branches pointing downwards), the nodes being indexed puzzles (which may have one or more solutions).
floor 81 : all the N different complete solution grids (N is huge, but no one is asking you to cross the whole forest), each indexed by the empty sequence; notice that all the puzzles at this floor have 81 clues;
floor 80: each puzzle P at floor 81 sprouts 81 branches pointing to floor 80, one for each clue C in P; the head of this C branch will be the puzzle obtained from P by removing clue C and indexed by the 1-element sequence (C); notice that all the puzzles at floor 80 have 80 clues;

Now, recursively, given floor n (all puzzles of which have n clues and are indexed by sequences of length 81-n), build floor n-1 as follows:
each puzzle P at floor n sprouts n branches pointing to floor n-1, one for each clue C in P; the head of the C branch will be the puzzle obtained from P by removing clue C and indexed by the (81-n)-element sequence obtained by appending C to the end of the sequence indexing P; notice that all the puzzles at floor n-1 have n-1 clues and that the index of a puzzle is the (ordered) sequence of deletions that led to it.

It is easy to see that, at floor n, there are N * 81! / n! indexed puzzles, each of which has its underlying (non indexed) puzzle identical to that of (81 - n)! indexed puzzles at the same floor (including itself).

Along each branch of this forest, there is one and only one (indexed) minimal puzzle. Above it, all the puzzles have redundant clues; below it, all the puzzles have multiple solutions.
Consider now the border B (i.e. the set of indexed minimal puzzles). This border crosses all the branches of our forest at different floors.

If we consider that each puzzle at each floor has the same probability of being reached from the top in (81 - n) steps and these probabilities sum up to 1 at each floor, then any indexed puzzle at floor n has probability 1/ {(N * 81! / n!) of being reached.
And any non-indexed puzzle at floor n has probability Pn = 1/ {(N * 81! / n!) * (81 - n)!} = 1/N * 1/81! * n! * (81 - n)!
[correcetd a typo; thanks to Red Ed; see his next post]

For small n, this is a inimaginably small, umanageable number.
But the ratio Pn+1 / Pn is very simple: **Pn+1 / Pn = (n + 1) / (81 - n).**

From which we can see that the probability induced on B by these probabilities is biased.
As we know this bias, we can correct it easily by applying to the probabilities induced on B correction factors cf(n), whose absolute value is not important, but which satisfy **cf(n+1) / cf(n) = (81 - n) / (n+1)**.

2) Top down generators and the set of paths among the single solution puzzles
Consider now a top-down generator. The first phase is the generation of unbiased complete grids, which we know how to do.
For each initial complete grid, the second phase of the generator consists of following some path donwards in our forest until it reaches an indexed minimal puzzle on the border B, where it stops.
The generator works according to the probabilities on our forest of indexed puzzles instead of having a uniform distribution on minimal puzzles.
But, as we know its bias precisely, it is easy to apply correction factors.

What does this mean in practice?
If you have a variable X, let
on(n) be the observed number of puzzles with n clues,
X(n) be the observed mean value of X for puzzles with n clues.
The raw (biased) mean of X is computed classically as sum(X(n) * on(n)) / sum(on(n))
The corrected (unbiased) mean of X must be computed as: **unbiased-mean(X)**

= sum(X(n) * on(n) * cf(n)) / sum(on(n) * cf(n)).

This formula shows that the cf(n) sequence needs be defined only modulo a multiplicative factor.
For convenience, I chose cf(20) =1 in my computations. This gives the following sequence of correction factors:
cf-sequence[19...31] = 0.333333333333333 1 2.9047619047619 7.92207792207792 20.3218520609825 49.1111424807077 111.973404856014 241.173487382183 491.279326148891 947.467271858576 1731.57811753464 3001.40207039337 4937.79050290523
It may be shocking to consider that some puzzles must be given a weight 3000 times greater than other puzzles (ratio between 30-clue and 20-clue puzzles), but that's how it is.
A consequence of this is that statistics on minimal puzzles must rely on very large samples.

3) Applications

Let's use the raw computations for 2 collections (sudogen0_1M and rabrnd1m) of 1 million puzzles each, generated by 2 top-down generators with very different first phases.
For details on these collections and computations, see the "rating rules / puzzles" thread or my web page:
http://www.carva.org/denis.berthier/HLS/Classification/index.html.
We get the following results.

3.1) The mean number of clues of minimal puzzles = 25.39

sudogen0_1M: raw-average = 24.380591 unbiased-average = 25.3910685435253
rabrnd1M: raw-average = 24.384134 unbiased-average = 25.3920009372608
The 2 collections lead to the same result.

This is 1 more than the raw-average.

3.2) The mean SER of minimal puzzles = 4.06

sudogen0_1M: raw-average = 3.7722223 unbiased-average = 4.06321801478134
rabrnd1M: raw-average = 3.76660230000001 unbiased-average = 4.06047314710775

3.3) The mean NRCZT-rating of minimal puzzles = 2.09

sudogen0_1M: raw-average = 1.94113060000003 unbiased-average = 2.09015177491554
rabrnd1M: data not available

3.4) More generally, one can compute the real distribution of minimal puzzles

It is merely the product of the observed distribution and the correction factors, namely on(n) * cf(n) (normalised, of course, by sum(on(n) * cf(n)).

4) Remarks
In the " how many minimal sudokus has an average grid" thread, another approach has been taken by Red Ed, who tries to estimate the number of minimal puzzles with n clues, a very hard problem. His estimation of the average number of clues is 26.4 (but he mentions that this is must be taken "with a pinch of salt"), much above the 25.39 value computed here..
In the present approach, no estimation of these numers is needed. Only very simple computations lead to the solution.

5) A remark on bottom-up generators
A similar analysis for bottom-up generators is more difficult (but I din't have time to really think of it), because these generators are not purely bottom-up. Staring with 0 clues, they add clues until they reach a single solution puzzle, but after that they delete clues until they reach a minimal puzzle.

Edited 14/17/09: BEWARE:
- the model described here neglects some essential aspects of the workings of a "classical" top-down generator;
- the initial idea of using a biased generator with correction factors remains, but is has to be applied to a modified top-down generator, which I called a controlled-bias generator;
- this post is thus superseded by the new model for these generators, here: http://www.sudoku.com/boards/viewtopic.php?t=14615&start=134.

Last edited by denis_berthier on Sun Jul 19, 2009 9:02 am; edited 4 times in total

Back to top profile pm www

Red Ed

Posted: Sat Jul 11, 2009 10:30 am Post subject: Re: THE REAL DISTRIBUTION OF MINIMAL PUZZLES quote

Joined: 06 Jun 2005
Posts: 608

Thank you for your interest in the distribution of minimal puzzles.

Your first mistake is here:

denis_berthier wrote:

> And any non-indexed puzzle at floor n has probability Pn = 1/ {(N * 81! / n!) * (81 - n)!} = 1/N * 1/81! * n! / (81 - n)!

In fact, Pn = 1/ {(N * 81! / n!) / (81 - n)!} = 1/N * 1/81! * n! * (81 - n)!

Of course, this formula holds only for a sampling process that doesn't care whether the puzzles it finds have single or multiple solutions.

Back to top profile pm

denis_berthier

Posted: Sat Jul 11, 2009 10:37 am Post subject: Re: THE REAL DISTRIBUTION OF MINIMAL PUZZLES quote edit

Joined: 19 Jun 2007

Red Ed wrote:

Posts: 745
Location: Paris, France

> Thank you for your interest in the distribution of minimal puzzles.
> Your first mistake is here:
>
> **denis_berthier wrote:**
>
> > And any non-indexed puzzle at floor n has probability Pn = 1/ {(N * 81! / n!) * (81 - n)!} = 1/N * 1/81! * n! / (81 - n)!
>
> In fact, Pn = 1/ {(N * 81! / n!) / (81 - n)!} = 1/N * 1/81! * n! * (81 - n)!

Yes, you're right, I made a typo when I copied from my sheet of paper.
But the ratios are correct.

> **Red Ed wrote:**
>
> > Of course, this formula holds only for a sampling process that doesn't care whether the puzzles it finds have single or multiple solutions.

As the generator stops when a minimal puzzle is reached, what happens below B is immaterial. But in the forest, yes, puzzles may have several solutions.

Last edited by denis_berthier on Sat Jul 11, 2009 10:43 am; edited 1 time in total

Back to top profile pm www

Red Ed

Posted: Sat Jul 11, 2009 10:42 am Post subject: quote

Thank you for your continued interest.

Joined: 06 Jun 2005
Posts: 608

Your second mistake is to confuse subgrids with (proper, i.e. single-solution) puzzles. Your correction factors apply only to subgrids, not proper puzzles: their use in section 3 (collections of proper puzzles) is thus invalid.

Back to top profile pm

denis_berthier

Posted: Sat Jul 11, 2009 10:45 am Post subject: quote edit

> **Red Ed wrote:**
>
> > Thank you for your continued interest.
> >
> > Your second mistake is to confuse subgrids with (proper, i.e. single-solution) puzzles. Your correction factors apply only to subgrids, not proper puzzles: their use in section 3 (collections of proper puzzles) is thus invalid.

Joined: 19 Jun 2007
Posts: 745
Location: Paris, France

We've been cross-posting. I make no confusion. Answer in my previous post.

Back to top profile pm www

Red Ed

Posted: Sat Jul 11, 2009 10:50 am Post subject: quote

> **denis_berthier wrote:**
>
> > As the generator stops when a minimal puzzle is reached, what happens below B is immaterial.

Joined: 06 Jun 2005
Posts: 608

In that case, your first mistake is in fact to assert that Pn = 1 / { N*choose(81,n) }. There are N*choose(81,n) indexed *subgrids* at level n. There are substantially fewer (indexed) puzzles, and even fewer (indexed) minimal puzzles, at that level. So, to repeat myself: your formulae apply to a subgrid

sampler, not a puzzle sampler.

Back to top profile pm

denis_berthier

Posted: Sat Jul 11, 2009 10:57 am Post subject: quote edit

Joined: 19 Jun 2007
Posts: 745
Location: Paris, France

Red Ed wrote:

> **denis_berthier wrote:**
>
> > As the generator stops when a minimal puzzle is reached, what happens below B is immaterial.
>
> In that case, your first mistake is in fact to assert that Pn = 1 / { N*choose(81,n) }. There are N*choose(81,n) indexed *subgrids* at level n. There are substantially fewer (indexed) puzzles, and even fewer (indexed) minimal puzzles, at that level. So, to repeat myself: your formulae apply to a subgrid sampler, not a puzzle sampler.

You should read the definitions: the forest is a forest of grids with possibly multiple solutions.
But above border B, all the puzzles have only one solution (and below they ahve more).
You're playing on words: you're just naming subgrids what I've named puzzle. I never said that the puzzles in the forest had to be valid (i.e. with exactly one solution).
If that suits you better, replace everywhere "puzzle" by "subgrid".

Back to top profile pm www

Red Ed

Posted: Sat Jul 11, 2009 11:03 am Post subject: quote

Joined: 06 Jun 2005
Posts: 608

Denis wrote:

> X(n) be the observed mean value of X for puzzles with n clues.

Well then, how is X(n) defined for a subgrid of n clues that is not a minimal puzzle?

Back to top profile pm

denis_berthier

Posted: Sat Jul 11, 2009 11:05 am Post subject: quote edit

Joined: 19 Jun 2007
Posts: 745
Location: Paris, France

Red Ed wrote:

> **Denis wrote:**
>
> > X(n) be the observed mean value of X for puzzles with n clues.
>
> Well then, how is X(n) defined for a subgrid of n clues that is not a minimal puzzle?

For both the biased and the unbiased means, we are interested in the values of X(n) only on B. B is a probability space in itself. It carries 2 different probability measures. You can consider cf as their relative density.

Back to top profile pm www

Red Ed

Posted: Sat Jul 11, 2009 11:16 am Post subject: quote

No, Denis. Just no. You have defined correction factors for a sampler on one

Joined: 06 Jun 2005
Posts: 608

set (all indexed subgrids) and then, with a magical wave of the hands, applied them to samplers on a very small subset (indexed minimal puzzles). You can only do your X(n) computations when X(n) is defined over *all* indexed subgrids.

Back to top profile pm

denis_berthier

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

Joined: 19 Jun 2007
Posts: 745
Location: Paris, France

> **Red Ed wrote:**
>
> No, Denis. Just ... no. You have defined correction factors for a sampler on one set (all indexed subgrids) and then, with a magical wave of the hands, applied them to samplers on a completely different set (minimal puzzles). The argument does not carry over.

The correction factors define the relative density of the two measures on B - nothing more, nothing less.

Back to top profile pm www

Red Ed

Posted: Sat Jul 11, 2009 11:27 am Post subject: quote

Joined: 06 Jun 2005
Posts: 608

> **denis_berthier wrote:**
>
> The correction factors define the relative density of the two measures on B - nothing more, nothing less.

No, they define the "relative density" of n and n+1 clue objects in the set of all indexed subgrids. It's an unjustified leap of faith to assume that the same is true in the subset B. Are there only 65/17 times more (indexed) 17-clue puzzles than (indexed) 16-clue puzzles? No, of course not.

Back to top profile pm

eleven

Posted: Sat Jul 11, 2009 1:03 pm Post subject: quote

Joined: 10 Feb 2008
Posts: 364

What about bottom up generation without dropping clues ?

If you have a table, which shows, how often a puzzle with n clues was tried, how many of those puzzles were minimal, unique, with multi solutions and with no solution, can you calculate then the bias from it ?

Back to top profile pm

Red Ed

Posted: Sat Jul 11, 2009 1:17 pm Post subject: quote

Joined: 06 Jun 2005
Posts: 608

I don't trust bottom-up generators: they don't generate clue structures that mimic those in ordinary grids (as evidenced by the 0.5-clue difference in minimal puzzle size as compared to top-down generators).

Back to top profile pm

eleven

Posted: Sat Jul 11, 2009 1:47 pm Post subject: quote

Joined: 10 Feb 2008
Posts: 364

I see, though without dropping clues i got an average of 24.065 for about 80000 puzzles, so just the half difference.

Back to top profile pm

Display posts from previous: All Posts Oldest First Go


Sudoku Players' Forums
Forum Index ->
General/puzzle

All times are GMT
Goto page 1, 2, 3 ... 10, 11, 12 Next

Page 1 of 12

Stop watching this topic

Jump to: General/puzzle Go

You **can** post new topics in this forum
You **can** reply to topics in this forum
You **can** edit your posts in this forum
You **can** delete your posts in this forum
You **can** vote in polls in this forum