NEW RESULTS IN DEEP-SEARCH BEHAVIOUR

J. Renze Steenhuisen[1]

Delft, The Netherlands

ABSTRACT

This article is a follow-up on previous work done on deep-search behaviour of chess programs. The program CRAFTY was used to repeat the go-deep experiment on positions taken from previous experiments to push the search horizon to 20 plies. The same experimental setup was used to search, among others, a set of 4,500 positions, from the opening phase, to a depth of 18 plies. Our results showed that the chance of new best moves being discovered decreases exponentially when searching to higher depths, and decreases faster for positions closer to the end of the game. This contribution brings the understanding of deep-search behaviour and the prediction of performance a step further to completion.

1. INTRODUCTION

Searching is elementary to most game-playing programs, and it is generally believed that searching more deeply achieves ever stronger play. However, so far the precise relation between search depth and playing strength is not known, despite the rich history of research on this topic. Previous research on deep-search behaviour can be divided into two approaches: *self play* and *go deep*.

In self-play experiments, two otherwise identical programs are matched with one having a handicap in search depth or search time. Thompson (1982; 1983) introduced this type of experiment to computer chess in the early 1980s to measure the performance gain of searching more deeply. Using this method, he reported that the playing strength of BELLE increased almost linearly with search depth by 200-250 ELO points. These results have been confirmed by other researchers using different programs (Schaeffer, 1986; Szabo and Szabo, 1988; Berliner *et al.*, 1990; Mysliwietz, 1994; Junghanns *et al.*, 1997). Self-play experiments are also conducted in other games, such as Lines of Action (Billings and Björnsson, 2003).

Yet, intuition suggests that searching to increased depths should eventually lead to diminishing returns. In self-play experiments, diminishing returns should be visible by lower scoring rates at higher search depths. However, most self-play experiments do not quantify the differences in playing strength with high statistical confidence (Mysliwietz, 1994; Heinz, 1999b; Heinz, 2000b). Heinz (2000a) estimated that at least 1,000 games are needed per match to assess diminishing returns in self-play with 95% statistical confidence. Heinz (2001) conducted a large-scale self-play experiment with 3,000 games per match using FRITZ 6 at fixed iteration depths of 5 to 12 plies. The results proved the existence of diminishing returns for additional search in computer chess with 95% statistical confidence.

Go-deep experiments consider the best moves from different iteration depths of a set of positions. Newborn (1985) discovered that the results of self-play experiments are closely correlated with the rate at which the best move changes from one iteration to the next. Based on the self-play experiments with BELLE (Condon and Thompson, 1983) and TECH III (Szabo, 1984), Newborn formulated a hypothesis. Let $RI(d+1)$ denote the rating improvement when increasing search depth from level d to level $d+1$, and $BC(d)$ the expectation of finding a best move at level d different from the best move found at level $d-1$, then:

$$RI(d+1) = \frac{BC(d+1)}{BC(d)} \cdot RI(d) \tag{1}$$

[1]Department of Computer Science, Delft University of Technology, The Netherlands, Email: J.R.Steenhuisen@tudelft.nl

The go-deep experiment was introduced for determining $BC(d)$ for higher values of d. In 1997, PHOENIX (Schaeffer, 1986) and THE TURK (Junghanns *et al.*, 1997) were used to record best-move changes at iteration depths up to 9 plies. In the same year, Hyatt and Newborn (1997) let CRAFTY search to an iteration depth of 14 plies. Heinz (1998) repeated their go-deep experiment with DARKTHOUGHT and reported similar results.

He let $FB(d)$ (fresh best) denote the expectation that a best move has not been considered best at any iteration previous to d, and suggests the following approximation:

$$\frac{BC(d+1)}{BC(d)} \approx \frac{FB(d+1)}{FB(d)} \text{ holds for } d \geq 7. \tag{2}$$

Although much research has been conducted on deep-search behaviour, many open questions remain. In the remainder of this article, we will address some of the open questions. Before elaborating on the details, we start by briefly describing the questions and results achieved so far in the paragraphs below.

Hyatt and Newborn (1997) suggest that $BC(d)$ stabilises at 15 to 17% at higher search depths. Heinz (1998) questions this suggestion because the results of DARKTHOUGHT show a drop at the end of the data curve. However, these conclusions are based on results of searching up until a depth of 14 plies. In Section 2, we present new go-deep results with CRAFTY as a sequel to the go-deep experiments by Hyatt and Newborn (1997) and by Heinz (1998), extending the search depth to 20 plies. Based on the new results, we show that $BC(d)$ does not stabilise but continues to decrease beyond 14 plies.

Although statistical significance has been discussed thoroughly for self-play experiments (Haworth, 2003; Heinz, 1998; Heinz, 1999b; Heinz, 2001; Heinz, 2003), this has been given little attention in go-deep experiments. In Section 3, a formula is provided for determining lower and upper bounds on the different rates (e.g., best change, fresh best) for any %-level of statistical confidence. Using this formula, we show that it is worthwhile to increase the number of positions from near 350 to near 5,000 for reducing the size of the confidence window.

Previous work in deep-search behaviour has been done on a variety of sets of positions. However, it remains to be proven whether results from different test sets can be compared. In Section 4, we compare the $BC(d)$ results of different sets of test positions. Our results show that significant different results are achieved with different sets. Furthermore, the results indicate that the best-change rates are smaller for positions closer to the endgame.

One step towards finding a relation between search depth and playing strength is to model the deep-search behaviour itself. Therefore, Heinz (1999a) applied the least squares method to fitting several functions on the best-change data of CRAFTY (Hyatt and Newborn, 1997) and DARKTHOUGHT (Heinz, 1998). The results suggest that the go-deep behaviour is best modelled by the piece-wise constant/linear model, instead of the more intuitive exponential model. In Section 5, we revisit the modelling of the deep-search behaviour. Based on new data with tighter bounds, we show that the intuitive exponential model fits the data very well.

Based on the data from CRAFTY and DARKTHOUGHT, Heinz (1998) suggests that Equation 2 is a valid approximation. However, he just mentioned it briefly without providing any proof. In Section 6, an attempt is made to verify Heinz' approximation. Using the bootstrap technique from statistics, we provide a first analysis of the correctness of Heinz' approximation and show that verifying it is much harder than it looks.

2. CRAFTY GOES DEEPER

As a sequel to the initial go-deep experiment by Hyatt and Newborn (1997), we let CRAFTY search the same set of positions to an iteration depth of up to 20 plies. A detailed description of the experimental setup can be found in Appendix A. The 347 original test positions were obtained through ftp://ftp.cis.uab.edu/pub/hyatt/plytest/positions.gz, and corrected according to Heinz (1998). After correction, a total of 343 positions remained.

In order to formalise the measurements done on the deep-search behaviour of chess programs, we use the same definitions as provided by Heinz (1998). Let $B(d)$ denote the best move after iteration d, we then define the following best-move properties[2]:

[2]The factors (d-2) best and (d-3) best are included for comparison of our results to other publications only.

- **Best Change** $B(d) \neq B(d-1)$
- **Fresh Best** $B(d) \neq B(j), \forall j < d$
- **(d-2) Best** $B(d) = B(d-2)$ and $B(d) \neq B(d-1)$
- **(d-3) Best** $B(d) = B(d-3)$ and $B(d) \neq B(d-2)$ and $B(d) \neq B(d-1)$

We note that the fresh-best definition does not consider best moves to be fresh best when this move has already been best at some previous iteration. However, it might be a good idea to consider best moves to be fresh-best moves when they are best for a "fresh reason", even when it was the best move in some previous iteration.

Some positions were not able to complete all iterations up to a search depth of 20 plies (e.g., the interval between power shortages and system reboots was too small and no checkpointing was used). The original test set was a collection of 106 opening positions, 107 middle-game positions, and 130 end-game positions. One of these positions contained only one legal move, and therefore has neither a best change nor any of the other properties. An additional eleven positions are mate-in-X positions, which have no chance of having a best change when searching deeper than the mating line. The estimation of, for instance, the best change is based on the actual observation of its occurrence. Therefore, only the completed iterations have been taken into account, and iterations deeper than the mating lines in mate-in-X positions have not been taken into account[3]. In Table 1, the new results of CRAFTY going deep on all 343 corrected original test positions are summarised. Every column states the estimated probabilities (in %) and their estimated standard errors $SE = \sqrt{BC(d) \cdot (1 - BC(d))/(n_d - 1)}$ (n_d being the number of observations at search depth d). These results resemble the results of the previous go-deep experiments closely. Note that the rates for fresh best, $(d-2)$ best, and $(d-3)$ best are conditional to the occurrence of a best change.

Search Depth	Best Change in % (SE)	Fresh Best in % (SE)	$(d-2)$ Best in % (SE)	$(d-3)$ Best in % (SE)
2	40.52 (2.65)	100.00 (0.00)	-	-
3	38.48 (2.63)	76.52 (3.70)	23.48 (3.70)	-
4	28.57 (2.44)	62.24 (4.92)	28.57 (4.59)	9.18 (2.93)
5	30.70 (2.50)	59.05 (4.82)	22.86 (4.12)	8.57 (2.74)
6	29.91 (2.48)	52.94 (4.97)	30.39 (4.58)	6.86 (2.52)
7	26.69 (2.40)	49.45 (5.27)	34.07 (5.00)	5.49 (2.40)
8	28.15 (2.44)	41.67 (5.06)	33.33 (4.84)	10.42 (3.13)
9	20.23 (2.18)	49.28 (6.06)	26.09 (5.33)	5.80 (2.83)
10	19.06 (2.13)	38.46 (6.08)	26.15 (5.49)	12.31 (4.11)
11	17.89 (2.08)	42.62 (6.38)	22.95 (5.43)	8.20 (3.54)
12	16.47 (2.01)	17.86 (5.16)	33.93 (6.38)	14.29 (4.72)
13	12.94 (1.82)	29.55 (6.96)	27.27 (6.79)	9.09 (4.38)
14	13.02 (1.83)	40.91 (7.50)	18.18 (5.88)	0.00 (0.00)
15	12.72 (1.82)	32.56 (7.23)	20.93 (6.28)	9.30 (4.48)
16	12.20 (1.79)	29.27 (7.19)	19.51 (6.27)	14.63 (5.59)
17	10.45 (1.67)	31.43 (7.96)	20.00 (6.86)	5.71 (3.98)
18	11.14 (1.73)	29.73 (7.62)	16.22 (6.14)	13.51 (5.70)
19	9.06 (1.67)	48.15 (9.80)	14.81 (6.97)	11.11 (6.16)
20	7.34 (1.62)	21.05 (9.61)	31.58 (10.96)	5.26 (5.26)

Table 1: Results of CRAFTY (2004) for all 343 corrected original test positions.

The second column of Table 1 shows the best-change rates ($BC(d)$) at search depths (d) up to 20 plies for all 343 corrected original test positions. A noticeable difference with CRAFTY's previous go-deep results is that the drop below 20% (although arbitrarily chosen) occurs one ply later (cf. Hyatt and Newborn, 1997). Moreover, the results show that the best-change rate decreases at high search depths.

The fresh-best rate shows directionless wavering between 20 to 50% from a search depth of 7 plies onwards, which is similar to the previous results with CRAFTY and DARKTHOUGHT. The observed fresh-best rate at ply 12 seems to be an exception. It is remarkable that the rate at which fresh-best moves are discovered remains high, even at search depths of up to 20 plies. The provided estimated probabilities are conditional to the occurrence

[3]Iterations beyond the mating line should not be taken into account, because it is not part of the probability we are interested in. In other words, taking these additional iterations into account is faking statistics (cf. Hyatt and Newborn, 1997).

of a best change. Therefore, the number of observations is reduced proportional to the number of occurring best changes, which has its influence on the standard error.

3. CONFIDENCE BOUNDS

Statistical significance has been discussed thoroughly for self-play experiments in computer chess (Haworth, 2003; Heinz, 1998; Heinz, 1999b; Heinz, 2001; Heinz, 2003). In go-deep experiments we would like to provide confidence bounds on the values for best-change rate (as has partially been done by Heinz (1998)), fresh-best rate, $(d-2)$-best rate, and $(d-3)$-best rate. The problem of providing lower and upper bounds for these rates is a typical problem in standard statistics, where such a rate is called a binary-valued random variable.

Determining confidence bounds on a binary-valued random variable based on m successes in a sample size of n observations is done by modelling the counting of successes as a binomial distribution. For sufficiently large values n and m the binomial distribution can be approximated by a normal distribution. Lower and upper bounds on the success rate of a binary-valued random variable can be provided for arbitrary %-level of statistical confidence. Provided n samples and m successes, as described above, and a value λ associated with the single-sided fail rate of the $\mathcal{N}(0,1)$ normal distribution, the lower and upper bounds on the success rate for a given %-level of statistical confidence are given by:

$$\frac{m + \frac{\lambda^2}{2} \pm \lambda \cdot \sqrt{m \cdot (1 - \frac{m}{n}) + \frac{\lambda^2}{4}}}{n + \lambda^2} \tag{3}$$

Throughout this article we will use 95%-level of confidence ($\lambda = 1.96$) to provide lower and upper bounds on the presented data. Tighter lower and upper bounds on the estimated value are found for larger numbers of n. Note that the confidence bounds of the best-change rate tighten faster than those of the fresh-best rate related to the number of positions that have been searched. This is due to the precondition of an occurring best change before a fresh best can be observed.

As an example, consider the best-change rate to be estimated at 20% at a certain search depth. The resulting 95%-confidence bounds for samples of 350 and 5,000 positions are $[16.15, 24.51]$ and $[18.91, 21.13]$, respectively. For the fresh-best rate of the same positions the number of samples is reduced to the number of positions in which a best change occurred (i.e., 70 and 1,000, respectively). Considering an estimated fresh-best rate of 30%, the resulting bounds become $[20.54, 41.54]$ and $[27.24, 32.91]$. It is clear from this example that increasing the number of positions from 350 to 5,000 reduces the confidence window to a quarter of its size.

4. DIFFERENT TEST SETS

While a diversity of positions is used as a starting position in self-play experiments, the number of the original set of go-deep positions is rather limited. The test set originally used by Hyatt and Newborn (1997) consisted of 106 opening, 107 middle-game, and 130 end-game positions. Although using such a collection gives insight in the go-deep behaviour in chess in general, we are interested in the difference in behaviour in different game phases. Newborn (1985) already noticed that the best-change rates of the *Win at Chess* (Reinfeld, 1958) positions are lower than those of the positions he selected from the *Encyclopedia of Chess Openings* (Matanović, 1974-1979). We are interested in quantifying the change in the best-change rates when using different sets of test positions, taken from different phases of the game. To this extend, we let CRAFTY go deep on 4,500 of the *Encyclopedia of Chess Openings* (ECO) positions (18 plies), the 300 *Win at Chess* (WAC) positions (20 plies), and 939 of the *1001 Best Ways to Checkmate* (BWTC) positions (Reinfeld, 1971) (20 plies). The BWTC set features 62 mate-in-1 positions, which were removed from the data because best-move changes cannot be observed in these positions. In Figure 1, the results are depicted of letting CRAFTY go deep on these test positions together with the new results of the original test positions (Crafty-343) for comparison. The observed best-change rates with the ECO, WAC, and BWTC sets are tabulated in Appendix B.

A few observations can be made from Figure 1. First, the lines of ECO and Crafty-343 closely resemble each other, as do the lines of WAC and BWTC. The latter can be explained because both test sets mainly consist of mate-in-X positions. Apparently, the opening positions from the ECO set and the collection of positions from

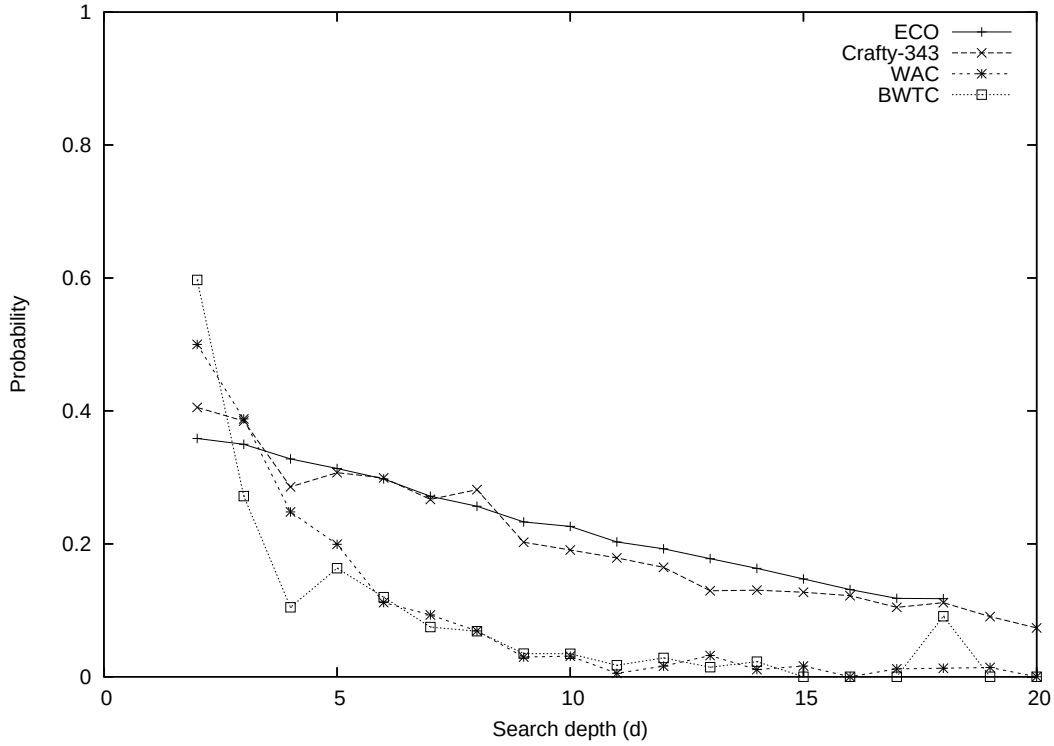

Figure 1: Go Deep results of CRAFTY on three different sets of test positions.

different phases closely resemble each other. Second, the best-change rates at $d = 2$ largely differ from each other while generally decreasing for growing values of d. Third, the small-to-large order of the best-rates of the different test sets seems to reverse when comparing small d to large values of d.

Figure 1 confirms the findings of Newborn (1985), that the best-change rates in the set of WAC positions are smaller than those in the ECO positions. However, we are interested in the difference in go-deep behaviour of mate-in-X positions, because these are well-defined groups. We expect that best-change rates are larger for larger values of X. We conducted six experiments with mate-in-X positions (for $2 \leq X \leq 7$). Because these positions all feature mating sequences, the number of plies in which a best-move change can occur is limited. After a shortest mate sequence is found, the program will not have a change of mind. Therefore, the best-change rate will remain zero (fresh best, $(d-2)$ best, and $(d-3)$ best are undefined). In Figure 2, the best-change rates are plotted for six sets of mate-in-X problems each containing at least 500 problem instances in comparison to the Crafty-343 data. A complete overview of all data can be found in Appendix B.

It was expected that best-change rates of mate-in-X should eventually be smaller than those of mate-in-(X+1), which seems to be the case except for mate-in-3 at $d = 5$. From this figure, we observe that the best-change rate decreases to the point where it intersects with the horizontal axis (equals zero) because the mating line is found. Furthermore, we see that the mate-in-X lines for increasing X intersect with the horizontal axis at increasing depths. Some questions remain, such as whether and at which point the 'true' best-change rate intersects with the horizontal axis. Moreover, what is the reason for a faster decreasing best-change rate?

In Table 2, the results are listed of letting CRAFTY go deep on the 4,500 positions from the ECO set. These results provide tighter confidence bounds than previous results (cf. SE values of Table 1). From these results we arrived at three tentative conclusions. First, instead of wavering directionless in the range 30 to 50% (Heinz, 1998), the fresh-best rate seems to stabilise at 25 to 30%. Second, our results confirm the findings by Heinz (1998), that the $(d-2)$-best rates are wavering in the narrow range of 25 to 35%. The average of the $(d-2)$-best rates in the ECO set is 31.1%[4], which suggests that modern chess programs feature odd-even instabilities for approximately 30% (cf. Heinz, 1998). Third, we conclude that the sum of the fresh-best and $(d-2)$-best rates does not seem to

[4]Heinz (1998) incorrectly reports that the averages of the $(d-2)$-best rates of CRAFTY and DARKTHOUGHT are 26.5% and 24.4%, respectively. This should have been 28.7% and 26.4%, because the value at depth 2 should not be taken into account.

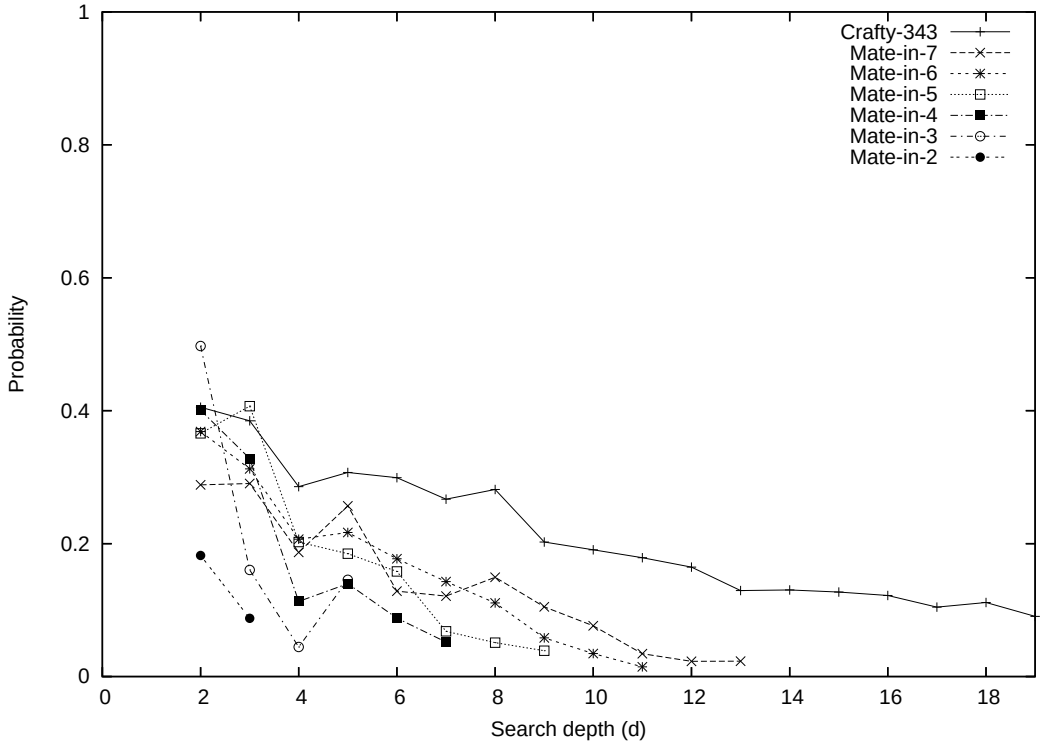

Figure 2: Best-change rates of mate-in-X positions.

Search Depth	Best Change in % (SE)	Fresh Best in % (SE)	$(d-2)$ Best in % (SE)	$(d-3)$ Best in % (SE)
2	35.84 (0.55)	100.00 (0.00)	-	-
3	34.97 (0.55)	80.31 (0.77)	19.69 (0.77)	-
4	32.78 (0.54)	61.46 (0.97)	29.11 (0.91)	9.42 (0.58)
5	31.33 (0.53)	57.55 (1.01)	30.45 (0.94)	6.52 (0.50)
6	29.76 (0.52)	48.52 (1.05)	30.65 (0.97)	9.42 (0.61)
7	27.15 (0.51)	46.33 (1.10)	32.87 (1.03)	7.96 (0.59)
8	25.66 (0.50)	42.08 (1.12)	31.82 (1.05)	9.81 (0.67)
9	23.29 (0.48)	38.10 (1.15)	33.88 (1.12)	9.51 (0.70)
10	22.61 (0.48)	32.12 (1.12)	35.42 (1.15)	9.74 (0.71)
11	20.29 (0.46)	33.46 (1.20)	34.88 (1.21)	9.04 (0.73)
12	19.24 (0.45)	30.18 (1.20)	32.77 (1.23)	9.33 (0.76)
13	17.76 (0.44)	31.22 (1.26)	32.92 (1.28)	8.49 (0.76)
14	16.30 (0.42)	27.25 (1.26)	33.12 (1.33)	10.37 (0.86)
15	14.72 (0.41)	26.27 (1.31)	32.68 (1.40)	10.60 (0.92)
16	13.11 (0.50)	27.97 (1.84)	31.16 (1.90)	10.39 (1.25)
17	11.79 (0.48)	26.68 (1.91)	30.22 (1.99)	12.31 (1.42)
18	11.74 (0.48)	29.46 (1.98)	26.27 (1.91)	9.38 (1.26)

Table 2: Results of CRAFTY (2004) for the 4,500 positions from the ECO test set.

be bounded by the range 65 to 75% but decreases steadilly[5] (cf. Heinz, 1998).

5. MODELLING DEEP-SEARCH BEHAVIOUR

In Sections 2 and 4, new data have been presented on the go-deep behaviour of CRAFTY. Heinz (1999a) modelled the observed best-change behaviour of CRAFTY and DARKTHOUGHT in the first 14 plies of the 343 corrected original test positions. He found that piece-wise constant/linear models provide far better interpolations for both programs than exponential models.

In this section, we use the new best-change data from letting CRAFTY go deep on the 4,500 positions from the *Encyclopedia of Chess Openings* (see Table 2). These new data are better in both quantity (18 instead of 14 plies) and quality (4,500 instead of 343 positions), and should therefore result in a better model for the go-deep behaviour of CRAFTY.

In the previous section, we saw that the decrease of the best-change rate is called to a halt at a certain point. This implies that the function fitting the data best might intersect with the horizontal axis, from which point on the best change should be zero. Therefore, if $BC_i(d)$ fits the data best and intersects with the horizontal axis, then $max(BC_i(d), 0)$ models the data even better. The experimental data will be fitted to three different functions.

1. $BC_1(d) = a \cdot b^d$
2. $BC_2(d) = a \cdot b^d + c$
3. $BC_3(d) = a \cdot d + b$

These functions take depth d as an input parameter and return an approximation of the best-change rate. Parameters a, b, and c denote free variables, which will be determined by iteratively minimising the sum of squared errors (least-squares fit). The program GNUPLOT, freely available through http://www.gnuplot.info/, was used for applying the least-squares fit. The following values were found for the free variables.

1. $BC_1(d) = 43.3875 \cdot 0.932613^d$
2. $BC_2(d) = 74.3291 \cdot 0.971886^d - 33.5908$
3. $BC_3(d) = -1.60289 \cdot d + 38.8725$

Using these functions, we determined the interpolated values and the squared error with the observed best-change rates. These are displayed in Table 3.

The linear model fits the data better than the plain exponential model, based on the least-square error. Clearly, the exponential function with a vertical offset (i.e., $BC_2(d)$) has the smallest summed squared error and therefore fits the data best[6]. However, this is as expected, based on the additional free variable.

Previous attempts at modelling the go-deep behaviour assumed linearity (Newborn, 1985) or concluded that piece-wise constant/linear models (the model is partitioned into multiple functions for different ranges) fitted the data best (Heinz, 1999a). We do not try to model our data with a piece-wise constant/linear model because we deem it to be a mere approximation of the true model and to have little predictive value. Furthermore, we note that the choice on the number of pieces in such a model is rather arbitrary. After redoing the modelling done by Heinz (1999a), we found it remarkable that only the plain exponential model ($BC_1(d)$) was tested, because it restricts the model to go to zero in the limit while he concludes that it should actually be a positive term. While the linear model is able to model every possible line (except a vertical line), the plain exponential model is restricted. On the data of CRAFTY used by Heinz (1999a), that model leads to a summed squared error of 20.40, while our $BC_2(d)$ only has an error of 15.82. The conclusions would have been the same, since our $BC_2(d)$ gets a positive c term for both the used data of CRAFTY and DARKTHOUGHT. Note that this c term has a large negative value after fitting against our new data of CRAFTY.

An interesting observation is that the best fitted model predicts that the intersection with the horizontal axis is near $d = 28$.

[5] Note that this seems to be in conflict with the previous two observations. Further research is needed on this part.

[6] Although not shown, this was the case in Heinz (1999a) as well when using our three models.

Search Depth	BC(d) Data in % (SE)	$BC_1(d)$	$BC_2(d)$	$BC_3(d)$
2	35.84 (0.55)	1.90	0.78	-0.17
3	34.97 (0.55)	0.22	-0.33	-0.91
4	32.78 (0.54)	0.04	-0.05	-0.32
5	31.33 (0.53)	-0.72	-0.47	-0.47
6	29.76 (0.52)	-1.21	-0.71	-0.50
7	27.15 (0.51)	-0.53	0.14	0.50
8	25.66 (0.50)	-0.83	-0.08	0.39
9	23.29 (0.48)	-0.13	0.62	1.16
10	22.61 (0.48)	-1.01	-0.31	0.23
11	20.29 (0.46)	-0.15	0.44	0.95
12	19.24 (0.45)	-0.46	-0.04	0.40
13	17.76 (0.44)	-0.24	-0.05	0.27
14	16.30 (0.42)	0.04	-0.03	0.13
15	14.72 (0.41)	0.52	0.15	0.11
16	13.11 (0.50)	1.10	0.40	0.12
17	11.79 (0.48)	1.46	0.39	-0.17
18	11.74 (0.48)	0.62	-0.84	-1.72
$\sum_2^{18} \Delta(d)^2$	-	11.94	3.19	7.39

Table 3: Interpolation errors of least-squares fits of functions to best-change data of CRAFTY.

6. HEINZ' APPROXIMATION

As described in Section 1, Heinz (1998) had strong empirical evidence that the ratio of two succeeding iterations of the best-change and the fresh-best rate are approximately equal (see Equation 2). Unfortunately, this strong empirical evidence was neither provided explicitly, nor has anything been said about the level of statistical confidence. This section makes a start at verifying the approximation.

As mentioned in Section 3, determining the confidence bounds for the best-change rate is much easier than for the fresh-best rate. However, based on our experiments we are able to determine both the average and its confidence bounds for any %-level of statistical certainty. Dividing two binary-valued random variables (e.g., estimates for BC($d+1$) and BC(d)) results in a new random variable with unknown %-level confidence bounds. Confidence bounds on these rates can be found by using the bootstrap technique from statistics. This technique generates new data from the observations by resampling with replacement from the original sample. For instance, determining the average of two succeeding best-change rates is done by dividing the new found BC($d+1$) and BC(d) after applying the bootstrap procedure. Repeating this process many times provides us with the distribution around the average. The %-level confidence bounds can be estimated from this distribution. These bounds are more accurate when the bootstrap and division procedure are repeated more often.

For gaining insight into the validity of Heinz' approximation, we use the data from the ECO go-deep experiment. After generating 10,000 bootstrap samples, the mean and their respective 95%-confidence intervals are as depicted in Figure 3.

From this figure, it is evident that the ratio of best-change rates shows a much stabler behaviour than that of the fresh-best rate. Furthermore, the ratio of best-change rates seems to have a downwards trend while the ratio of fresh-best rates seems to go upwards. Nevertheless, it may not be concluded that Heinz' approximation does not hold because the confidence intervals have too much overlap. More data are needed to reduce this uncertainty.

Finally, we would like to relate these results to the obtained results with FRITZ 6 by Heinz (2001) in order to shed some light on Newborn's Hypothesis (see Equation 1). Assuming FRITZ 6 and CRAFTY have comparable rating improvements[7], the data from Table 2 in Heinz (2001) can be used to calculate the ratio of succeeding rating improvements for different iteration depths (see Figure 3). Note that half of the rating-improvement ratios

[7]This is a very bad assumption because we know from the literature that programs with less knowledge benefi t more from deeper search than their more knowledgeable counterparts.

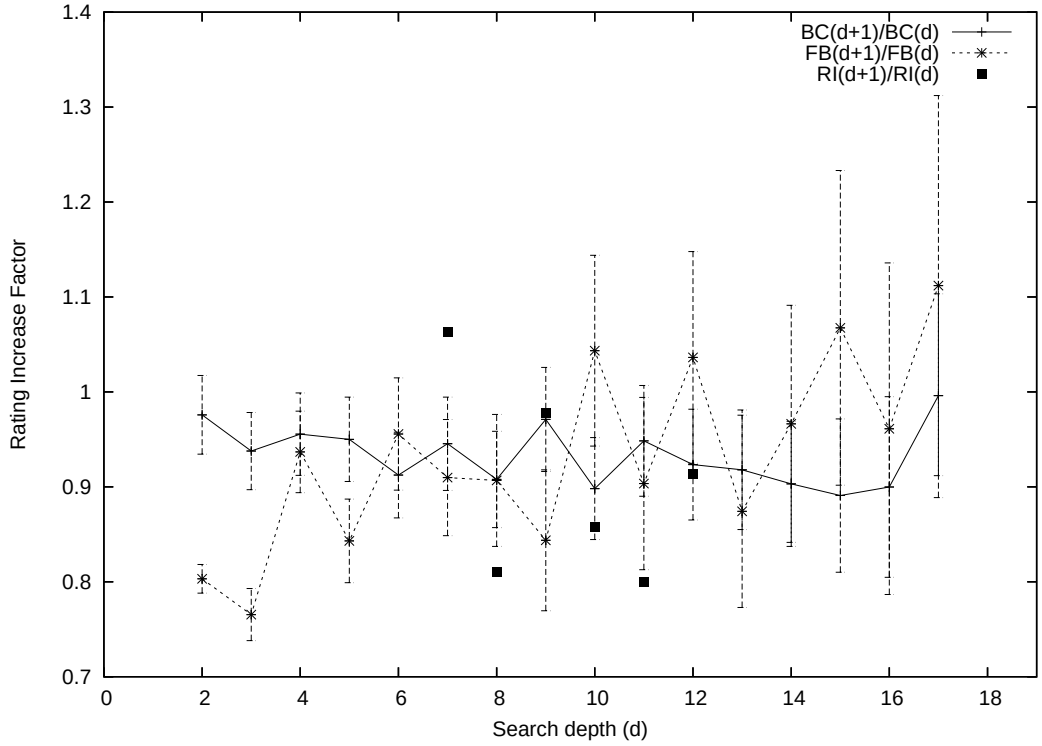

Figure 3: Heinz' approximation for the rating increase factor.

are outside the 95%-confidence interval of the best-change ratios, and are always outside the 95%-confidence interval of the fresh-best ratios. A large-scale self-play experiment has to be conducted in order to relate these ratios to the improvement ratio, with the same settings as used in our go-deep experiment (see Appendix A).

7. CONCLUSIONS

New results have been presented of letting CRAFTY go deep on 4,500 positions from the Encyclopedia of Chess Openings. These results quantify the deep-search behaviour with a high level of statistical confidence. They show that the best-change rate continues to decrease with increasing search depth, instead of stabilising as was suggested by others (Hyatt and Newborn, 1997; Heinz, 1998; Heinz, 1999a). However, the results also show that the fresh-best rate seems to stabilise at 25 to 30%, instead of wavering directionless as was suggested by Heinz (1998). The results further show that CRAFTY suffers from an odd-even instability of approximately 30%, which is even higher than the estimation of Heinz (1998).

The speed with which the best-change rate decreases depends on the test set used. This is likely based on the game phase of the positions in the set. Positions from the opening phase result in relative large best-change rates and a slow decrease, while in sets of end-game positions fast decreases were observed.

Based on the new deep-search results, we verified the results of modelling the best-change behaviour of CRAFTY as obtained by Heinz (1999a). It turns out that the exponential function with a vertical offset is the best model for fitting the experimental data.

Finally, the question was addressed whether Heinz' approximation for the rating-improvement factor of Newborn's hypothesis is valid. First, it was noted that the best-change rates have tighter confidence bounds than the fresh-best rates. Second, it was not possible to conclude whether Heinz' approximation holds based on the obtained results. However, comparison to published results on the rating improvement of FRITZ 6 suggests that these are not estimated correctly by the ratio of fresh-best rates.

8. REFERENCES

Berliner, H. J., Goetsch, G., Campbell, M. S., and Ebeling, C. (1990). Measuring the Performance Potential of Chess Programs. *Artificial Intelligence*, Vol. 43, No. 1, pp. 7–20. ISSN 0004–3702.

Billings, D. and Björnsson, Y. (2003). Search and Knowledge in Lines of Action. *Advances in Computer Games: Many Games, Many Challenges* (eds. H. J. van den Herik, H. Iida, and E. A. Heinz), pp. 231–248, Kluwer Academic Publishers. ISBN 1–4020–7709–2.

Condon, J. H. and Thompson, K. L. (1983). BELLE. *Chess Skill in Man and Machine* (ed. P. W. Frey), pp. 201–210, Springer-Verlag, New York, N.Y. ISBN 0–387–90790–4/3–540–90790–4.

Haworth, G. M. (2003). Note: Self-Play: Statistical Significance. *ICGA Journal*, Vol. 26, No. 2, pp. 115–118. ISSN 1389–6911.

Heinz, E. A. (1998). DARKTHOUGHT Goes Deep. *ICCA Journal*, Vol. 21, No. 4, pp. 228–244. ISSN 0920–234X.

Heinz, E. A. (1999a). Modelling the "Go Deep" Behaviour of CRAFTY and DARKTHOUGHT. *Advances in Computer Chess 9* (eds. H. J. van den Herik and B. Monien), pp. 59–71, Universiteit Maastricht. ISBN 90–6216–5761.

Heinz, E. A. (1999b). Self-Play Experiments in Computer Chess Revisited. *Advances in Computer Chess 9* (eds. H. J. van den Herik and B. Monien), pp. 73–91, Universiteit Maastricht. ISBN 90–6216–5761.

Heinz, E. A. (2000a). A New Self-Play Experiment in Computer Chess. Technical Report 608, Massachusetts Institute of Technology, Laboratory of Computer Science, United States of America. Technical Memo No. 608 (MIT-LCS-TM-608).

Heinz, E. A. (2000b). *Scalable Search in Computer Chess (Algorithmic Enhancements and Experiments at High Search Depths).* Vieweg-Verlag, Braunschweig, Germany. ISBN 3–528–05732–7.

Heinz, E. A. (2001). Self-Play, Deep Search and Diminishing Returns. *ICGA Journal*, Vol. 24, No. 2, pp. 75–79. ISSN 1389–6911.

Heinz, E. A. (2003). Follow-Up on Self-Play, Deep Search, and Diminishing Returns. *ICGA Journal*, Vol. 26, No. 2, pp. 75–80. ISSN 1389–6911.

Hyatt, R. M. and Newborn, M. M. (1997). CRAFTY Goes Deep. *ICCA Journal*, Vol. 20, No. 2, pp. 79–86. ISSN 0920–234X.

Junghanns, A., Schaeffer, J., Brockington, M. G., Björnsson, Y., and Marsland, T. A. (1997). Diminishing Returns for Additional Search in Chess. *Advances in Computer Chess 8* (eds. H. J. van den Herik and J. W. H. M. Uiterwijk), pp. 53–67, Universiteit Maastricht. ISBN 90–6216–2347.

Matanović, A. (ed.) (1974-1979). *Encyclopedia of Chess Openings.* Chess Informant, Beograd, Jugoslavija.

Mysliwietz, P. (1994). *Konstruktion und Optimierung von Bewertungsfunktionen beim Schach.* Ph.D. thesis, University of Paderborn, Germany.

Newborn, M. M. (1985). A Hypothesis Concerning the Strength of Chess Programs. *ICCA Journal*, Vol. 8, No. 4, pp. 209–215.

Reinfeld, F. (1958). *Win at Chess.* Dover Publications, Inc., New York. ISBN 0–486–41878–2.

Reinfeld, F. (1971). *1001 Brilliant Ways to Checkmate.* Wilshire Book Company. ISBN 0–879–80110–7.

Schaeffer, J. (1986). *Experiments in Search and Knowledge.* Ph.D. thesis, University of Waterloo. Also printed as Technical Report (TR 86-12), Department of Computer Science, University of Alberta, Alberta, Canada.

Szabo, A. (1984). Computer Chess Tactics and Strategy. M.Sc. thesis, University of British Colombia.

Szabo, A. and Szabo, B. (1988). The Technology Curve Revisited. *ICCA Journal*, Vol. 11, No. 1, pp. 14–20. ISSN 0920–234X.

Thompson, K. L. (1982). Computer Chess Strength. *Advances in Computer Chess 3* (ed. M. R. B. Clarke), pp. 55–56, Pergamon Press Ltd. ISBN 0–08–026898–6.

9. APPENDICES

APPENDIX A: EXPERIMENTAL SETUP

The chess program CRAFTY version 19.6 was used with a 768 MByte transposition table and a 24 MByte pawn hash table, equipped with all 3-4-men EGTBs with 1377 KByte for indices and decompression tables, and no openingbook. Fourteen stand-alone computers were used (3.06 GHz Intel P4, 512 KByte cache, 1 GByte RAM).

APPENDIX B: BEST-CHANGE RATES FOR DIFFERENT TEST SETS

Search Depth	ECO: BC(d) in % (SE)	WAC: BC(d) in % (SE)	BWTC: BC(d) in % (SE)
2	35.84 (0.55)	50.00 (3.17)	59.70 (1.60)
3	34.97 (0.55)	38.80 (3.09)	27.19 (1.45)
4	32.78 (0.54)	24.80 (2.74)	10.45 (1.00)
5	31.33 (0.53)	19.92 (2.61)	16.33 (1.47)
6	29.76 (0.52)	11.16 (2.15)	11.98 (1.78)
7	27.15 (0.51)	9.30 (1.99)	7.48 (1.47)
8	25.66 (0.50)	6.90 (1.78)	6.86 (1.77)
9	23.29 (0.48)	2.96 (1.19)	3.48 (1.30)
10	22.61 (0.48)	3.09 (1.25)	3.45 (1.70)
11	20.29 (0.46)	0.52 (0.52)	1.72 (1.21)
12	19.24 (0.45)	1.60 (0.92)	2.82 (1.98)
13	17.76 (0.44)	3.19 (1.29)	1.43 (1.43)
14	16.30 (0.42)	1.07 (0.75)	2.27 (2.27)
15	14.72 (0.41)	1.62 (0.93)	0.00 (0.00)
16	13.11 (0.50)	0.00 (0.00)	0.00 (0.00)
17	11.79 (0.48)	1.18 (0.83)	0.00 (0.00)
18	11.74 (0.48)	1.29 (0.91)	9.09 (9.09)
19	-	1.40 (0.99)	0.00 (0.00)
20	-	0.00 (0.00)	0.00 (0.00)

Table 4: BC(d) of CRAFTY on three different sets of test positions.

Search Depth	Mate-in-2 in %	Mate-in-3 in %	Mate-in-4 in %	Mate-in-5 in %	Mate-in-6 in %	Mate-in-7 in %
2	18.22	49.73	40.07	36.59	36.85	28.83
3	8.77	16.04	32.79	40.68	31.24	29.05
4	-	4.45	11.29	20.23	20.67	18.69
5	-	14.55	13.97	18.49	21.67	25.68
6	-	-	8.78	15.82	17.73	12.84
7	-	-	5.19	6.81	14.29	12.10
8	-	-	-	5.11	11.08	14.93
9	-	-	-	3.89	5.82	10.50
10	-	-	-	-	3.46	7.63
11	-	-	-	-	1.46	3.44
12	-	-	-	-	-	2.29
13	-	-	-	-	-	2.31

Table 5: BC(d) of CRAFTY on different sets of mate-in-X positions.