DETERMINING THE STRENGTH OF CHESS PLAYERS BASED ON ACTUAL PLAY

Diogo R. Ferreira[1]

Oeiras, Portugal

ABSTRACT

The increasing strength of chess engines and their fine-tuned ability to evaluate positions provide the opportunity to use computer analysis for assessing the strength of players on a move-by-move basis, as a game is being played. As each player makes a move on the board, a sufficiently strong engine will be able to determine whether that move has contributed to improve or degrade the position. This change in position evaluation is defined as the gain per move. In this article, we assume that it is possible to characterise a player by a distribution of gain over several moves, preferably across several games. We present an approach to estimate the strength of players based on their distributions of gain. This estimated strength is expressed in terms of a perceived Elo rating, which may differ from a player's actual rating. However, the average of the perceived Elo ratings is equal to the average of the actual Elo ratings for a set of players. A critical factor in the approach is the need to determine the strength of the engine used for analysis. Once this is obtained, it is possible to carry out all sorts of comparisons between players, even players from different eras who never competed against each other. There are many potential applications but here we focus mainly on developing the approach, using a few recent tournaments as well as some historical events for illustrative purposes.

1. INTRODUCTION

In chess, as in most other sports, the strength of players or competing teams can be estimated based on the result of games between them. However, in chess, unlike many other sports, it is possible to assess the strength of each move, at least subjectively, based on some sort of positional evaluation. In principle, by evaluating the positions on the board before and after a move was made, it is possible to assess whether the move was good or bad. Of course, straightforward approaches to position evaluation – such as summing up the value of material on the board – will not work since, for example, in the middle of a piece exchange the material may appear to be temporarily unbalanced. A position should therefore be evaluated on the basis of its potential, and this can be done by studying the main lines that arise from that position. Assuming that both players will always choose the best possible moves, the problem of evaluating a given position is recursive since it requires evaluating the new positions that can arise from the given one. Chess programs make use of a recursive algorithm known as *minimax*, which performs a kind of tree-based search in order to find the best move.[2] The best move has an assigned score that corresponds to the best position that can arise out of the current one, assuming that both players play best moves throughout. A more detailed explanation of how computers play chess goes beyond the scope of this work and can be found elsewhere (e.g., Levy and Newborn, 1990; Campbell, Hoane, and Hsu, 2002), but this general overview should suffice to understand the kind of analysis pursued here. Chess programs are becoming amazingly strong, to the point that there is little hope that a world-class player can outplay the best computer programs available today. After the much publicized match between Kasparov and DEEP BLUE in 1997 (Hsu, 2002), which ended in the machine's favour, chess engines kept evolving to the point that they now seem to belong to an entirely different league.[3] Disappointing as it may seem, this also creates new opportunities

[1] IST – Technical University of Lisbon, Avenida Prof. Dr. Cavaco Silva, 2744-016 Oeiras, Portugal. Email:diogo.ferreira@ist.utl.pt

[2] While the origin of the minimax algorithm can be traced back to von Neumann (1928), the reader may find a more convenient introduction in Russell and Norvig (2010).

[3] A quick look at computer chess ratings lists such as CCRL (http://www.computerchess.org.uk/) and SSDF (http://ssdf.bosjo.net/) shows that the top 10 is dominated by engines with an estimated Elo rating above 3000 (March 2012).

for studying the game and improving the skill of human players. Here, we are interested in taking advantage of the strength of computer programs to assess the strength of human players, on a move-by-move basis. When Arpad Elo, inventor of the Elo rating system, wrote in his book (Elo, 1978): “Perhaps a panel of experts in the art of the game could evaluate each move on some arbitrary scale and crudely express the total performance numerically (...)”, it was hard to imagine that three decades later there would be computer programs that excel at evaluating positions on the board, helping humans to pinpoint exactly where something went wrong. The analysis capabilities of current programs are such that there is hardly a way to surprise a chess engine with a good move; what one can hope for is only to avoid poor moves that may compromise the position. From this perspective, a human player facing (theoretically) perfect play by a machine can expect at most to be able to keep the position balanced; anything else are just inaccuracies and bad moves that turn the position in the machine’s favour. So far (for some, unfortunately), such machine does not yet exist, and throughout the history of chess, human creativity and ingenuity were able to create masterpieces that elude even the strongest engines available today.

2. AN EXAMPLE

In 1956, a 13-year-old player known as Bobby Fischer produced one of such masterpieces, when playing with black pieces against a leading player at the time. For its brilliancy, this particular game would become known as “the game of the century” (Burgess, Nunn, and Emms, 2010). Figure 1 presents the actual move list, as well as an evaluation of the position after each move. The program used to perform this analysis was the HOUDINI chess engine, which, at the time of writing, is reportedly the strongest engine available. It leads several rankings of chess programs, and it has consistently won over other top engines in multiple-game matches. The particular version used in this work was HOUDINI 1.5a 64-bit configured with a fixed depth per move, set to 20, meaning that the program will look ahead 20 plies, where a ply is a move from either white or black. For example, when analysing the position arising after move 11.♗g5, the engine will look ahead possible moves up to move 21. (Whether this is sufficient to obtain an accurate evaluation will be discussed later.)

The position evaluation is in units of pawns, where a pawn is worth 1.0. For example, after white’s move 11.♗g5, the position is evaluated as -1.09, the negative sign meaning that the advantage is in black’s favour. In this case, black is assessed to have a positional advantage that is roughly equivalent to one pawn, even though material is balanced, as can be seen in Figure 2(a). In this position, Fischer played the startlingly brilliant 11...♘a4!!, perhaps the move that the game is best known for. However, according to the analysis in Figure 1 this is exactly what the engine was waiting for. Both before and after the move, the position is evaluated as -1.09, meaning that Fischer played as the engine expected. In this sense, the *gain* achieved in playing this move is 0.0, as indicated in the fourth column of Figure 1. However, for the opponent, the previous move 11.♗g5? can be considered to be a mistake, since it decreased the position evaluation by -1.37 pawns. In general, an increase in the position evaluation is a positive gain for white, while a decrease is a positive gain for black.

On move 17, Fischer player another brilliant move by receding his bishop to 17...♗e6!! and therefore sacrificing his queen which, as shown in Figure 2(b), is being threatened by white’s bishop on c5. In the analysis of Figure 1, the move 17...♗e6 is reported as incurring in a small loss (-0.11). Actually, the engine recommends 17...♗e6 as the best move, so the small loss can be explained by the fact that as the engine proceeds to analyse the next move, it looks ahead one more ply and concludes that the position had been slightly overestimated for black. Such small variations or updates in the evaluation occur quite often and can be observed throughout the game. Larger values of loss usually indicate that the engine would prefer a different move. Such is the case of 26.h3 (-2.60), where the engine prefers 26.a3. Also, instead of 34.♘e5 (-10.25) the engine prefers 34.♘e1 or 34.♘d4. And after 35.♔g1 (-11.18) the engine suddenly discovers a forced mate; for that reason, 35.♘d3 would be preferred. In contrast, there is 34...♔g7 (3.73) with a significant positive gain, which is due simply to an update in the position evaluation, since the engine also selects 34...♔g7 as the best move. In Figure 2(c), a different situation arises: here, Fischer could have mated in 4 moves with 37...♖e2+ 38.♔d1 ♗b3+ 39.♔c1 ♗a3+ 40.♔b1 ♖e1# rather than mating in 5 moves as it happened in the game. However, the move 37...♗b4+ does not change the position evaluation, since white is still under forced mate.

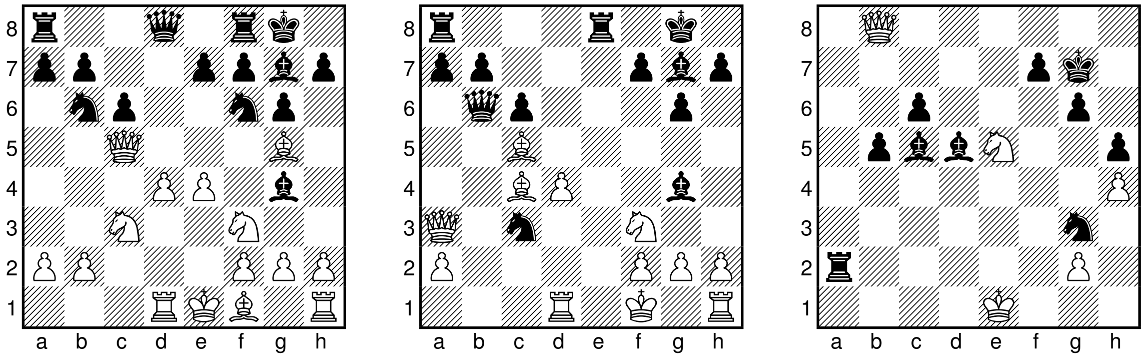
“The Game of the Century”
Donald Byrne – Robert James Fischer
New York, 1956

Move	Evaluation	Player	Gain
1 ♘f3	0.12	Byrne	-0.01
1...♘f6	0.12	Fischer	0.00
2 c4	0.07	Byrne	-0.05
2...g6	0.23	Fischer	-0.16
3 ♘c3	0.10	Byrne	-0.13
3...♗g7	0.21	Fischer	-0.11
4 d4	0.19	Byrne	-0.02
4...O-O	0.23	Fischer	-0.04
5 ♗f4	0.06	Byrne	-0.17
5...d5	0.16	Fischer	-0.10
6 ♕b3	0.07	Byrne	-0.09
6...d×c4	0.05	Fischer	0.02
7 ♕×c4	0.05	Byrne	0.00
7...c6	0.09	Fischer	-0.04
8 e4	0.05	Byrne	-0.04
8...♘bd7	0.18	Fischer	-0.13
9 ♖d1	0.15	Byrne	-0.03
9...♘b6	0.18	Fischer	-0.03
10 ♕c5	0.24	Byrne	0.06
10...♗g4	0.28	Fischer	-0.04
11 ♗g5	-1.09	Byrne	-1.37
11...♘a4	-1.09	Fischer	0.00
12 ♕a3	-1.12	Byrne	-0.03
12...♘×c3	-1.11	Fischer	-0.01
13 b×c3	-1.10	Byrne	0.01
13...♘×e4	-1.11	Fischer	0.01
14 ♗×e7	-1.68	Byrne	-0.57
14...♕b6	-1.30	Fischer	-0.38
15 ♗c4	-1.26	Byrne	0.04
15...♘×c3	-1.39	Fischer	0.13
16 ♗c5	-1.48	Byrne	-0.09
16...♖fe8+	-1.22	Fischer	-0.26
17 ♔f1	-1.26	Byrne	-0.04
17...♗e6	-1.15	Fischer	-0.11
18 ♗×b6	-6.71	Byrne	-5.56
18...♗×c4+	-6.62	Fischer	-0.09
19 ♔g1	-6.64	Byrne	-0.02
19...♘e2+	-6.64	Fischer	0.00
20 ♔f1	-6.64	Byrne	0.00
20...♘×d4+	-6.64	Fischer	0.00
21 ♔g1	-6.64	Byrne	0.00
21...♘e2+	-6.82	Fischer	0.18
22 ♔f1	-6.82	Byrne	0.00
22...♘c3+	-6.82	Fischer	0.00
23 ♔g1	-6.82	Byrne	0.00
23...a×b6	-6.80	Fischer	-0.02
24 ♕b4	-6.83	Byrne	-0.03
24...♖a4	-6.86	Fischer	0.03
25 ♕×b6	-7.41	Byrne	-0.55
25...♘×d1	-7.24	Fischer	-0.17
26 h3	-9.84	Byrne	-2.60
26...♖×a2	-10.05	Fischer	0.21
27 ♔h2	-10.47	Byrne	-0.42
27...♘×f2	-10.64	Fischer	0.17
28 ♖e1	-11.99	Byrne	-1.35
28...♖×e1	-11.57	Fischer	-0.42
29 ♕d8+	-11.57	Byrne	0.00
29...♗f8	-11.57	Fischer	0.00
30 ♘×e1	-11.14	Byrne	0.43
30...♗d5	-12.01	Fischer	0.87
31 ♘f3	-12.75	Byrne	-0.74
31...♘e4	-12.90	Fischer	0.15
32 ♕b8	-13.37	Byrne	-0.47
32...b5	-13.37	Fischer	0.00
33 h4	-13.37	Byrne	0.00
33...h5	-13.84	Fischer	0.47
34 ♘e5	-24.09	Byrne	-10.25
34...♔g7	-27.82	Fischer	3.73
35 ♔g1	-39.00	Byrne	-11.18
35...♗c5+	-39.00	Fischer	0.00
36 ♔f1	-39.00	Byrne	0.00
36...♘g3+	-39.00	Fischer	0.00
37 ♔e1	-39.00	Byrne	0.00
37...♗b4+	-39.00	Fischer	0.00
38 ♔d1	-39.00	Byrne	0.00
38...♗b3+	-39.00	Fischer	0.00
39 ♔c1	-39.00	Byrne	0.00
39...♘e2+	-39.00	Fischer	0.00
40 ♔b1	-39.00	Byrne	0.00
40...♘c3+	-39.00	Fischer	0.00
41 ♔c1	-39.00	Byrne	0.00
41...♖c2#	-39.00	Fischer	0.00

Figure 1: An analysis of the game of the century, using HOUDINI 1.5a 64-bit with depth 20.

3. THE VALUE OF CHECKMATE

Following the analysis of Figure 1, it is possible to obtain a crude estimate of the strength of both players in this game by averaging their respective gains per move. Fischer played with an average gain of approximately 0.10, while Byrne played with an average of -0.86. Since the gain is usually negative, some authors refer to this metric as the *mean loss per move* (see Guid and Bratko, 2006). In this particular game, the average gain was positive for Fischer, which does not necessarily mean that Fischer played at a stronger level than HOUDINI 1.5a at depth 20. In the next section, we will delve into a more rigorous approach to estimate player strength based on the gain per move. But before that, the attentive reader may have noticed in the analysis of Figure 1 that we have used a value of -39.0 to denote the fact that white is checkmated.



(a) Position before 11…♘a4!!

(b) Position before 17…♗e6!!

(c) Position before 37…♗b4+

Figure 2: Notable positions in the analysis of the game of the century.

The relative value of chess pieces is fairly well established. With the value of the pawn set to unity, the queen is worth 9, rooks are worth 5, bishops and knights are worth 3. However, for the value of the king there is no common agreement. Theoretically, its value would be infinite, since it can never be captured or exchanged; Shannon (1950) suggested assigning a value of 200 to make it more important than other considerations. In practice, chess programs assign an arbitrary value such as 100 or even 1000. For the purpose of calculating the gain per move, this value becomes relevant and involves a careful decision. A too high value for the king implies a huge penalty for the player who lets the king be checkmated (in this sense, it would be better to resign rather than running the risk of reaching such a position). In contrast, a value for the king that is too low may not reflect the actual importance of that piece in determining the outcome of the game.

In the analysis of Figure 1, the engine considers that Byrne brought his king under forced checkmate on move 35.♔g1. Given the previous position evaluation of -27.82, the gain (or loss) incurred in playing 35.♔g1 is $g = v - (-27.82)$ where v is the evaluation score associated with white being checkmated. Clearly, g should be negative for Byrne when playing 35.♔g1, but it should not be so large as to have a disproportionate influence on the average gain per move of that player. After all, Byrne could have chosen to resign earlier, and should not be punished too hard for deciding to play on. At the start of the game, the sum of all material on the board (except the king) amounts to 39.0 (in units of pawn). Considering that the king is worth the sacrifice of any and every other piece, it can be worth as much as 39.0. In practice, it is very uncommon to see evaluation scores outside the range $[-39.0; 39.0]$ (at least for the chess engine used here). So, in the interest of not punishing players too hard for checkmate, we decided to settle on 39.0 as the maximum penalty.

It is interesting to note that with checkmate being worth 39.0, a difference in the average gain of players of 1.0 pawn per move would mean that the game is expected to end in checkmate before move 40, which is a standard time control in chess. As an example, in the game of the century, and according to the analysis of Figure 1, Fischer played with an average gain per move of 0.10, while Byrne played with an average gain of -0.86; the difference is 0.96 and black was mated in 41 moves. The choice of 39.0 as the maximum penalty is debatable, but it does not have a profound impact in our approach. It just represents an attempt to take all moves into account when calculating the distribution of gain for a player. A second option would be to discard moves outside a certain range of evaluation score, as Guid and Bratko (2006) did, albeit for different reasons.

Also, the engine depth used to carry out the analysis is of a lesser concern than could be supposed initially. First, we are interested in the position evaluation, and not in the best move; so even if the choice of a best move changes due to an increase in search depth, this is irrelevant if the overall position evaluation does not change. Second, if the search depth is 24 instead of 20, the engine finds a forced mate one move earlier, at 34.♘e5; this means that the position evaluation could have reached -39.0 sooner, but still a negative gain would have been attributed to Byrne in any case, which would result in roughly the same average gain per move. Third, even if by virtue of an increased depth the engine finds that the weaker player played even worse and the stronger player did not play so good, this will have a limited impact on the relative strength of both players, since it is calculated in terms of the difference of gain. For these reasons, the discussion on the effect of search depth will not be pursued here further; instead, our main focus is on developing an approach to estimate the strength of players, and the same approach can be applied with increased search depth.

4. CALCULATING THE EXPECTED SCORE

Most rating systems rely on the assumption that the individual performances of a player are normally distributed. Such is the case of the Elo rating system (Elo, 1978), as well as more recent systems such as TrueSkill (Herbrich, Minka, and Graepel, 2007), which has also been applied to chess (Dangauthier *et al.*, 2008). Other approaches, such as the USCF Rating System (Glickman and Doan, 2011) and Whole-History Rating (Coulom, 2008), are based on the Bradley-Terry model (Bradley and Terry, 1952), which in turn relies on the assumption that the ratio between the strength of two players follows a logistic distribution. However, the distribution of the gain per move obtained through an analysis similar to the one presented in Figure 1 does not appear to follow any of these distributions. In fact, neither the normal nor the logistic distribution fit the peakedness and skewness (asymmetry) found in the distribution of gain.

Figure 3 shows the histograms of gain for each player. The frequency values in these histograms have been normalised, meaning that the sum of the heights of all bars is 1.0. For example, Byrne has 13 out of 41 moves with a gain of exactly 0.0, yielding a normalised frequency of $13/41 \cong 0.32$ at the origin. Also, the width of each bar is 0.01, meaning that the normalised frequency concerns values that are within a *centipawn* (one hundredth of a pawn) apart. In fact, HOUDINI provides its position evaluation in units of centipawn with no decimal places, so the maximum resolution is effectively 0.01, or 1 centipawn. Besides being normalised, the horizontal axes in Figure 3 are in units of centipawn; this makes the area under each histogram sum up to 1.0 and, according to a frequentist approach to probability, the histogram can be interpreted as a probability mass function.

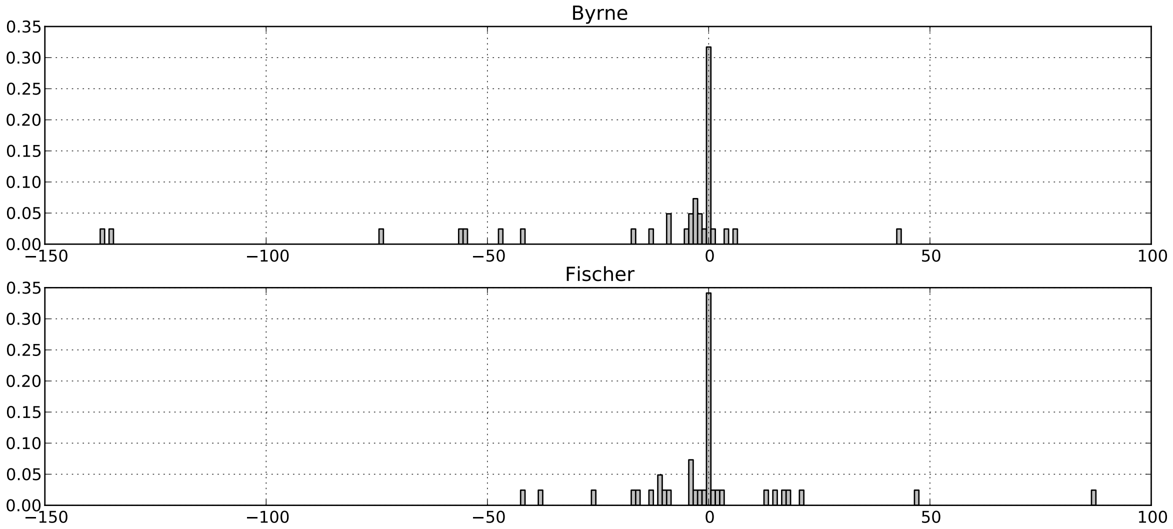


Figure 3: Normalised histograms of gain for each player, in intervals of 1 centipawn, plotted in the range of -150 to 100 centipawns.

The probability mass function that describes the distribution of gain has two salient characteristics. The first is its peakedness: when a player makes the move that the engine thinks is best, and this may happen quite often, the position evaluation is likely to remain unchanged, and hence there might be a significant number of moves with zero gain; in contrast, moves that are not as strong will end up having a gain that is more evenly distributed on the tails of the distribution. The second characteristic of the distribution of gain is its asymmetry: usually it is hard to find the best move, but it is relatively easier to find sub-optimal moves; when analysing moves with a strong engine, it is more likely that the gains lie in the negative range rather than in the positive range, since it is difficult to surprise the engine with a move that is better than the one it chooses, but it is comparatively easier to play a move that is worse.

Even if the move actually played is better than the move chosen by the engine, the reason for that may be that the engine (mistakenly) judges that the move played is worse; in this case, the player may be unjustly attributed a negative gain. Yet, the opposite effect may occur as well: a lack of depth may result in an excessive positive gain, as in 34…♔g7 (3.73) in Figure 1. In any case, such deviations have a limited impact with regard to the relative strength of players, which is to be determined based on a measure of difference between distributions of gain.

If the distribution of gain could be approximated by a normal distribution, then the *distribution of the difference* between two distributions of gain would also follow a normal curve. But since the normal distribution does not fit the distribution of gain, one could look for other distributions that provide a better fit. Given the lack of a function that provides an appropriate fit to the distribution of gain, in this work we use the normalised histogram above as an empirical density function (Waterman and Whiteman, 1978). In this setup, the histogram becomes an approximation to the probability mass function of gain as a discrete random variable. We can then make use of the well-known fact that the probability distribution of the difference between two random variables is the cross-correlation of their distributions. Let X and Y be two discrete random variables with probability distributions f_X and f_Y, respectively; then the probability distribution of $X - Y$ is given by the cross-correlation:

$$f_{X-Y}[n] = (f_Y \star f_X)[n] \triangleq \sum_{m=-\infty}^{\infty} f_Y[m] \cdot f_X[n+m] \tag{1}$$

Figure 4 provides an intuitive explanation of cross-correlation. If $n = 0$, this corresponds to probability of X being equal to Y, i.e., $P(X = Y)$, and can be computed by multiplying the values of both functions and summing all those terms. As n increases, f_X slides to the left a number of n units (i.e., centipawns) and the same multiplication followed by summation is applied. As the overlap between f_X and f_Y decreases, so does the value of f_{X-Y}. Eventually, one may come to a point where the functions do not overlap anymore, and $f_{X-Y}[n] = 0$ beyond that point. This means that as the value of n increases towards $+\infty$, or conversely when it decreases towards $-\infty$, the probability of observing such large difference between X and Y is eventually zero. Therefore, the values of n for which f_{X-Y} is nonzero can be found in a delimited range, which facilitates computation.

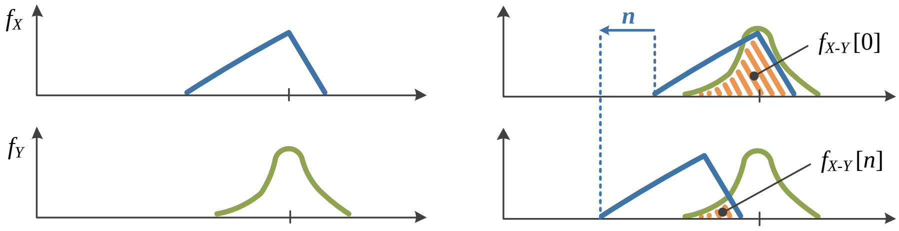


Figure 4: An intuitive explanation of the concept of cross-correlation.

To compute the expected score in a game between two players, we recall that in chess there are three possible outcomes, with a win being worth one point, a draw being worth half a point, and a loss being worth zero. Based on these considerations, the expected score can be computed as:

$$\begin{aligned} p_{XY} &\triangleq 0.5 \cdot P(X = Y) + 1 \cdot P(X > Y) \\ &= 0.5 \cdot f_{X-Y}[0] + \sum_{n=1}^{\infty} f_{X-Y}[n] \end{aligned} \tag{2}$$

Substituting the two instances of f_{X-Y} in Eq. (2) by the expression in Eq. (1), and carrying out the computations with f_X and f_Y representing the distribution of gain for Byrne and Fischer, respectively, the result is an expected score of 0.345. Conversely, using now f_X as the distribution of Fischer and f_Y the distribution of Byrne, the result is an expected score of 0.655. The sum of both results is exactly 1.0 as expected.

In the Elo rating system (Elo, 1978), a difference of d rating points corresponds to a percentage expectancy $p = \Phi(\frac{d}{200\sqrt{2}})$ where $\Phi(z)$ is the cumulative distribution function (CDF) of the standard normal distribution. Given the value of p, it is possible to find the corresponding value of $z = \Phi^{-1}(p)$ in a table of the standard normal CDF, and then compute $d = z \cdot 200 \cdot \sqrt{2}$. For the example above, a percentage expectancy of 65.5% corresponds to a rating difference of 113 points in favour of Fischer.

In summary, the relative strength between two players can be calculated according to the following algorithm:

Algorithm 1 Determine the relative strength between two players i and j

1. Run a computer analysis with a fixed depth per move for a game (or set of games) between two players i and j, taking note of the position evaluation before and after each move.
2. Compute the gain g_k for each move k. Let s_{k-1} and s_k be the position evaluation before and after move k, respectively. If the move was played by white, compute the gain as $g_k = s_k - s_{k-1}$; if the move was played by black, compute the gain as $g_k = -(s_k - s_{k-1})$.
3. Obtain the distributions of gain f_i and f_j for players i and j, respectively, by building the normalised histogram of gain for each player, using the smallest interval allowed by the resolution of the position evaluation (usually, 1 centipawn or 0.01 pawns).
4. Compute the expected score p_{ij} between player i and j through $p_{ij} = 0.5 \cdot (f_j \star f_i)[0] + \sum_{n>0}(f_j \star f_i)[n]$ where $(f_j \star f_i)[n]$ is the cross-correlation between f_j and f_i.
5. Compute the rating difference d_{ij} between player i and j by $d_{ij} = 200\sqrt{2} \cdot \Phi^{-1}(p_{ij})$ where $\Phi^{-1}(z)$ is the inverse CDF of the standard normal distribution.

5. AN ANALYSIS OF SOME RECENT TOURNAMENTS

In the previous example, the relative strength of players (in terms of expected score and rating difference) was estimated based on computer analysis of a single game. Ideally, one would like to perform such analysis over several games, in order to obtain a more accurate estimate of a player's strength. The idea is that, by taking the analysis of several games together, one can build a histogram that is a better approximation of the "true" distribution of gain. However, one should bear in mind that such an approach may suffer from some undesirable effects, such as the particular conditions in which each game was played, and also the fact that player strength varies inevitably over time, an effect that will become more noticeable when the games to be analysed cross a relatively long period of time (e.g., weeks or even months). We start with the analysis of a single tournament that was played over a period of 10 days and that included some of the world's top players.

The London Chess Classic 2011 was the third in a series of tournaments taking place in London. The first London Chess Classic was held in December 2009 and the second in December 2010; both were won by Magnus Carlsen. Here we focus on the 2011 event, which took place from 3rd–12th December 2011, and had the participation of all four players with a rating of 2800+ at that time: Carlsen (2826), Anand (2811), Aronian (2802), and Kramnik (2800). The other players were Nakamura (2758), Adams (2734), Short (2698), McShane (2671), and Howell (2633). The tournament had a special interest due to the fact that it featured the world champion Anand in one of his last tournaments before defending the title against Boris Gelfand in May 2012.

Having 9 players, the London Chess Classic 2011 was a round-robin tournament where each player played every other once. There were a total of 9 rounds and 4 games per round, with a different player having a "bye" in each round, so each player effectively played 8 rounds, against 8 different opponents. Table 1 presents the cross-table of results, with players ordered by decreasing total score. The original scoring system adopted in the London Chess Classic assigned 3 points to a win, 1 to a draw, a zero to a loss, but here we keep the traditional scoring of 1, 1/2, and zero, respectively. This does not change the fact Kramnik was the winner, as he had both the highest total score and the largest number of wins.

The results of individual games are given in parenthesis in each cell of Table 1. Also within each cell is an analysis carried out based on the distribution of gain for each player, where the distribution was obtained by aggregating the gain per move across *all* games of that player in the tournament. Again, Houdini 1.5a 64-bit with a fixed depth of 20 was used to analyse every game. In each cell of Table 1, the second line gives the expected score using the cross-correlation between the distributions of gain for both players. It can be observed (1) that the values for the expected score are in the range $[0.41, 0.59]$ and (2) that they have an interesting correlation with the actual result of the game, despite of being calculated based on the overall distribution of each player.

Below the expected score, in the third line of each cell in Table 1, is the rating difference according to the Elo system for the given expected score. This provides the basis for a second analysis that aims at estimating the strength of each player in terms of a perceived Elo rating. The original Elo rating for each player is given in

	Player	Elo	1	2	3	4	5	6	7	8	9	Total	Estimated strength
1	Kramnik	2800	–	(½) 0.52 +15	(½) 0.58 +57	(1) 0.57 +51	(½) 0.54 +28	(½) 0.56 +43	(1) 0.57 +50	(1) 0.59 +63	(1) 0.59 +67	(6.0) 4.52	2790 -10
2	Carlsen	2826	(½) 0.48 -15	–	(1) 0.56 +40	(½) 0.55 +37	(½) 0.52 +11	(½) 0.54 +26	(½) 0.55 +36	(1) 0.56 +46	(1) 0.57 +51	(5.5) 4.32	2774 -52
3	Nakamura	2758	(½) 0.42 -57	(0) 0.44 -40	–	(½) 0.50 -3	(1) 0.46 -27	(1) 0.48 -17	(½) 0.50 -4	(1) 0.51 +4	(1) 0.52 +12	(5.5) 3.82	2734 -24
4	McShane	2671	(0) 0.43 -51	(½) 0.45 -37	(½) 0.50 +3	–	(½) 0.47 -24	(½) 0.48 -13	(1) 0.50 -1	(1) 0.51 +8	(1) 0.52 +14	(5.0) 3.86	2737 +66
5	Anand	2811	(½) 0.46 -28	(½) 0.48 -11	(0) 0.54 +27	(½) 0.53 +24	–	(½) 0.52 +13	(1) 0.53 +23	(½) 0.55 +32	(½) 0.55 +39	(4.0) 4.17	2762 -49
6	Aronian	2802	(½) 0.44 -43	(½) 0.46 -26	(0) 0.52 +17	(½) 0.52 +13	(½) 0.48 -13	–	(1) 0.52 +12	(½) 0.53 +22	(½) 0.54 +29	(4.0) 4.02	2750 -52
7	Short	2698	(0) 0.43 -50	(½) 0.45 -36	(½) 0.50 +4	(0) 0.50 +1	(0) 0.47 -23	(0) 0.48 -12	–	(½) 0.51 +9	(1) 0.52 +15	(2.5) 3.87	2738 +40
8	Howell	2633	(0) 0.41 -63	(0) 0.44 -46	(0) 0.49 -4	(0) 0.49 -8	(½) 0.45 -32	(½) 0.47 -22	(½) 0.49 -9	–	(½) 0.51 +8	(2.0) 3.75	2729 +96
9	Adams	2734	(0) 0.41 -67	(0) 0.43 -51	(0) 0.48 -12	(0) 0.48 -14	(½) 0.45 -39	(½) 0.46 -29	(0) 0.48 -15	(½) 0.49 -8	–	(1.5) 3.67	2722 -12

Table 1: An analysis of London Chess Classic 2011.

the third column of Table 1; the last column provides an estimate (and the corresponding difference to the actual rating) based on the rating differences recorded in each game. The procedure is described next.

Let d_{ij} be the estimated rating difference in a game between player i and player j, whose strength in terms of estimated ratings r'_i and r'_j, respectively, are to be determined. The way to determine r'_i and r'_j is to make the difference $\delta_{ij} = r'_i - r'_j$ as close as possible to d_{ij}. Let the squared error of an estimated r'_i be defined as:

$$e_i \triangleq \sum_j (\delta_{ij} - d_{ij})^2 = \sum_j (r'_i - r'_j - d_{ij})^2 \tag{3}$$

Then it is possible to obtain an estimate for the rating r'_i of each player i by choosing r'_i so as to minimise the squared error e_i when the estimated ratings r'_j of all opponents j are given. Of course, the ratings r'_j are themselves calculated using the same procedure, so the estimation of all ratings becomes an iterative algorithm that runs until all ratings converge.

Algorithm 2 Determine the strength of each player i, in terms of a perceived Elo rating r'_i, for a set of players in a tournament, where the actual Elo ratings r_i are known

1. Initialise each r'_i with the actual Elo rating r_i of player i.
2. For every player i, find r'_i by minimising the squared error e_i defined in Eq. (3).
3. If there was a change to any r'_i in the previous step, go back to step 2.
4. After the ratings r'_i have converged, add a quantity $\Delta = \bar{r} - \bar{r}'$ to each rating r'_i where $\bar{r}$ is the average of the actual Elo ratings r_i and $\bar{r}'$ is the average of the estimated ratings r'_i.

The last step is used to ensure that the average $\bar{r}'$ of the estimated ratings is equal to the average $\bar{r}$ of the actual Elo ratings. Such step is necessary since there are multiple solutions to the above minimisation problem; in fact, it is possible to find another solution $r''_i = r'_i + C$ just by adding the same constant C to every rating r'_i. By ensuring that the average $\bar{r}'$ is equal to that of the actual ratings $\bar{r}$, we make the estimated ratings r'_i become closer, on average, to the actual Elo ratings r_i.

The results of applying Algorithm 2 are presented in the last column of Table 1. On one hand, we see that Kramnik had an estimated performance that is close to his actual Elo rating, while other top players seem to have slightly underperformed. The English players, on the other hand, outperformed themselves in face of such competition, except for Adams who, however, did not play as bad as his number of losses may suggest. During the tournament, it was apparent that McShane was performing extremely well, and at a certain point he was leading the tournament together with Kramnik when the two faced each other in round 8, Kramnik coming out with a win.

The same set of 2800+ players, together with Nakamura, had been playing another tournament the previous month, in November 2011. It is interesting to compare the estimated strength of these players in both tournaments, in order to determine whether there is consistency in the distribution of gain of a given player. For that purpose, we built the overall distribution of gain for each player in each tournament and computed the expected score between these player-tournaments. Table 2 presents the results together with the corresponding rating differences. If the distribution of gain for each player would be consistent across tournaments, then it would be expected to see noticeably smaller values of rating difference in the diagonal of Table 2, when compared to other cells. However, that is not the case. The conclusion to be drawn from here is that the proposed approach can hardly be used to predict the outcome of future games; this is not so much a problem of the approach itself, but a matter of the inherent unpredictability of the outcome of chess games. Previous initiatives, in particular the chess rating competitions organised by Jeff Sonas, have shown just how difficult making such predictions can be.[4]

	Anand @TalMemorial	Aronian @TalMemorial	Carlsen @TalMemorial	Kramnik @TalMemorial	Nakamura @TalMemorial
Anand @London	0.485 -11	0.497 -2	0.463 -26	0.526 +18	0.528 +20
Aronian @London	0.465 -25	0.479 -15	0.442 -41	0.509 +7	0.512 +9
Carlsen @London	0.503 +2	0.516 +11	0.480 -14	0.545 +32	0.547 +33
Kramnik @London	0.532 +22	0.543 +31	0.501 +1	0.569 +49	0.571 +50
Nakamura @London	0.443 -41	0.455 -32	0.425 -53	0.485 -11	0.488 -9

Table 2: Comparison between players at London Chess Classic 2011 and Tal Memorial 2011.

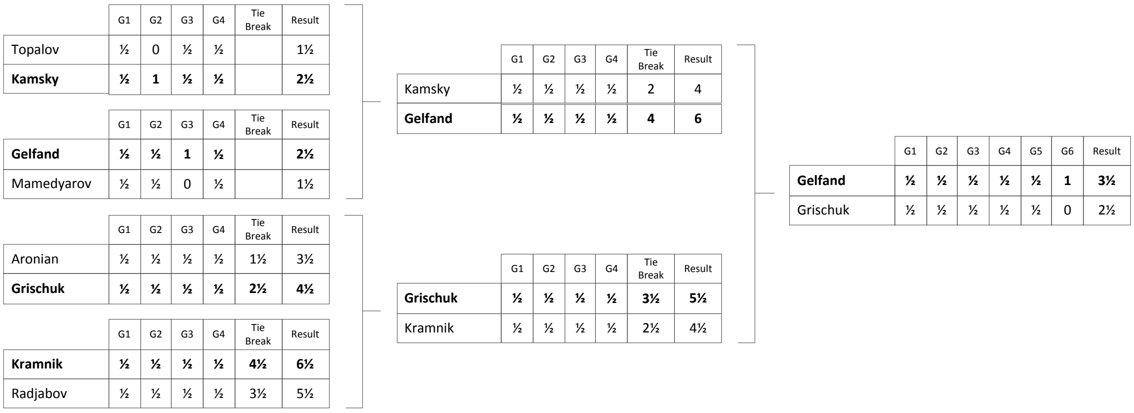
To obtain a better estimate of the relative strength of two players, one may want to look at a series of games between those players. For that purpose, it becomes useful to look at chess events that consist of multiple-game matches. Such is the case of the Candidates 2011 tournament, which took place in May 2011. This was a knockout tournament with 4-game matches between players at the quarter-finals (8 players) and semi-finals (4 players), and a 6-game final match with two players. In case of tied results after a match, there would be additional tie-break games in rapid and blitz format. The players were Aronian (2808), Kramnik (2785), Topalov (2775), Mamedyarov (2772), Grischuk (2747), Radjabov (2744), Gelfand (2733), and Kamsky (2732), where the ratings refer to May 2011. Figure 5 shows the results and pairings for this event.

Regarding this tournament, we used the distribution of gain across each multiple-game match between two players, including also the rapid and blitz tie-break games if that was the case. We then computed the expected score and rating difference as in Algorithm 1, obtaining the results shown in Table 3. Following the procedure of Algorithm 2, we obtained a set of estimated ratings by minimising the squared error of the rating differences. In this case, the relative order of the estimated ratings reflects much better the actual outcome of the tournament, with Gelfand earning the right to challenge the world champion Anand in a match to take place in May 2012.

6. DETERMINING THE ENGINE STRENGTH

Throughout the analysis in the previous sections, we have used Houdini 1.5a 64-bit with a fixed search depth of 20 plies. The question now arises as to what is the estimated strength of this engine. If the distribution of gain for the engine would be available, then it would be possible to compute the expected score p and rating difference d between a player and the engine, by following the same procedure as in steps 4–5 of Algorithm 1. If the rating r

[4] Jeff Sonas organised the first competition to predict chess results in 2010 (http://www.kaggle.com/c/chess), and a second competition in 2011 (http://www.kaggle.com/c/chessratings2).



	G1	G2	G3	G4	Tie Break	Result
Topalov	½	0	½	½		1½
Kamsky	**½**	**1**	**½**	**½**		**2½**

	G1	G2	G3	G4	Tie Break	Result
Gelfand	**½**	**½**	**1**	**½**		**2½**
Mamedyarov	½	½	0	½		1½

	G1	G2	G3	G4	Tie Break	Result
Aronian	½	½	½	½	1½	3½
Grischuk	**½**	**½**	**½**	**½**	**2½**	**4½**

	G1	G2	G3	G4	Tie Break	Result
Kramnik	**½**	**½**	**½**	**½**	**4½**	**6½**
Radjabov	½	½	½	½	3½	5½

	G1	G2	G3	G4	Tie Break	Result
Kamsky	½	½	½	½	2	4
Gelfand	**½**	**½**	**½**	**½**	**4**	**6**

	G1	G2	G3	G4	Tie Break	Result
Grischuk	**½**	**½**	**½**	**½**	**3½**	**5½**
Kramnik	½	½	½	½	2½	4½

	G1	G2	G3	G4	G5	G6	Result
Gelfand	**½**	**½**	**½**	**½**	**½**	**1**	**3½**
Grischuk	½	½	½	½	½	0	2½

Figure 5: Pairings and results for the Candidates 2011 tournament.

	Player	Elo	1	2	3	4	5	6	7	8	Estimated strength
1	Gelfand	2733	–	0.52 +12	–	0.55 +32	0.55 +34	–	–	–	2793 +60
2	Grischuk	2747	0.48 -12	–	0.51 +5	–	–	–	0.56 +43	–	2783 +36
3	Aronian	2808	–	0.49 -5	–	–	–	–	–	–	2778 -30
4	Kamsky	2732	0.45 -32	–	–	–	–	0.52 +12	–	–	2760 +28
5	Mamedyarov	2772	0.45 -34	–	–	–	–	–	–	–	2759 -13
6	Topalov	2775	–	–	–	0.48 -12	–	–	–	–	2748 -27
7	Kramnik	2785	–	0.44 -43	–	–	–	–	–	0.51 +7	2740 -45
8	Radjabov	2744	–	–	–	–	–	–	0.49 -7	–	2733 -11

Table 3: Estimated strength of players at the Candidates 2011 tournament.

of the player is known, then with both r and d it would possible to obtain an estimate for the engine strength as $r - d$. The problem is that the distribution of gain for the engine is unknown.

However, one can obtain an approximation to that distribution based on the following idea: the strength of the engine is equal to the strength of a player who would agree with the engine on every move and position evaluation; in principle, the gain per move of such player would always be zero, leading to a histogram with a single peak of height 1.0 at the origin. In reality, this is an approximation since the gain will not be zero all the time. For an engine with fixed search depth, as each move is made the engine is able to look one ply further into the game, eventually adjusting its own position evaluation. In general, using a fixed depth will always lead to less-than-perfect play, so there will be a certain tendency for the actual distribution to be slightly skewed towards negative gain. In any case, the gain per move refers to a player's own moves, so if a player chooses the same move as the engine, the position evaluation is likely to remain the same, or change only slightly; the largest changes will come from the opponent's (i.e., non-engine) moves. This means that assuming a constant zero gain for the engine is a reasonable approximation to its distribution of gain, but still it is only an approximation, of which the effects will become more apparent in the next section.

In this section, the procedure to estimate engine strength is based on Algorithm 3.

Table 4 illustrates the application of Algorithm 3 to the analysis of the London Chess Classic 2011 presented in Section 5. The Elo rating r_i of each player at the time of the tournament is given in the second column, while the estimated strength r'_i comes directly from Table 1. As before, the distribution of gain for each player was obtained by aggregating the gain per move across all games in the tournament. The expected score p_i between player and

Algorithm 3 Determine the engine strength based on the rating differences between the engine and a set of players whose Elo ratings and/or perceived ratings are known.

1. Given the distribution of gain for player i, and assuming that the distribution of gain for the engine can be described by a histogram with a single peak of height 1.0 at the origin, compute the expected score p_i between player i and the engine according to step 4 of Algorithm 1.
2. From the expected score p_i find the rating difference difference d_i between player i and the engine, according to step 5 of Algorithm 1.
3. If the actual Elo rating r_i of player i is known, obtain an estimate for engine strength as $r_e = r_i - d_i$.
4. For comparison, if the estimated strength r'_i of the player is available, obtain an estimate for the engine strength in terms of $r'_e = r'_i - d_i$.
5. If data is available on multiple players, apply steps 1–4 for each player i and calculate the average for both estimates $\bar{r}_e$ and $\bar{r}'_e$.

Player	Elo rating (r_i)	Estimated strength (r'_i)	Expected score against engine (p_i)	Difference in rating points (d_i)	Estimated rating of engine ($r_i - d_i$)	Estimated strength of engine ($r'_i - d_i$)
Kramnik	2800	2790	0.403	-69	2869	2859
Carlsen	2826	2774	0.396	-74	2900	2848
Nakamura	2758	2734	0.320	-132	2890	2866
McShane	2671	2737	0.337	-119	2790	2856
Anand	2811	2762	0.371	-93	2904	2855
Aronian	2802	2750	0.335	-120	2922	2870
Short	2698	2738	0.340	-116	2814	2854
Howell	2633	2729	0.304	-145	2778	2874
Adams	2734	2722	0.315	-136	2870	2858
				Average:	*2860*	*2860*

Table 4: Engine strength computed both by Elo rating and estimated strength of players at the London Chess Classic 2011.

engine, and the corresponding rating difference d_i, are shown in the fourth and fifth columns. Since the expected score is below 0.5 in every case, the rating difference is negative; this suggests that the engine is stronger than any of the players. The estimated rating for the engine, calculated as $r_e = r_i - d_i$, is shown in the next-to-last column, where the results are spread over a relatively wide range of values. A more consistent estimate can be obtained by making use of the estimated strength r'_i of each player, as in the last column of Table 4. However, taking the average of both estimates yields exactly the same result.

From this analysis, we can say that HOUDINI 1.5a 64-bit, with a fixed depth of 20, plays with an estimated strength of $\bar{r}_e = 2860$. A more accurate estimate could be obtained by applying the same procedure to a larger set of players and games, but even more important would be to have a better characterisation of the engine's distribution of gain. This, however, becomes a lesser concern if the search depth is increased. Over time, with increasing computing power and search depth, the single-peak approximation to the distribution of gain will become more accurate. In this work, we will have to contend with this preliminary estimate of engine strength.

7. DETERMINING PLAYER STRENGTH ON A MOVE-BY-MOVE BASIS

The approach delineated in the previous sections provides a means to estimate the strength of a player on a move-by-move basis. Such possibility is of interest, for example, to determine the strength of an player as a game is taking place. By analysing the moves in real time during the course of the game, one can obtain an instant assessment of the player strength, or at least the perceived strength of the players in that game. Given that there are websites that provide computer analysis of tournament games in real time[5], it is possible to obtain an assessment of the player strength directly on top of that analysis.

[5] An example is: http://chessbomb.com/

Assuming, as we did in the last section, that the distribution of gain for the engine is given by a histogram with a single-peak at the origin, we can compute the expected score and rating difference between player and engine by considering the moves that have been played up to a certain moment. Assume, for example, that while a player is pondering on the 20th move, the previous 19 moves have already been analysed (up to a certain depth) in terms of position evaluation and gain per move. Then the distribution of gain for that player, up to move 19, is already known. This can be used to compute the expected score p_i and the rating difference d_i between player i and the engine. If an estimate $\bar{r}_e$ for the engine strength is available, then $r'_i = \bar{r}_e + d_i$ provides an indication of the player's strength up to that point in the game. Algorithm 4 summarises these steps.

Algorithm 4 Determine the strength of a player after a number of moves, based on the rating difference between player and engine, when an estimate of engine strength is available.

1. Find the distribution of gain for player i, considering all moves that have been made by that player over the board, up to the present stage.

2. Assuming that the distribution of gain for the engine is described by a histogram with a single peak of height 1.0 at the origin, compute the expected score p_i between player i and the engine according to step 4 of Algorithm 1.

3. From the expected score p_i find the rating difference difference d_i between player i and the engine, according to step 5 of Algorithm 1.

4. If an estimate $\bar{r}_e$ for the engine strength is available, obtain an estimate r'_i of player strength through $r'_i = \bar{r}_e + d_i$.

Figure 6 illustrates the application of Algorithm 4 to the analysis of the game shown earlier in Figure 1. For the first six moves, the engine assigns a negative gain to Byrne, which results in an expected score of zero and a rating difference of $d_i = -\infty$; then, as the game develops, the rating difference eventually settles on $d_i = -185$. In contrast, Fischer begins with a move of zero gain, and therefore has an initial expected score of $p_j = 0.5$ and rating difference $d_j = 0$. This decreases sharply in the first few moves, which have a negative gain. Then, as more samples of gain are collected, the rating difference rises up to $d_j = -43$.

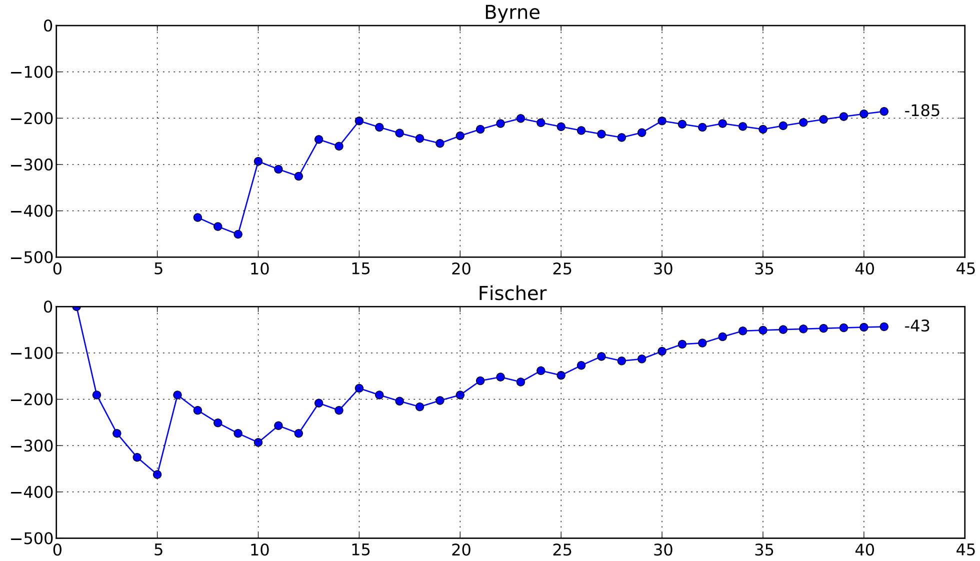


Figure 6: Estimated rating difference between player and engine after each move in the game of the century.

Considering the estimate of engine strength $\bar{r}_e = 2860$ that was obtained in the previous section, these results suggest that Fischer performed with an estimated strength of $r'_j = 2860 - 43 = 2817$, while Byrne performed with an estimated strength of $r'_i = 2860 - 185 = 2675$. These values should be interpreted with great care. First, these perceived ratings are heavily dependent on the estimate of the engine strength which, for the reasons discussed in the previous section, is not fully precise. Second, these results suggest a rating difference of 142 points, while the analysis of Section 4 indicated a difference of only 113 points. Although both values are in the same sort of magnitude, a closer agreement would be desirable. Our preliminary experiments on this issue reveal that if the distribution of gain for the engine is, as supposed, slightly skewed towards negative gain, the two estimates become more consistent. A more detailed analysis is to be carried out in future work.

8. FISCHER, KASPAROV, AND DEEP BLUE

In 1971, a well known candidates tournament took place, when Fischer qualified for the world championship 1972 against Spassky. At that time, the Elo rating system had just been adopted by FIDE[6], so it is interesting to compare the 1971 ratings with the perceived strength of players by today's standards. For this purpose, we built the distribution of gain for each player in the candidates tournament 1971 by considering all games of that player in the tournament. We then applied Algorithm 4 to obtain an estimate of the strength for each player.

Table 5 shows the results of applying such procedure, with players ranked by estimated strength. In this knockout tournament, Fischer beat Taimanov by 6–0 in a 6-game match, then Larsen by 6–0 again, and finally Petrosian by 6.5–2.5 over 9 games. Overall, his distribution of gain results in an expected score of $p_i = 0.376$ against the engine, which corresponds to a rating difference of $d_i = -89$ points; this, in turn, results in an estimated strength of $r'_i = 2860 - 89 = 2771$. For the reasons discussed before, the estimated strength of players given in Table 5 should be regarded as a rough estimate rather than a precise result. Still, these results are on average 100 points above the 1971 Elo ratings for the same players, which is consistent with the claim that Elo ratings have suffered an inflation of about 100 points in this time period (Howard, 2006).

	Player	Elo (1971)	1	2	3	4	5	6	7	8	9	10	p_i d_i	Estimated strength (2011)
1	Fischer	2760	–	0.53 +24	–	–	–	0.54 +28	–	0.6 +74	–	–	0.376 -89	2771
2	Petrosian	2640	0.47 -24	–	0.51 +5	0.52 +12	–	–	–	–	–	–	0.353 -107	2753
3	Korchnoi	2670	–	0.49 -5	–	–	–	–	0.52 +13	–	–	–	0.352 -107	2753
4	Hübner	2590	–	0.48 -12	–	–	–	–	–	–	–	–	0.351 -108	2752
5	Portisch	2630	–	–	–	–	–	–	–	–	–	0.53 +25	0.341 -116	2744
6	Taimanov	2620	0.46 -28	–	–	–	–	–	–	–	–	–	0.337 -119	2741
7	Geller	2630	–	–	0.48 -13	–	–	–	–	–	–	–	0.331 -124	2736
8	Larsen	2660	0.4 -74	–	–	–	–	–	–	–	0.52 +15	–	0.330 -124	2736
9	Uhlmann	2580	–	–	–	–	–	–	–	0.48 -15	–	–	0.318 -134	2726
10	Smyslov	2620	–	–	–	–	0.47 -25	–	–	–	–	–	0.282 -163	2697
	Average:	*2640*												*2741*

Table 5: Estimated strength of players in Candidates 1971, using engine strength as reference.

By the time Fischer met Spassky for the world championship match in 1972, his Elo rating was 2785 while Spassky, the second highest rated player at the time, had a rating of 2660, a difference of 125 points. An analysis by Algorithm 4, using the distribution of gain over all games in that world championship match, suggests that Fischer performed with an estimated strength of $r'_i = 2779$ and Spassky with an estimated strength of $r'_j = 2745$.

For comparison, we present a similar analysis for all world championship matches between Kasparov and Karpov in Table 6. Apart from the 1984 match, which was interrupted after 48 games without a declared winner, Kasparov secured the title in the remaining four encounters. It is worth noting that the estimated strength of both players does not change much across these matches, while their Elo ratings have a more noticeable fluctuation. Also, the performance of Fischer in 1972 appears to be above that of both Kasparov and Karpov in their matches.

[6] *Fédération Internationale des Échecs* (http://www.fide.com/)

	1984		1985		1986		1987		1990	
	Elo	Strength	Elo	Strength	Elo	Strength	Elo	Strength	Elo	Strength
Kasparov	2715	2754	2700	2760	2740	2758	2740	2748	2800	2756
Karpov	2705	2754	2720	2749	2705	2758	2700	2755	2730	2754

Table 6: An analysis of world championship matches between Kasparov and Karpov.

In 1992, Fischer returned to play an unofficial rematch against Spassky. Apart from the unconventional circumstances in which the match took place, there was some expectation with regard to Fischer's performance since he had been away from competition for twenty years and did not even have an Elo rating anymore. Having a look at the games, Kasparov, the world champion at the time, is quoted as having said: "Maybe his [Fischer's] strength is around 2600 or 2650."[7]. An analysis based on Algorithm 4 of the 30 games that were played in that 1992 rematch suggests that Fischer performed with an estimated strength of 2731.

Meanwhile, in 1997, Kasparov himself was involved in a different kind of rematch, in this case against IBM's computer DEEP BLUE (Campbell, Hoane, and Hsu, 2002). Kasparov had won a previous 6-game match against the machine in 1996 by a score of 4–2, and was now playing a second match against a much stronger version. Kasparov won the first game, lost the second, drew the following three games, and eventually lost in the sixth and final game, allowing IBM to claim that DEEP BLUE defeated the world chess champion. An analysis by Algorithm 4 of the 6 games played in this 1997 rematch suggests that Kasparov played with an estimated strength of 2754 (the same level of strength as in the Kasparov-Karpov matches), while DEEP BLUE performed with an estimated strength of 2741. This seems to be an indication that Kasparov played better than DEEP BLUE, a conclusion that finds support in claims that Kasparov could have drawn game 2 and could have won game 5 at some point. Yet, a direct computation of the expected score between Kasparov and DEEP BLUE according to Algorithm 1 gives a rating difference in the opposite direction, a difference of just 3 Elo points in favour of DEEP BLUE.

9. RELATED WORK

The results presented in the previous section should not be taken lightly and should be confirmed (or denied) by far more evidence than we are able to develop here. A more in-depth comparison of player performances should also take into account the complexity of positions arising over the board during the course of the game or match. After all, in some positions there are only a few legal moves and there is little room for going wrong, while in other positions it becomes very hard to make the correct choice among a set of possible moves. Guid and Bratko (2006) have taken this factor into account by determining the *complexity* of a position based on the difference between the best move evaluation and the second-best move evaluation. Regan and Haworth (2011) also considered that when a move has a clear standout evaluation, it is more likely to be selected by a player than when it is given slight preference over several alternatives. This kind of considerations are missing here, but they were not part of our original purpose, which was to develop an approach for estimating strength based solely on the distribution of gain for each player. However, the analysis that we performed here with HOUDINI 1.5a 64-bit at depth 20 should be more accurate than that of either Guid and Bratko (2006) or Regan and Haworth (2011), who used CRAFTY with depth 12 and RYBKA with depth 13, respectively.

Guid and Bratko (2006) were the first to develop a systematic and extensive study on the strength of chess players based on computer analysis. Their approach can be summarized as follows.

- A modified version of the engine CRAFTY was used to perform the required analysis at a fixed search depth of 12 plies (13 plies in the endgame).
- Analysis started on the 12th move to try to reduce the effect of opening theory/preparation.
- Moves with an evaluation outside the range $[-2, 2]$ pawns were discarded to allow suboptimal play in won or lost positions.
- The main criterion for comparing players is a single statistic designated as *mean loss* (equivalent to the average of gain across a number of moves).

[7] In several sources across the Internet, this quote is reported to having first appeared in a book by Fred Waitzkin.

- A second criterion for comparing players takes into account the complexity of each position. It was observed that when the complexity is low the players tend to agree with the "best" move chosen by the engine (40%–60% of the times), whereas for the most complex positions there is virtually no agreement (less than 5%). The distribution of best move percentage over a range of complexity values was used to characterise each player's *style*. In turn, this distribution was used to calculate the expected best move percentage in a hypothetical scenario where all players would face the same position complexity.
- In an analysis of all world championship matches, Capablanca came on top according to the main criterion (mean loss), closely followed by Kramnik. On the second criterion (similar position complexity), it was Kramnik who came on top, closely followed by Capablanca. Here, the results suggest that Capablanca had a style of play that favoured positions of comparatively lower complexity.

The fact is that Guid and Bratko (2006) joined two goals in a single work, by (1) developing their approach and (2) also trying to find the strongest world chess champion of all time. While the approach itself was well received, the results regarding the world champions faced a number of criticisms, namely that the engine was too weak to perform such analysis and that the number of positions was too low to characterise the players. The same authors addressed those criticisms in a second publication (Guid, Perez, and Bratko, 2008) where they showed how a weaker engine could still provide a sensible ranking of stronger players. In Guid and Bratko (2011) the authors carried out the same experiments with stronger engines (SHREDDER and RYBKA) but with the same search depth of 12 plies. The results were very similar to those obtained with CRAFTY.

In connection with the work by Guid and Bratko (2006) , we highlight the following differences to our own work.

- We used HOUDINI 1.5a which is one of the strongest engines currently available (only superseded by HOUDINI 2.0c at the time of this writing), with a search depth of 20 plies.
- We carried out the analysis from move 1 (actually, analysis begun with the starting position, so that the gain of the first move could be determined). With respect to opening theory/preparation, we consider it to be an integral part of strength of play.
- With regard to suboptimal play in won/lost positions, we considered that playing strength should be determined based on whatever a player brings about on the board. If a player plays suboptimally, and this changes the position evaluation, then it should have an effect on the perceived strength of play.
- Our method for comparing players is based on the distribution of gain expressed as a histogram with the maximum possible resolution (in this work, we used intervals 1 centipawn). This is in contrast with other approaches that consider only one or two parameters (e.g., mean loss, or mean and standard deviation).
- Rather than an expected error or best move percentage, our method provides an expected score between players that can be directly translated into an estimated rating difference. To the best of our knowledge, this is the first time that computer analysis can be translated directly into the Elo-rating scale.
- There is no consideration whatsoever of the complexity of the positions. This is perhaps one of the few issues in which the present approach is lacking in comparison with both Guid and Bratko (2006) and Regan and Haworth (2011).

We did not venture into an analysis of world champions as in Guid and Bratko (2006) because the world championship matches may not be fully representative of a player's repertoire, and it is remarked that the strength of players varies over time, which makes it difficult to determine when a player was at his/her peak. Instead, we looked at the performance of selected players in some recent tournaments (Section 5) and historical events (Section 8). Since our estimated strength is expressed in terms of a perceived Elo rating, it is possible to bring players from different eras to comparable terms. Also, we observed an effect that can be possibly related to the phenomenon of ratings inflation (last row in Table 5). However, once again we warn the reader about the assumptions that underlie the present approach – namely the approximation to the engine's distribution of gain, and also the uncertainty in the estimated engine strength –, which can make the analysis presented in Section 8 rather speculative in nature. For these reasons, we believe that these approaches still have to be further improved and developed before a definite assessment of world champions can be carried out.

Another branch of related work can be found in Regan and Haworth (2011), Regan, Macieja, and Haworth (2011), and Regan (2012) , where the authors ran an analysis on an extensive set of games using RYBKA 3 at depth 13. Their approach, however, is slightly different. We mention five differences.

- They considered the evaluation of several possible moves for each position. Each possible move m_i has an evaluation score $e(m_i)$ and moves are ordered by decreasing evaluation, so that m_0 is the move with the highest score $e(m_0)$, and $e(m_{i+1}) \leq e(m_i)$.
- A player is represented by a model which provides the probability p_i of the player choosing a move which is $e(m_0) - e(m_i)$ below the evaluation $e(m_0)$ of the best move m_0. The model is based on two main parameters: the sensitivity s and the consistency c.
- From the analysis of more than 150,000 games of rated players, values for s and c have been found for players with 2700, 2600, 2500, etc., down to 1600 Elo. This was done by fitting the model parameters to the results obtained from the analysis with RYBKA 3 at depth 13.
- In Regan and Haworth (2011) the authors conclude that there is a relationship between Elo ratings and their model parameters s and c, which suggests that these are able to capture a measure of playing strength. However, the relationship is such that it also suggests that there has been ratings deflation, rather than inflation, over the years.
- It was only in Regan *et al.* (2011) and Regan (2012) that the authors established a correspondence between their model parameters and an estimated strength, which they call *Intrinsic Performance Rating* (IPR). Again, it was obtained by data fitting over the analysis performed with RYBKA 3 at depth 13.

With regard to the work by Regan and Haworth (2011) , we can make similar remarks as we did in connection to the work by Guid and Bratko (2006). Namely, that we performed an analysis with a stronger engine running to a greater depth; that we characterize players based on their full distribution of gain rather than on a model with one or two parameters; and that our method provides a direct measure of expected score and rating difference between players, without the need for data fitting as in Regan and Haworth (2011) . However, in order to provide an estimate of player strength in terms of an absolute Elo rating, we had to calibrate the approach on a set of players with known ratings, and we did that in step 4 of Algorithm 2. This means that the results can be shifted up or down by an arbitrary amount, but the difference between the strength of players will be maintained.

10. CONCLUSION

In this work we have developed an approach to estimate the strength of players based on the difference in position evaluation (gain) before and after each move.

One of the immediate problems we had to deal with is how to define the position score when there is a forced mate on the board; we addressed this problem in Section 3.

In Section 4, we presented the foundation for the proposed approach by explaining how to compute the expected score and rating difference between two players based on the cross-correlation of their distributions of gain. With such rating difference, it is possible to estimate the strength of a player in terms of a perceived Elo rating, if the rating of the opponent is known.

In order to work with known ratings, we turned to the analysis of some recent tournaments in Section 5. From the known Elo ratings and an analysis of gain per move, we were able to estimate the strength of players in terms of a perceived Elo rating. We calibrated the results so that the average of the perceived Elo ratings coincides with the average of the actual Elo ratings.

In Section 6 we used a similar approach to estimate the strength of the chess engine itself. In this case, the perceived Elo rating for the chess engine is calculated as an average of the rating difference towards a set of players with known ratings. Since the rating difference is calculated from the cross-correlation of distributions of gain, we need to know the gain per move for the engine. Following the rationale explained in Section 6, we approximate this distribution by a constant zero gain.

Having determined the engine strength, this can be used to estimate the strength of players on a move-by-move basis, during the course of a game. Even though such an estimate may be imprecise, we illustrate the approach in the analysis of some historical events in Section 8. An analysis based on the distribution of gain per move can bring players from different eras to comparable terms. The results can be surprising, and Section 8 just scratches at the surface of what can be done with this kind of approach.

The main contribution of this work can be found in the relationship that was established between gain per move and expected score in Section 4. Basically, this bridges the gap between two different domains by providing a means to translate quality of play into a perceived Elo rating. In future work, we will look into finding a more reliable estimate for engine strength, namely by characterising its actual distribution of gain. This, however, will become a lesser problem as computing power and analysis depth are continuously increasing.

11. REFERENCES

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs, I. the method of paired comparisons. *Biometrika*, Vol. 39, No. 3/4, pp. 324–345.

Burgess, G., Nunn, J., and Emms, J. (2010). *The Mammoth Book of the World's Greatest Chess Games.* Running Press.

Campbell, M., Hoane, A. J., and Hsu, F.-H. (2002). Deep Blue. *Artificial Intelligence*, Vol. 134, Nos. 1–2, pp. 57–83.

Coulom, R. (2008). Whole-History Rating: A Bayesian Rating System for Players of Time-Varying Strength. *Computers and Games*, Vol. 5131 of *Lecture Notes in Computer Science*, pp. 113–124. Springer.

Dangauthier, P., Herbrich, R., Minka, T., and Graepel, T. (2008). TrueSkill Through Time: Revisiting the History of Chess. *Advances in Neural Information Processing Systems 20.* MIT Press.

Elo, A. (1978). *The rating of chessplayers, past and present.* Arco Pub.

Glickman, M. E. and Doan, T. (2011). *The USCF Rating System.* United States Chess Federation.

Guid, M. and Bratko, I. (2011). Using Heuristic-Search Based Engines for Estimating Human Skill at Chess. *ICGA Journal*, Vol. 34, No. 2, pp. 71–81.

Guid, M. and Bratko, I. (2006). Computer analysis of world chess champions. *ICGA Journal*, Vol. 29, No. 2, pp. 65–73.

Guid, M., Perez, A., and Bratko, I. (2008). How Trustworthy is Crafty's Analysis of Chess Champions? *ICGA Journal*, Vol. 31, No. 3, pp. 131–144.

Herbrich, R., Minka, T., and Graepel, T. (2007). TrueSkill: A Bayesian Skill Rating System. *Advances in Neural Information Processing Systems 19.* MIT Press.

Howard, R. (2006). A complete database of international chess players and chess performance ratings for varied longitudinal studies. *Behavior Research Methods*, Vol. 38, pp. 698–703.

Hsu, F.-H. (2002). *Behind Deep Blue: Building the Computer that Defeated the World Chess Champion.* Princeton University Press.

Levy, D. N. L. and Newborn, M. (1990). *How Computers Play Chess.* Computer Science Press.

Regan, K. (2012). Intrinsic Ratings Compendium (Working Draft). Department of CSE, University at Buffalo.

Regan, K. and Haworth, G. (2011). Intrinsic Chess Ratings. *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, AAAI Press.

Regan, K., Macieja, B., and Haworth, G. (2011). Understanding Distributions of Chess Performances. *Proceedings of the 13th ICGA conference on Advances in Computer Games*, Tilburg, The Netherlands.

Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach.* Pearson, 3rd edition.

Shannon, C. E. (1950). Programming a Computer for Playing Chess. *Philosophical Magazine*, Vol. 41, No. 314.

von Neumann, J. (1928). Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, Vol. 100, pp. 295–320.

Waterman, M. and Whiteman, D. (1978). Estimation of probability densities by empirical density functions. *International Journal of Mathematical Education in Science and Technology*, Vol. 9, No. 2, pp. 127–137.