Rating the proficiency of Baduk/Go players

A. Cieplý, Prague, Czechia

I Introduction

Many Baduk/Go[1] players are interested how their strength relates to the strength of other players, either their direct or potential opponents. There is also some interest in calibrating a performance of top players and in establishing their relative strength. In many competitive sports the national and governing international organizations adopted various rating systems that fulfill these tasks.

Tournament Go can be viewed in close resemblance with another mind sport: western chess. In chess, the rating system was developed by A. Elo in 1950s and accepted worldwide in the following two decades [1]. Today, it is impossible to imagine tournament chess without *ELO ratings*. Although similar rating systems were introduced in Go by the American Go Association, European Go Federation (EGF), and by Chinese Weiqi Association (only for professional players) a worldwide accepted Go ratings are still missing. A kind of international rating scale is used at internet Go servers but it also varies from one server to another, is not suitable for tournament play and is often criticized for its instability. We would like to stress that any rating system applied to Go should relate to the existing scale of professional and amateur grades tied to a system of handicaps used to equalize the odds favouring the stronger player in a game with a weaker opponent. Historically, the amateur grades are distributed in steps about equal to a difference of one handicap stone while the distribution of professional grades is more dense (approximately one handicap stone per 2.5 grades). Unification of both systems into one scale is problematic as there are not many regular games between the two groups of players. In addition, the grade and promotion systems used by various Baduk/Go associations exhibit significant regional differences.

In this report we present the basic ideas of the rating system used by the EGF and discuss its possible extrapolation towards a uniform scale of professional and amateur ratings. We also show some statistical analysis of the data collected at tournaments

[1] Usually we use the simple word *Go* for the name of the game, sometimes replacing it by *Baduk/Go* to make a clear distinction between a game and other meanings of the English word "go". It neither means that we prefer the Japanese name of the game nor that we disregard the name Wei-Qi used in China.

included in the ratings database.

II Rating Theory

The rating system is derived from the ELO rating system used by the International Chess Federation (FIDE). It is based on the idea that one can define a probability of winning a game (so called winning expectancy) P_{xy} depending on the difference of opponent's ratings $D_{xy} = R_x - R_y$. Here P_{xy} is the probability for player X scoring a win over player Y. Let us assume a third player Z with rating R_z. Then the odds of player X to score over Z (win/loss ratio) are

$$P_{xz}/P_{zx} = (P_{xy}/P_{yx})(P_{yz}/P_{zy}) \tag{1}$$

and the corresponding rating difference is

$$D_{xz} = D_{xy} + D_{yz} \quad . \tag{2}$$

The equations (1) and (2) are easily satisfied for P_{xy}/P_{yx} taken in exponential form

$$P_{xy}/P_{yx} = e^{\lambda D_{xy}} \quad , \tag{3}$$

where λ stands for arbitrary constant. If we do not consider *jigos* for a moment, then $P_{yx} = 1 - P_{xy}$ and we get

$$P_{xy} = 1/(1 + e^{-\lambda D_{xy}}) \tag{4}$$

which is the well known logistic function. The parameter λ is to be related to a rating scale. A typical behaviour of $P(D)$ is shown in Figure 1 where the probability was calculated with the parameter λ fixed at the value $\lambda = 1/100$. This setting gives about 27% probability for beating a one grade stronger opponent if one assumes 100 rating points for a proficiency difference corresponding to one amateur grade. It is also consistent with statistical observations giving a small (if not negligible) chance of beating an opponent more than four grades stronger.

We further fix the rating scale by assuming that an average 1 dan player (amateur) should have a rating R(1d)=2100.[2] This gives R(20k)=100 for a rating of 20 kyu, the

[2]Making a distinction between the professional and amateur grades we use 1p for the professional *shodan* and 1d for the amateur grade. A similar notation applies to the higher grades. If not specified otherwise any further reference to dan grades/players will refer to amateur ranks.

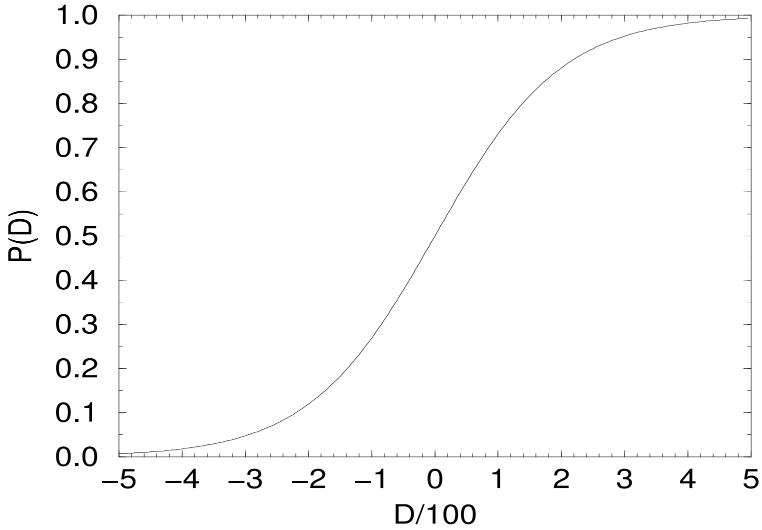

Figure 1: Expected winning probability dependence on the rating difference D.

lowest amateur grade in most European countries. Although, the ratings can in principle be even lower (and reach values below zero), the strength of such lowly rated players cannot be reliable and it is quite plausible to set the rating bottom to R(20k).

Example 1: The player with Go rating $R = 2050$ can be regarded either as a weak 1d or a strong 1k.

Example 2: The ratings of top European amateur players (comparable in strength with lower professional grades) reach values about $R = 2760$.

Since stronger players play more consistently than the weaker ones, the probability of beating a one grade weaker opponent tends to rise with the player's grade. This fact can be reflected in the model by an appropriate dependence of parameter λ on rating. The specific form of this function is to be determined by the available statistical data and shall be discussed below.

In a single even game the rating of a player changes by

$$R_{\rm new} - R_{\rm old} = C(S - P) \quad , \tag{5}$$

where S is the achieved result ($S = 1$, 0 or 0.5 in case of jigo), P stands for the winning expectancy given by Eq.(4) and the factor C characterizes a magnitude of the variation. Since the performance of top players is more stable than the performance of weaker players it is essential to require a higher stability of the ratings at the top and allow for larger rating variations in the region of lower ratings. This means that the parameter C should be a decreasing function of the rating.

The complete setting of our model is shown in Table I where the corresponding probabilities of beating a one grade stronger opponent are given as well. As one can see, 20 kyu is expected to win about 40% games with one grade stronger opponents while the top amateur players should win only 20% of their games with 100 rating points stronger opponents.

The system also allows the inclusion of handicap games, assuming that the rating difference D is reduced by $100(H-0.5)$, where H is the number of given handicaps. Note, that it can happen that the winning expectancy P of a weaker player is larger than the one (equal to $1-P$) of the stronger player (i.e. the weaker player is expected to win the game) if the number of given handicaps (reduced by 0.5) is larger than the absolute value of $(R_x - R_y)/100$.

Table I: The dependence of parameters C and $a = 1/\lambda$ on the rating R. For convenience the winning expectancies P_1 (in percents) for beating one grade (= 100 points) stronger opponent are shown as well. We use $C = 10$ and $a = 70$ for $R > 2700$.

R	C	a	$P_1[\%]$	R	C	a	$P_1[\%]$	R	C	a	$P_1[\%]$
100	116	200	37.8	1000	70	155	34.4	1900	31	110	28.7
200	110	195	37.5	1100	65	150	33.9	2000	27	105	27.8
300	105	190	37.1	1200	60	145	33.4	2100	24	100	26.9
400	100	185	36.8	1300	55	140	32.9	2200	21	95	25.9
500	95	180	36.5	1400	51	135	32.3	2300	18	90	24.8
600	90	175	36.1	1500	47	130	31.7	2400	15	85	23.6
700	85	170	35.7	1600	43	125	31.0	2500	13	80	22.3
800	80	165	35.3	1700	39	120	30.3	2600	11	75	20.9
900	75	160	34.9	1800	35	115	29.5	2700	10	70	19.3

Example 3: Both opponents have the same rating $R_x = R_y = 2400$. This gives $D_{xy} = 0$ and $P_{xy} = P_{yx} = 0.5$. If player X wins, his new rating will be

$R_{\text{new}}(X) = 2400 + 15(1 - 0.5) = 2407.5$

At the same time, the rating of player Y drops by 7.5, i.e.

$R_{\text{new}}(Y) = 2392.5$

Example 4: $R_x = 320$, $R_y = 400$ and player X wins:

$$a = 189, \quad P_{xy} = 0.396$$

$$R_{\text{new}}(X) = 320 + 104(1 - 0.396) = 383$$

$$R_{\text{new}}(Y) = 400 + 100(0 - 0.604) = 340$$

Example 5: $R_x = 1850$, $R_y = 2400$, player X takes 5 handicaps and wins:

$$D_{xy} = -100, \quad a = 90, \quad P(-100) = 0.248$$

$$R_{\text{new}}(X) = 1850 + 33(1 - 0.248) = 1875$$

$$R_{\text{new}}(Y) = 2400 + 15(0 - 0.752) = 2389$$

III Practical Implementation

The rating system was adopted by the Czech Go Association at the beginning of 1998. Originally it was designed to serve only the needs of Czech Go community. Later we decided to enlarge the tournament database by including other European tournaments and made it comparable with the former EGF database. The system has been used for computing the official EGF ratings since November 1998. Now the database consists of almost 1500 tournaments (including some play-off matches) held since the beginning of 1996. They are divided in three categories (according to the importance of the event) that imply the weight with which the event is counted. Fast games (less than 30 minutes) are not considered because their results are less consistent. We allow for handicap games since their inclusion helps to keep the correspondence between ratings and grades.

Every new player enters the system at a rating corresponding to his grade, and the rating development of all players is stored in the database. If a player's rating drops below 100, it is reset to $R = 100$ which is fixed as the bottom value. If a rank professed by the player has improved significantly (at least by 2 grades for amateur players or by 1 professional grade) with respect to the highest previously professed rank, the rating of the player is reset. This measure helps to deal with fast improving players and with players who participate at included tournaments only occasionally. To avoid undesirable oscillations in the bottom part of the rating list, the drop of a player's rating at one tournament is restricted to 100 points.

Depending on players' tournament results and on differences between ranking systems used in various countries, the correspondence between grades and ratings may not work quite well especially for lower kyu grades. However, it gives a relatively good measure of strength if the player has participated in at least 3-5 tournaments. If a player has not participated at any considered tournament for some time (this period is set to 2 years for dan players, 12 months for 1-10 kyu, and 6 months for 11-20 kyu), then the player's rating is no longer published. However, the rating is kept in the database and is used once the player appears again at any tournament in the future.

When including tournament results, the ratings are not reevaluated after each game (round) and the "new" ratings are computed from the "old" ones by summing all contributions from games that the player completed at the event. In other words, we assume that the ratings of players do not change during one tournament.

IV Testing the System

At the moment (April 2001), the database of EGF ratings includes about 140000 games, 12.7% of them being handicap games. There are over 9000 players listed in the database and about 4000 of them appear in the current rating ladder. This large data set is particularly useful for statistical analysis and for various tests of our rating model.

A detailed analysis of current EGF ratings with respect to grades professed by British and other European players was carried out by G. Kaniuk [2]. His fit of individual player ratings performed for players listed in the EGF rating ladder produced the relation

$$R = 96.7\, g + 2061.3 \quad , \tag{6}$$

where the variable g denotes the grade professed by the players in a scale fixed at $g = 0$ for 1d. This result is satisfactorily close to the nominal scale $R = 100\, g + 2100$. Especially, the coefficient of the term linear in g indicates that the grades professed by European players are fairly well distributed according to the relative strength of players. The lower value of the constant term either means that ratings are deflationary or it simply reflects a well known fact that players tend to profess grades slightly above their real proficiency. The latter case is often supported by inflationary promotion policies in some countries. Another analysis on average rating-grade differences performed by the author [3] for each country reveals that such statistics exhibit significant regional differences. In general, players from countries with strict and well designed promotion systems profess grades corresponding to relatively higher ratings.

Another analysis presented in Ref. [4] was driven by data collected on winning percentage dependence on the grade of lower ranked player [5]. In this statistics we consider separately the games between opponents with grade differences equal to n, $0 < n \leq 4$. In average, the rating difference of the players is $D_n = -100\,n$. The related winning probability P_n is given by Eq.(4). From the observed statistical probabilities one can infer the corresponding values of parameter λ,

$$\lambda = -\frac{1}{D_n} \ln \left(\frac{1 - P_n}{P_n} \right) \quad . \tag{7}$$

In Figure 2 the collected data are plotted against the grade of the weaker player. It is remarkable how well the data points obtained for various rating differences agree. This feature can be viewed as a confirmation of the particular form (the logistic function) adopted for winning expectancy in Eq.(4).

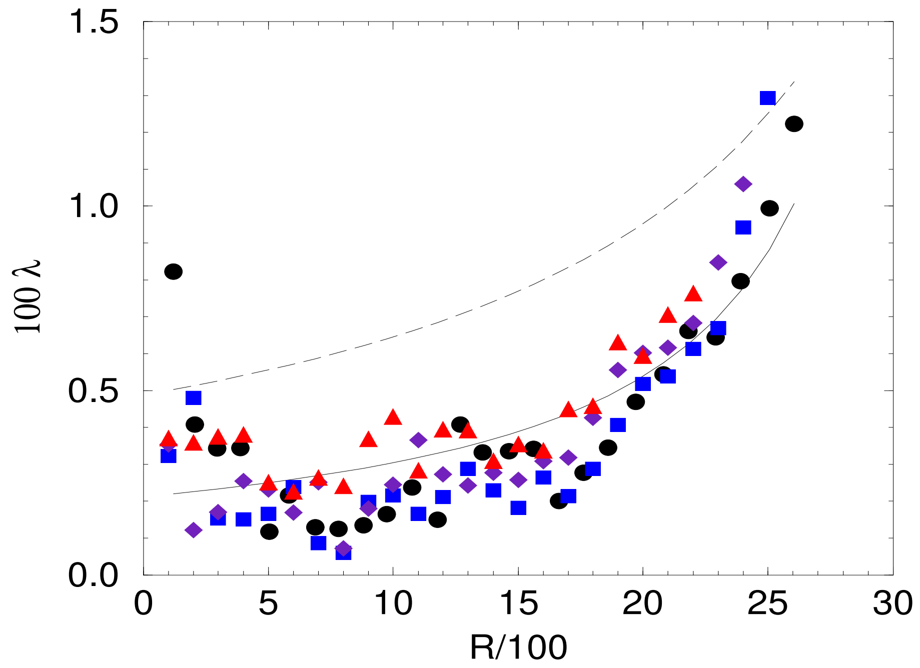

Figure 2: The dependence of parameter λ on rating R. The data points visualize the data collected from statistics on games with one grade difference (circles), two grades (squares), three grades (diamonds) and four grades (triangles) differences. The dashed line corresponds to the current parametrizarion of EGF rating system and the full line shows the parametrization fitted to the data.

Another point worth further discussion is the observed trend when going from lower grades to the higher ones. First, one should disregard the data in the region of $R < 500$, where new players with unreliably estimated grades enter the database and where the results are also affected by other factors (like the existence of rating/grade bottom) that may generate systematic deviations from the characteristics seen at higher ranks. In general, the data confirm the expected behaviour of parameter λ mentioned in Section II. A relatively steep increase of λ observed in region of dan players ($R \geq 2100$) corresponds to a significant drop of the probability to win a game with stronger opponent. An interesting conclusion can be drawn if λ is assumed in a hyperbolic form of the $1/R$ type, i.e.

$$\lambda = \frac{A}{B - R/100} \quad . \tag{8}$$

The parameters A and B can either be varied to achieve the best fit of the observed data or set to specific values determining a probability dependence intended within a framework of a given rating model. Extrapolating the observed dependence to higher ratings one notes that the adopted function (8) is singular at $R/100 = B$. For ratings very close to this point λ goes to infinity. In this limit the corresponding winning expectancy is zero even for very small rating differences. In other words, a hypothetical player with rating $R = 100\,B$ is expected to have a perfect score with lower rated opponents regardless how small the rating differences are. The proficiency of such player would be absolutely consistent and we can consider it as another definition of *Baduk/Go God*.[3]

The dashed line in Figure 2 corresponds to the parametrization used in EGF rating system ($A = 0.2$, $B = 41.0$) and specified in our Table I. A good fit to the data is achieved for $A = 0.07$ and $B = 33.0$ defining the λ dependence visualized by full line in the figure. Another type of parametrization assuming a more general exponential form instead of Eq.(8) was used in Ref. [2] and is not discussed here.

The discrepancy between the observed data and the parametrization used in our model (EGF rating system) implies a corresponding difference between the observed and expected winning probabilities. In Figure 3 we show a comparison of those quantities presented separately for dan and kyu players. The full line of ideal relation P(statistical) $=$ P(expected) is also drawn in the figure to guide reader's eye. It is seen that the variation represents some $3 - 6\%$ for dan players while more significant deviations are observed in the region of kyu players. The figure also shows that players in both groups

[3] Baduk/Go God is often regarded as a perfect player making the best moves available in any position on the board.

outperform (score better than expected) their higher ranked opponents. This feature is in accordance with the fact that a fraction of (fast) improving lower rated players is higher than the corresponding fraction of higher rated players. One should also take into account that the McMahon tournament system used at the vast majority of EGF events gives a preference to pairing lower rated players performing above average against higher rated opponents whose performance is below the average. Therefore, it is not so surprising that lower rated players do better than expected, as is reflected in our rating model.

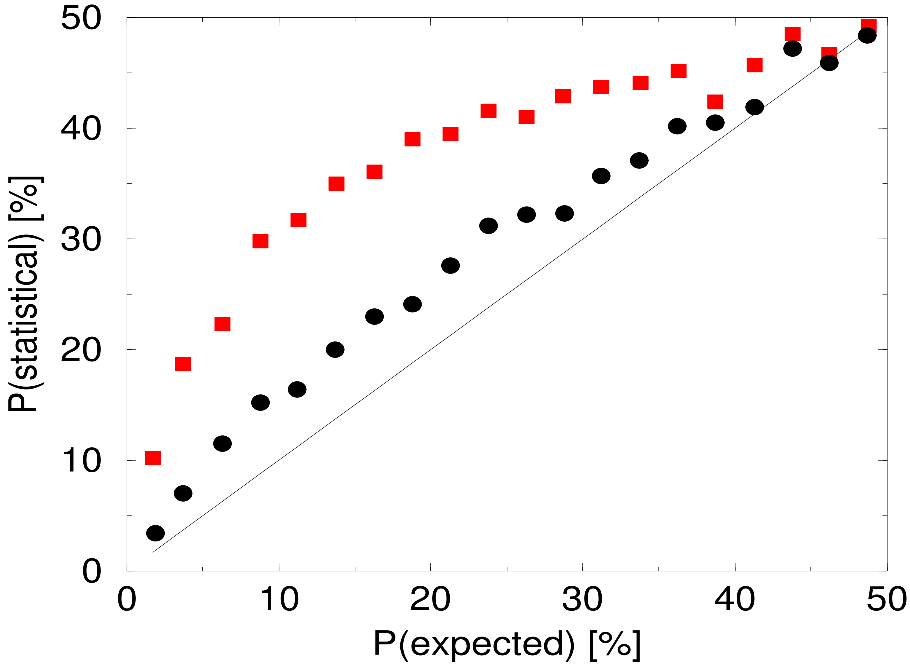

Figure 3: A comparison of statistical winning probabilities with the winning expectancies assumed by the rating model. The circles and squares visualize the data points for dan and kyu players, respectively.

One of the main problems inherent in any ELO-like rating system is a slow deflation of the ratings due to the incoming flux of new players. This process can be mediated by various means not discussed here. We just mention that the setting of EGF rating system described above guaranties that the total sum of all ratings is increasing in any game won by a lower rated player. This is another reason why the *EGF parametrization* adopted for λ is preferred to fitting the statistical data.

Finally, we briefly turn our attention to statistics on handicap games. It is accepted [6] that one handicap stone has a value about twice as large as *komi* used to equalize the advantage of the first move in even games. Since white plays first in handicap games he should also give komi to even the win/loss odds in an H-handicap game against H grades weaker opponent. However, the handicap games are traditionally played without any komi, which should favour the white player. This advantage is about equal to a half-stone (half-grade) difference, i.e. to about 50 rating points. Our treatment of handicap games mentioned in Section II takes it into account and white is expected to win about 60% of all games.

Table II summarizes statistical results obtained in over 10000 handicap games played with the handicap H equal to grade difference of the opponents. The statistical probabilities agree quite well with the expected value. It is remarkable that the data show no dependence on the number of handicap stones H. This can be viewed as confirmation of the theoretical assumptions stated above. Unfortunately, the statistical sample is not large enough for any more detailed analysis like a dependence on grade of the weaker player.

Table II: Statistical probabilities $P(H) = N_W/N_G$ achieved in N_G handicap games, N_W of them won by black. The data are presented separately for different numbers of handicap stones H, $0 < H \leq 9$.

H	1	2	3	4	5	6	7	8	9
N_W	1526	949	541	379	278	174	150	122	109
N_G	3684	2366	1343	890	690	443	383	288	255
$P(H)$	41.4	40.1	40.3	42.6	40.3	39.3	39.2	42.4	42.7

V Toward the Uniform Rating Scale

The present rating model can easily be extended to include professional games. We assume that 1p is about equal in strength to average European 7d, i.e. R(1p)=2700. Since an average 9p player can give about 2.5 handicap stones to an average 1p player, the professional grades are to be separated by about 30 rating points. Of course, this setting represents just a reasonable first guess and has to be tested in practise. This would require inclusion of many professional games in the database. In principle, one could include the many available game results going back in history maybe as far as times of Dosaku, and compare the proficiency of players past and present. In the near future, it may not look so difficult if the rapidly evolving databases of professional games are used as a source.

Although not many games are played between European amateurs and professional Go players there is one good example for testing the extrapolation of our rating model towards professional ratings. Guo Juan, a top European player and a regular participant at EGF tournaments used to be 5p in China years ago. Her EGF rating oscillates around 2760, a value we set for an average 3p player. Considering that Guo Juan may not exhibit her full strength in games with amateur players, it seems that our setting of professional ratings cannot be too much off. Unfortunately, the European professional players Catalin Taranu and Hans Pietsch are not active in Europe, so we cannot conclude anything from their ratings.

In an attempt to make some further tests we included the results of big international pro-events (Ing Cup, Samsung Cup, LG Baduk World Cup, Fujitsu World Cup, Chunlan Cup and a few others). The computed ratings are not very reliable as the number of included games is too low. However, one can get at least an idea about the proficiency of top world players. The list is headed by Yi Chang-ho[4] at R=3030, followed by Yu Chang-hyeok, Ma Xiaochun and Cho Hun-hyeon, all of them with ratings around 2970. The best Japanese player on the list is O Rissei at R=2960. A comparison with the hypothetical perfect player introduced in a previous section indicates that the Baduk/Go god may give about 3-4 handicap stones to the best human players, while they themselves are about 9 stones stronger than European 1d players. Of course, this should be regarded as a first approximation, and more detailed analysis performed on a much larger sample of professional games is needed to improve our knowledge.

VI Final Remarks

We have presented a rating model attempting to measure the strength of Baduk/Go players on a plausible theoretical basis supported by various statistical data. Many players tend to accept the rating system in so far as their ratings agree with their own subjective estimates. Subjective views may have some grounds but they are hardly proper tests of any rating model. The valid test lies in the success of qualitative predictions, in forecast

[4] We use McCune-Reischauer transliteration for Korean names.

of the scores of tournament games. The collected statistical data presented in this report seem to indicate that the rating model works sufficiently well and can be used as a reliable measure of players proficiency. It also looks as if both amateur and professional players can be accommodated in one rating scale making it suitable for uniform ranking of all Baduk/Go players.

Acknowledgements:
I am greatly indebted to L. Dvořák who wrote the computer code for the rating program. The present work would not be possible without his generous work, help and fruitful ideas. I would also like to thank G. Kaniuk for many valuable comments and discussions. Finally, we acknowledge the help of many people who have contributed to the ratings database by sending tournament results.

References

[1] A. Elo: *The Rating of Chess Players Past and Present*, Batsford Chess Books, 1978

[2] G. Kaniuk: *European ratings and UK grades*, British Go Journal 122 (2001) 40

[3] A. Cieplý: www page http://www.ujf.cas.cz/∼cieply/GO/rgdrep.html

[4] G. Kaniuk: *Report on UK ratings*, internal report written for BGA grading committee, 2000

[5] A. Cieplý: www page http://www.ujf.cas.cz/∼cieply/GO/statev.html

[6] public discussion on internet news conference rec.games.go and personal communication with J.-L. Gailly (1998)