The Go Ranking System of Ales Cieply

Franco Pratesi

Several federations of sports and games have adopted the Elo system for rating and ranking their players. Such being the case, we might appreciate that this system provides quantitative comparisons among various games and players. To this aim, we should keep its standard parameters unaltered and, in particular, the criterion of 76% winning probability between adjacent ranks. For Go however, the traditional handicap stones still represent the common basis for ranking, also after officially adopting Elo ratings. A direct proportionality between stone and point ranks has been assumed (with a 100-point rating difference corresponding to the interval between stone ranks), which seemingly has been confirmed by results of many games.

Let us examine some basic features of the particular Elo system, adapted to Go by Ales Cieply, which is now the basis of the European Official Ratings (EOR).

First, a discussion is needed on the limits of strength acknowledged in the EOR because it is not obvious how to fix them. Some uncertainty of this kind is typical of every application of the Elo system due to its 'open-ended floating scale'.

The bottom limit of the EOR has been fixed at 20 kyu. (Actually the system starts at 20k with 100 rating, so that ratings can be extended one step by assigning 0 to 21k, which may thus become the true starting point.) The choice is mainly justified by the fact that in collecting and processing results from Go tournaments this is the lowest rank to be acknowledged everywhere in Europe. In any case, if one extends the scale to lower limits, down to 30 or even 35k, the problem of players with strength lower than the last rank accepted would still exist.

The top limit of the EOR scale has been set at 7d, a reasonable choice for European Go players. Higher ranks may be needed for Asian pros, with their nine ranks separated by one third of a stone; such ranks have been provisionally set in the system at 30 point intervals, up to the highest value of 2940 for 9p.

However, in setting the top limit of the system, we may run into the philosophical problem of ranking the strength of God or, probably better defined, of the perfect player. This player of the 'first' rank is expected from game theory actually to exist for any two-player, zero-sum game with full information. With the standard Elo procedure, which assigns players to a lower rank when losing 76 percent of the games, fixing the 'second' strongest rank would not be easy either: by definition, a perfect player would win all the games against any weaker player!

Elo ranks seem thus hard to assign for both weakest and strongest Go players but in these ranges, stone-handicap ranks also have problems of their own. We can thus better limit our analysis to intermediate strengths, typical of ordinary players. An interesting question arises here about the correlation between handicap-stone and Elo ranks; at first sight it may seem to follow the linear dependence initially assumed, but matters become more complex if carefully examined.

Different from other cases, the Cieply system has implemented parameters that correct the winning expectation according to player strength. It is stated that 'strong players play more consistently than the weaker ones' and thus the system lets the winning probability decrease – against a 100 point stronger opponent – from about 0.40 at 20k to about 0.20 at 6d. This modified winning expectation is shown as Theory curve in Figure 1.

Now, thanks again to Ales Cieply, we have a way to check the basic assumptions of the system by using statistics of game results. As a matter of fact, a table of 'Statistics on winning probabilities in even games' (between players separated in ranking by one to four handicap stones) has been published in the web pages of the EGF at: http://www.european-go.org (go to News, European Official Ratings).

These statistics are very interesting and useful; however, the agreement is found to be unsatisfactory, as shown by the four curves in Figure 1, corresponding to one to four rank difference, as mentioned above.

The Theory curve, preliminarily fixed for a 100 point difference, is found better to agree with a 200 or 300 point difference.

The impression is that some mix-up may exist between 100 and 200 point ranks, the latter holding the 'original' meaning of 0.76 winning probability. We can then try and associate the Theory curve (for 100 point difference) with that experimentally observed for 200 point intervals (meaning here a two stone rank difference) but the agreement is still unsatisfactory.

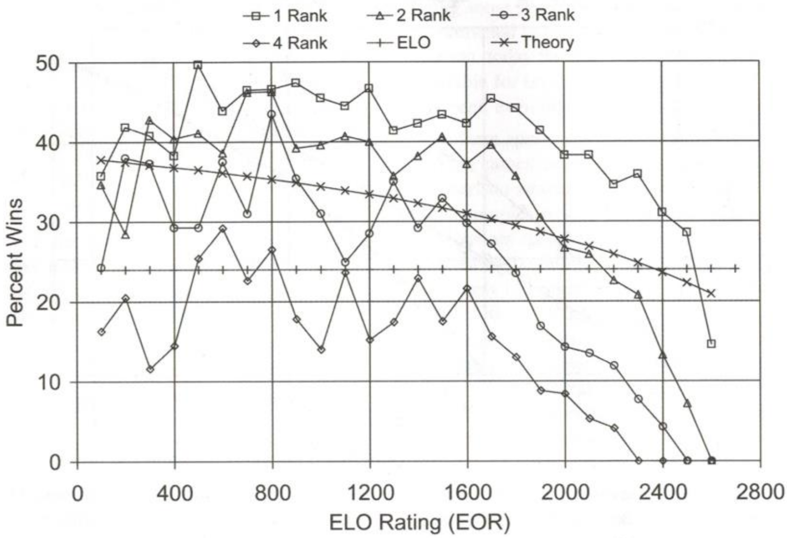

Figure 1 Winning expectation and statistics from actual games (from EGF)

The difference in slope observed with respect to the straight Elo line at 0.24, which would be found for 'original' Elo ranks, is much greater in the experimental than in the Theory curve. The need of a theoretical correction (with respect to a winning expectation independent of player strength measured in stone ranks) is confirmed, but its amount should actually be greater.

In particular, a much sharper decrease is observed at high strengths, where (with the approach to perfect play) even a single stone difference may be critical for winning expectations. We are apparently reaching a physical limit, where on the one hand the usual 'open-ended floating' scale of Elo ratings can hardly be extended or float and on the other hand smaller and smaller stone fractions would be needed for handicaps.

In the four experimental curves of Figure 1 some decrease is also unexpectedly observed in the range of the weakest players, but this should be confirmed by further data because here we find less games and overlapping with results from players weaker than 20k. Moreover, in this range some crossing is observed among the four curves, indicating that they are not fully reliable. These curves should at most reach (without any crossing, because a greater handicap does not plausibly lead to a better winning expectation) the 50% value, when players are so weak that the given stone handicap has no influence on the game result.

From the four curves, it is possible to transform the traditional stone or dan ranks into ‘original’ Elo (o-Elo) ranks, really corresponding to 0.76 winning probability. The single intervals between dan ranks are usually too small, except for the strongest players, and suitable o-Elo intervals correspond to multiple intervals of dan ranks, increasingly with decreasing player strength.

Let us try, with 21k as the starting point. In order to use a continuous scale, I have simply changed all kyu ranks into negative dan ones, after setting 1k at 0; thus 2k corresponds here to -1d, 3k to -2d, and so on, until 21k at -20d.

In order to keep the analogy with ratings encountered for Chess and Football, we may agree not to start here at 0, but already at 1000, leaving lower ratings, and ranks, for weaker players. Then suitable intervals of dan ranks can be derived so that our o-Elo ranks are both separated by 200 point intervals and correspond to 0.76 winning probability. In Figure 2, where the two kinds of ranks are compared (including the linear dependence assumed in the EOR-Elo system), we thus obtain curve New, with, for example, 211 = 1000, 10k = 1600, 1d = 2400, 6d = 3000. The number of subsequent ranks required is still greater than for Chess, but now the difference between the two major board games becomes less remarkable.

It would be possible to start from 21k set at 0, as for EOR, and a curve similar to Zero of Figure 2 would be obtained. Whichever the particular choice, it seems clear that the correspondence between dan and o-Elo ranks cannot obey a linear dependence over the whole range. A logarithmic function is more suitable to fit the points, as shown by the Log curve of the same Figure. For people loving analytic expressions (fortunately, they do exist), I can provide the following for the Log curve:

$$D = 23.882 \ln E - 185$$

with D for dan and E for o-Elo ranks. Now, I have reflected a long time before reaching it but, please, do not take this numerical approximation too seriously.

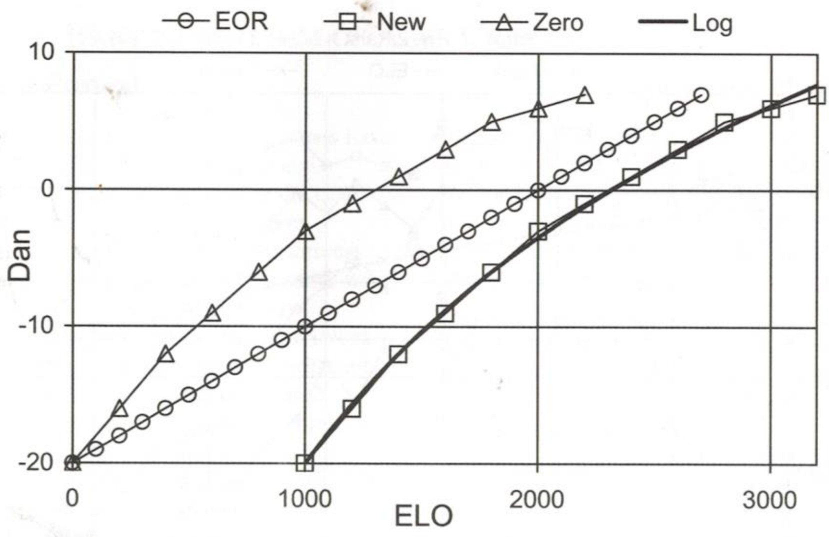

Figure 2 Correlation between Dan and Elo ranks

Of course, I would like to find out the correlation in a rigorous way, but for me this is a hard task. Among other problems, Elo ranks can either increase or decrease for any player in the course of time, whereas dan ranks tend to increase up to fixed values, because subsequent adjustments to lower ranks seldom occur. This may represent a further plus – in addition to the ability to compare different games directly – for using o-Elo instead of dan ranks.

We have studied systems used for ranking the players, but we could not avoid encountering some properties originating from the fundamental rules of the game. In my opinion, today there are two possible ways available for improvement, leading to different, even diverging, directions.

A uniform application of Elo ratings to various games can provide an immediate comparison among player strengths and game complexities. If this is the aim, fewer adjustments specific to the game should be inserted in the rating systems than actually are. I only have made a rough attempt above, so that Go could directly be compared with other games that use Elo ratings (Chess and Football in particular). This way, however, appears to be difficult because it would require a solid agreement among the experts of many federations – I guess I will have little to add here.

The other way is based on recognising that each game has its own properties to take into account, as much as possible. If this is the aim, Go has several characteristics that might be used, to begin with explicitly considering game scores, as we will possibly do in the next issue.