PROBABILITY OF WIN

Geoff Kaniuk *geoff@kaniuk.co.uk*

April 2011

Contents

List of Figures

List of Tables

1 INTRODUCTION

The probability of win is a very important parameter in any attempt to understand how well the various pairing algorithms perform, or how well a rating system measures a players strength, or how effective a tie-break is in ranking players. The game of Go is unique in that its handicapping system provides players with a built-in measure of relative strength. Although the grades of players assigned via club and tournament play are by no means precise, they do form a good starting point as is evidenced by the broadly linear relation that is seen in E.G.D's [1] correlation [2] of player's (more accurate) ratings [3] with grade.

The E.G.D system collects the win:games ratio for even games between players of different grades in every tournament [4]. This win ratio data is published for grade differences between 1 and 4 stones, and has now been collected for 14 years and includes half a million games.

The purpose of this article is to develop a model P_{win} for the probability of win between players of *any* strength and not just the integral grade differences necessarily collected by the system.

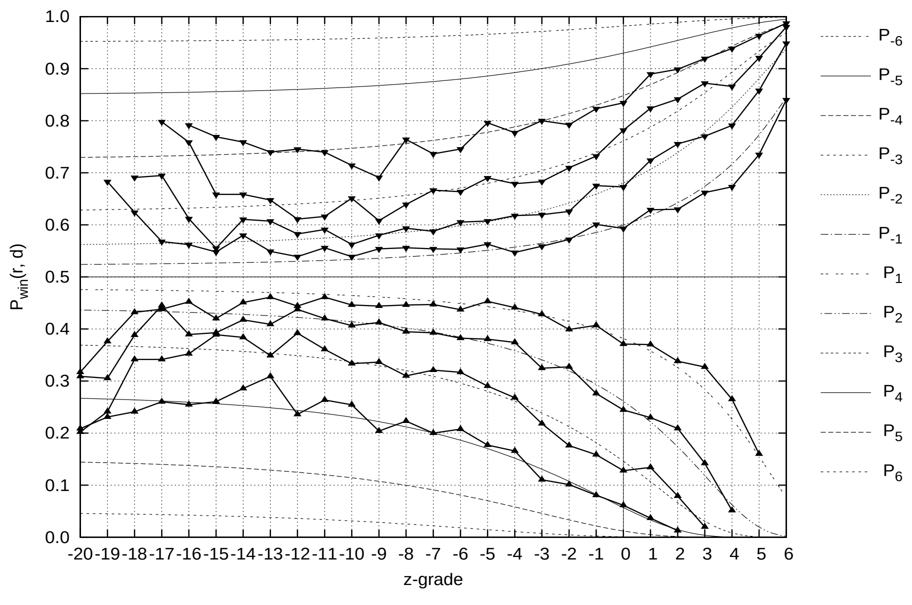

Figure 1: Raw data for the probability of win.

The above plot represents grade in *zero shodan* units, and the labels P_d identify the data for probability of win against a player d stones stronger when $d > 0$, or d stones weaker when $d < 0$. It shows that the measured probability of win is fairly well organised from about 12 kyu upwards, but below that the data would not easily be modelled. However, players below 12 kyu contribute less than 15% of of the total games to the statistics, and this is taken into account in the model.

In the next section we develop the general properties required of P_{win} leading to a clear statement of the form of the model. Subsequent sections deal with the the fitting procedure and the stability of P_{win} over time. In the final sections we present a graphical rendition of the probability of win between players of any rating and indicate some of the future work flowing from this presentation.

2 GENERAL PROPERTIES

In this discussion we assume that games between players are even (no handicap) and result in either a win or a loss, so there are no drawn (jigo) games. E.G.D publishes the win ratio $W(r, d)$ for games between players with grades r and $s = r + d$, where $d = 1 \cdots 4$, and the *weaker* player wins. As is evident from figure 1 there is missing data for games against 8 dan players. We do have games for players beating 7 dan players, so we are able to extend the E.G.D data to the case where 7 dan (or below) beat weaker players. The following shows how the extension is obtained.

Let a player's score be 1 for a win, and 0 for a loss. Define $P^{\gamma}(r, s)$ to be the probability distribution for a player with grade r scoring γ and the player's opponent with grade s scoring $1 - \gamma$. In this definition the point (r, s) lies in the region $\mathcal{R}$ covering the extended E.G.D data i.e. $\mathcal{R} = \{(r, s) | r = -20 \cdots 6, s = -20 \cdots 6\}$.

Identifying the win ratio as a probability of win, we have the following properties for P^{γ}:

$$\begin{aligned} P^1(r, r + d) &= W(r, d), \quad r = -20 \cdots 5, d = 1 \cdots 4 \\ P^0(r, s) &= 1 - P^1(r, s) \\ P^1(s, r) &= P^0(r, s) \end{aligned} \tag{1}$$

The second property follows from the definition of P^{γ} as a distribution, and the third from the requirement that a win for s is a loss for r. These properties show that the distribution $P^{\gamma}(r, s)$ is specified for all grades in $\mathcal{R}$, except for the case $r = s$, where we define $P^{\gamma}(r, r) = 0.5$ - the probability of win in an even game.

We wish to represent a player's strength as a continuous parameter in the model. In keeping with the Elo tradition followed by [3], we will base the probability distribution on the error function erf [5], i.e. we will assume that for all values of r and s the model has the general form :

$$p(r, s) \equiv P^1(r, s) = \tfrac{1}{2}[1 - erf(\Lambda(r, s))] \tag{2}$$

The properties expressed in equations (1) together with the reflection symmetry of erf imply that Λ is antisymmetric in r and s as shown in Appendix A.1. For the Elo system as used in Chess [6], the function $\Lambda(r,s) = Kd$, where K is a constant and $d = s - r$. That this model does not apply in the game of Go is amply demonstrated in figure 1. For example, the data shows clearly that it is much harder for a 2 dan to beat a 4 dan ($P = 22\%$) than it is for a 4 kyu to beat a 2 kyu ($P = 35\%$), so the probability of win in Go cannot be governed by a simple grade difference as Elo implies. Thus Λ needs to be a non-linear function of r, s and cannot depend on the difference $d = s - r$ alone.

There is one further condition governing the probability $p(r,s)$ (and hence the form of Λ) that arises from the nature of the game of Go: it is virtually impossible for a player to win in an even game against a player 9 stones or more stronger. Most pairing programs would regard such a pairing as unacceptable, so it should be difficult to find tournament examples of such games. Teaching and club game experience however, justify the statement as holding no matter what the grade of the weaker player is, and so we shall require that $p(r,s)$ satisfies:

$$\begin{aligned} p(r, r+d) &\approx 0 \\ p(r, r-d) &\approx 1 \\ d &\geq 9 \end{aligned} \tag{3}$$

3 DETAILS OF THE MODEL FOR Λ

We can express $\Lambda(r,s)$ in terms of the probability values $p(r,s)$ by inverting equation (2) to get:

$$\Lambda(r,s) = erf^{-1}(1 - 2p(r,s)) \tag{4}$$

The graphs displaying the form of $\Lambda(r,d)$ derived from the E.G.D data by applying (4) with $d = r - s$ are shown below, and as before, the family of curves is labelled by the grade difference d. These graphs suggest that each curve has an exponential shape, so my first numerical experiments to obtain a fit concentrated on forms like $\Lambda(r,d) = A + Be^{Kr}$, where A, B K are functions of d alone.

These models gave a good starting point, but it soon became clear that for each d, the curvature in the exponential function does not grow fast enough to cope with the accelerating rise in Λ for all grades above shodan. The natural extension is to consider further exponential terms forming a series:

$$\Lambda(r,d) = B_0 + B_1 e^{Kr} + B_2 e^{2Kr} + B_3 e^{3Kr} + \cdots ,$$

where all the coefficients depend on d only.

This improved the fit to each individual curve further, and the surprise was that the value of K did not vary much with d at all, so it is appropriate to regard K as a fixed constant independent of both r and d. No more than the third order in the exponential series was needed.

This suggests that for the case $r < s$ the model for Λ should take the form:

$$\Lambda(r,s) = \sum_{n=0}^{3} H_n(s-r)e^{nKr}$$

Since Λ is antisymmetric, this implies that for $r > s$

$$\Lambda(r,s) = -\Lambda(s,r) = \sum_{n=0}^{3} -H_n(r-s)e^{nKs}$$

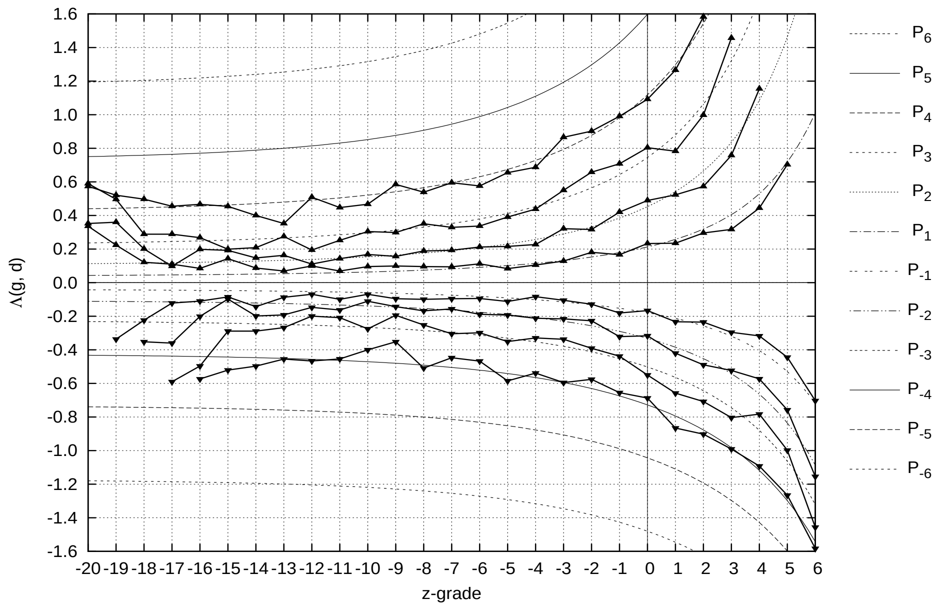

Figure 2: Graph of Λ derived from win ratios.

We can construct the antisymmetric function $h_n(x)$ from $H_n(x)$ by:

$$\begin{aligned} h_n(x) &= H_n(x), & x > 0 \\ &= -H_n(-x), & x < 0 \\ &= 0, & x = 0 \end{aligned}$$

Observing that $r = \min(r,s)$ when $r < s$ and that $s = \min(r,s)$ when $r > s$, the exponential terms simplify, and we arrive at the following single expression for the model:

$$\Lambda(r,s) = \sum_{n=0}^{3} h_n(s-r)e^{n\,K\,\min(r,s)} \tag{5}$$

A suitable form for the functions h_n is

$$h_n(x) = u_n x + v_n x^3, \quad n = 0 \cdots 3, \tag{6}$$

where u_n and v_n are *positive* constants specifying monotonic increasing cubics. This guarantees antisymmetry and helps to ensure that $|\Lambda|$ is large when $|d| > 9$ as required by the properties expressed in equations (3).

4 THE FITTING ALGORITHM

We use a weighted, non-linear least squares method for finding the parameters in the model for Λ presented in the previous section. It is convenient to implement the fitting procedure in the (r,d) co-ordinate system. The function to be minimised is:

$$F(\mathbf{q}) = \sum_{r_i=-20}^{6} \sum_{d_j=-4}^{4} w_{r_i} [\Lambda(r_i, d_j) - \Lambda_{r_i d_j}])^2$$

The model function $\Lambda(r, d) \equiv \Lambda(r, r + d)$ is fully specified via equations (5) and (6). The parameter $\mathbf{q}$ bundles all the free variables into one vector:

$$\mathbf{q} = (u_0, v_0, u_1, v_1, \cdots, K)$$

The weights are assigned on the basis of player population and grade. We form the cumulative sum of games played starting from the 20 kyu end:

$$W(r) = \sum_{n=-19}^{r} \sum_{d=1}^{4} G(n, d)$$

$G(r, d)$ is the count of number of games played between players of grades r and $r+d$ as published in [4]. The individual weights w_r are obtained by normalising the distribution $W(r)$ so that its maximum value is 1. Note that we ignore the data for 20 kyu players in the fitting procedure because this group contains other players whose grades are below 20 kyu.

The minimisation of F is carried out using the conjugate gradient algorithm [7]. Starting from the initial point in the table belowe, the algorithm converges within 20 cycles to an accuracy of 10^{-10} in the magnitude of ∇F at the minimum.

	$\mathbf{u_0}$	$\mathbf{v_0}$	$\mathbf{u_1}$	$\mathbf{u_3}$	**K**
initial	0.04	0.004	0.1	0.02	0.2
solution	0.0351224	0.00445376	0.156777	0.0164481	0.18818

Table 1: Coefficients for period 2001 to 2010.

A good fit was obtained ignoring the free variables v_1,u_2, v_2, and v_3. Firstly v_1 and v_3 were usually small, and secondly u_2, v_2 produced negative values for some years, leading to unjustified lumps in the model.

Considering only players in the range 12 kyu to 7 dan, the rms error between the measured win ratios and the computed probability of win is less than 0.02, for each value of $d = 1 \cdots 4$. A shodan has a probability of win against a 9 dan of 3.6×10^{-15} and against a 9 kyu of $1 - 1.5 \times 10^{-5}$.

The coefficients shown in table 1 were obtained for the winning statistics data covering the period January 2001 to December 2010. These coefficients are used to generate the families of P_{win} and Λ curves displayed in figures 1 and 2.

5 STABILITY OF THE MODEL OVER TIME

The measured win ratios change over time, but it is expected that these will settle down to a constant value over a long enough time period. The E.G.D system makes available winning statistics for even games over *any* specified time period. Experiments showed that the winning probabilities become reasonably stable over a 5 year period. It could be misleading simply to gather statistics from the inception of the system, since this may hide the existence of a long term variation. In order to expose any such feature we examine the trace of win ratios gathered for a moving 5 year window since 1996.

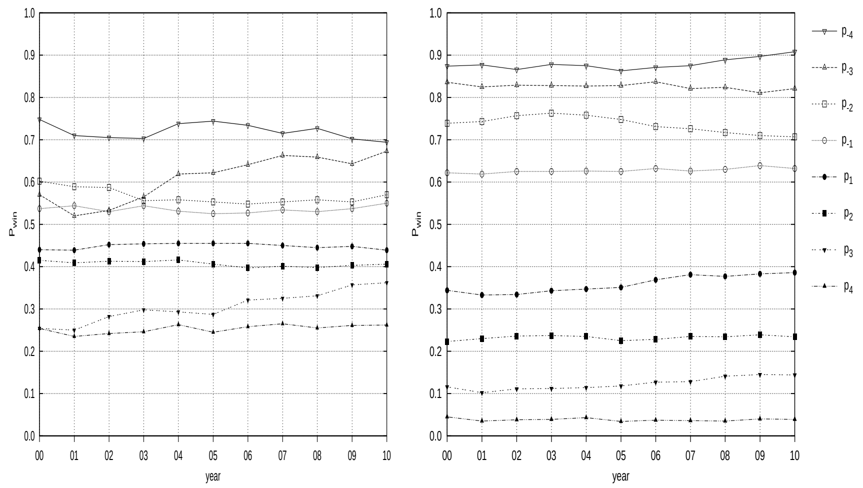

Figure 3: Measured P_{win} for 10 kyu (left) and 2 dan (right).

In these graphs, each point is a win ratio gathered over a 5 year time window ending on New Year's Eve of the year shown and starting 5 years earlier on New Year's day. Traces are shown for winning against players from 4 stones weaker (trace p_{-4}) to players 4 stones stronger (trace p_4). The trace showing the greatest variation is p_{-3} for the 10 kyu player, i.e. winning against 13 kyu players. The 2 dan players show improved stability, and for all players the system is reasonably stable from 2006 on.

We next examine the stability of the *coefficients* for the Λ model in the following trace produced for the same time periods as in figure 3.

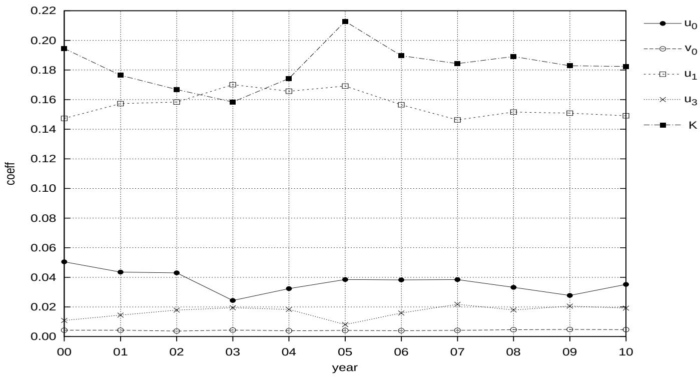

Figure 4: Trace of coefficients for fit to Λ.

The instability in the coefficients is very much in evidence prior to 2006, especially in the all important K governing the growth of the exponential terms. However, the main interest is the behaviour of P_{win} computed from the coefficients, rather than the variation in the coefficients themselves.

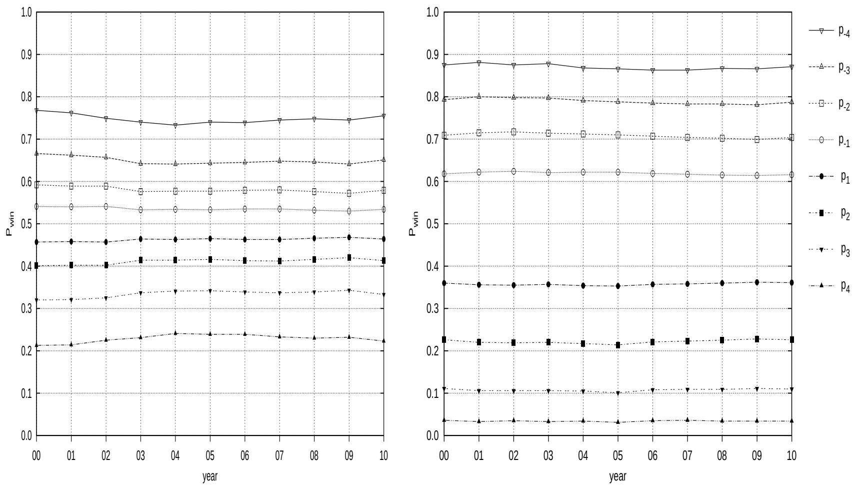

Figure 5: Model P_{win} 10 kyu (left) and 2 dan (right).

The traces above show the behaviour of our 10 kyu and 2 dan players resulting from the calculated P_{win} values. Clearly this is much more stable than the coefficient trace - thanks to the important damping property of erf.

6 GRAPHICAL PRESENTATION

The antisymmetry of Λ means that it has reflection symmetry about the line $r = s$. This suggests that a useful representation of the model comes about by viewing P_{win} referred to axes along and perpendicular to this line. We thus transform to canonical (g, d) co-ordinates (Appendix A.2) defined by:

$$g = \tfrac{1}{2}(r + s), \quad d = s - r \tag{7}$$

$$r = g - \tfrac{1}{2}d, \quad s = g + \tfrac{1}{2}d \tag{8}$$

In the following contour plot of $p(g, d) = \frac{1}{2}[1 - erf(\lambda(g, d)]$, probability is expressed as a percentage.

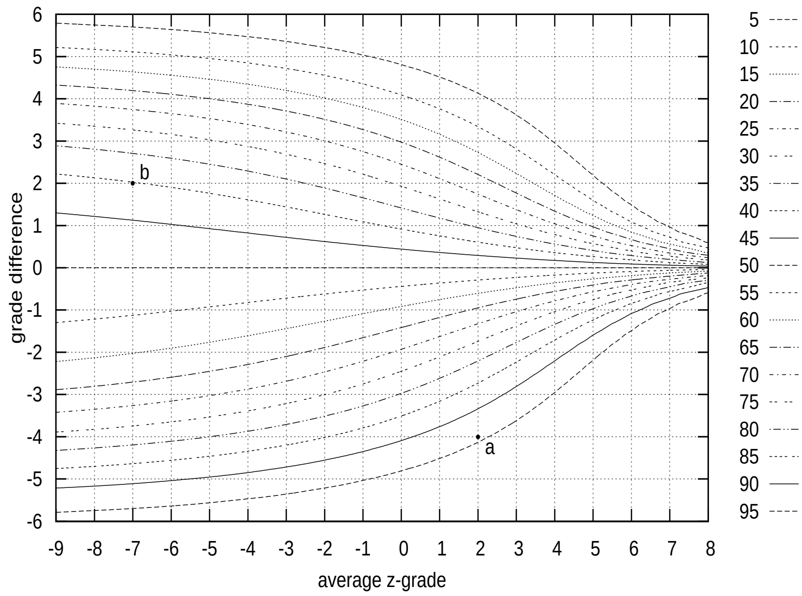

Figure 6: Contour plot of P_{win}.

For example, the 95% contour almost passes through the point **a** with average grade of 2 (in zero shodan units) and grade difference -4. So the stronger player has grade $2-(-2) = 5$ dan, and beats the weaker player with grade $2+(-2) = 1$ dan with 95% probability.

On the other hand, if an 8 kyu plays a 6 kyu, the average grade is -7, the grade difference is +2, and the nearest contour line passing through the corresponding point **b** shows that the probability of a win is close to 40%.

7 SUMMARY AND CONCLUSION

A model for the probability of win in games between players with arbitrarily different grades has been developed, and has the following properties:

- The model agrees with measured win ratios for players from 12 kyu to 7 dan. The root mean square deviation between calculated and measured winning probabilities is better than 0.02 for players above 12 kyu.
- The model has the property that the extrapolated probability of win for players with grades more than 9 stones apart yields a value near 0 for the weaker player and near 1 for the stronger player.
- With a 5 year time window for collecting win ratio statistics, we have shown that the model is stable over the period 2001 to 2010 with no evidence of any long term trend.
- The model has just 5 parameters and is easy to implement. It is especially useful for the purposes of simulation.

8 FURTHER WORK

Players below 12 kyu

It has not been established why, as seen in figure 1, players below 12 kyu contradict the observation that the stronger you are the harder it is to beat someone 1 stone stronger than you in an even game. The contribution to total games by players below 12 kyu is about 15% and so it is legitimate to give this data low weight when fitting for the stronger players. Although the behaviour is anomalous, it is not totally chaotic, and it is seen over a wide range of time periods.

Model detail

The general form of the model in equation (5) is a power series. We have obtained a good fit with the first 4 terms in the series. However when the second order term $n = 2$ was included, the solution invariably produced non-negligible negative coefficients. These cause unjustified lumps in P_{win}, and more data would be needed to justify increasing the model complexity. This issue would best be resolved by attempting to extract winning probabilities from E.G.D where the grade difference is 5 stones or more.

Tie-break analysis

Many tie-breaks have been used in European tournaments, and Herman Hiddema [8] has proposed a sound testing method for analysing their effectiveness. The simulation of tournaments could be carried out using the above model for the probability of win as it provides a realistic method for generating game results.

Pairing Algorithms

A player's measured rating is not completely independent of the algorithm used for the pairing. The rating change achieved at a tournament depends on the players strength on the day *and* on the opponents chosen by the pairing system.

Start with a population of players each with a true strength (equal to grade) fixed for all time. A simulation of the tournament is carried out computing results from the P_{win} model using the true strengths. At the end of the tournament the ratings of all players are updated according to the E.G.D rules.

After many such tournament simulations, each player will have acquired a measured rating, which may differ from the true strength. A suitable metric for measuring the *quality* of a pairing algorithm is the rms error between the true and measured ratings.

This gives us a method for *comparing* the quality of pairing algorithms. In particular, modern techniques [9], [10] rely on assigning weights to potential pairings for use in maximum weighted matching algorithms. The assignment of weights is dependent on a number of parameters, whose values have so far been determined empirically. We can exploit the P_{win} model via the definition of pairing quality to optimise the parameters determining the weights.

A ANTISYMMETRY IN THE MODEL

A.1 The function $\Lambda(r, s)$ is antisymmetric

It follows from the properties expressed in equations (1) that $p(s, r) = P^0(r, s)$, and consequently:

$$\begin{aligned} p(s,r) &= 1 - p(r,s) \\ &= 1 - \tfrac{1}{2}[1 - erf(\Lambda(r,s))] \\ &= \tfrac{1}{2}[1 + erf(\Lambda(r,s))] \end{aligned}$$

Interchanging r and s and invoking the antisymmetry of erf we see that

$$p(r,s) = \tfrac{1}{2}[1 - erf(-\Lambda(s,r))]$$

It then follows from equation (2) that $\Lambda(r, s) = -\Lambda(s, r)$.

A.2 Canonical transform of Λ

On transforming to (g, d) co-ordinates, the definition (5) produces:

$$\begin{aligned} \Lambda(r,s) &= \Lambda(g - \tfrac{1}{2}d, g + \tfrac{1}{2}d) \\ &= \sum_{n=0}^{3} h_n(d) e^{n\,K\,\min(g-\frac{1}{2}d, g+\frac{1}{2}d)} \end{aligned}$$

Observing that $d = |d|$ when $d > 0$, and $d = -|d|$ when $d < 0$, we see that the argument in the exponential term simplifies, giving the following expression for the transformed version of Λ:

$$\tilde{\Lambda}(g,d) = \sum_{n=0}^{3} h_n(d) e^{nK(g-\frac{1}{2}|d|)} \tag{9}$$

A.3 Approximation for erf

The following approximation [11] to erf has an accuracy better than 6×10^{-4}

$$\begin{aligned} erf(x) &\approx tanh(ax + bx^3) \\ a &= 1.129324 \\ b &= 0.100303 \end{aligned} \tag{10}$$

$$\begin{aligned} erf^{-1}(x) &\approx d\ sinh(\tfrac{1}{3} sinh^{-1}(c\ tanh^{-1}(x))) \\ d &= \sqrt{\tfrac{4a}{3b}} = 3.875 \\ c &= \tfrac{3}{ad} = 0.6856 \end{aligned} \tag{11}$$

References

[1] EGD, http://www.europeangodatabase.eu/EGD/

[2] EGD[1] - Country/Club statistics.

[3] EGD - EGF Rating System.

[4] EGD - Winning statistics.

[5] Milton Abramowitz and Irene A. Stegun, Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables, Dover New York,1965.

[6] Arpad Elo, The Rating of Chessplayers, Past and Present (1978), Arco.

[7] Conjugate gradient method, http://en.wikipedia.org/wiki/Conjugate_gradient_method.

[8] Herman Hiddema, Testing Tiebreakers, http://senseis.xmp.net/?TieBreaker%2FTestingTiebreakers.

[9] Christoph Gerlach, MacMahon pairing program, http://www.cgerlach.de/go/macmahon.html

[10] Luc Vannier, OpenGotha pairing program, http://vannier.info/jeux/gotournaments/opengotha.htm

[11] T. J. Fairclough, The Case of the Counterfeit Error Function, Mathematics Today, Vol 36 No 5, October 2000

[1] References to EGD quote the names of tabs on the main page, not the whole address