COMMENTS ON NORMALIZED ELO

1. Introduction

Normalized Elo is introduced here [4]. The primary motivation for normalized Elo is that it is a measure for the amount of games it takes to prove that one engine is stronger than another, with a given level of significance. In other words it is an *objective measure* of strength difference.

In this document we make some cosmetic changes to the terminology introduced in loc. cit. In particular what was called "normalized Elo" will now be called "normalized t-value" and we redefine normalized Elo as the normalized t-value multiplied by an appropriate normalization constant. This is done to make the comparison with ordinary logistic Elo more intuitive.

2. Background

The *normalized t-value* for the strength difference of two engines is defined as

$$\boxed{t_n := \frac{s - 1/2}{\sigma_{\mathrm{pg}}}}$$

where s is the expected score and σ_{pg} is the standard deviation of the expected score *per game*. In the trinomial case σ_{pg} is the standard deviation of the outcome distribution of a game, scored as $0, 1/2, 1$. In the pentanomial case, σ_{pg} is the standard deviation of the outcome distribution multiplied by $\sqrt{2}$, where we score the outcome of a game pair as $0, 1/4, 2/4, 3/4, 1$.

The justification for this convention is that, whatever testing system we use, the normalized t-value $\hat{t}_n$ of a test is defined to be the usual t-value divided by the square root of the number of games. Then t_n is the asymptotic expectation value of $\hat{t}_n$. More precisely, asymptotically we have

$$\boxed{\hat{t}_n \sim N(t_n, \frac{1}{N})}$$

where N is the number of games.

In the trinomial case or in the pentanomial case with a perfectly balanced book we have

$$\sigma_{\mathrm{pg}} = \frac{1}{2}\sqrt{1-d} \tag{2.1}$$

where d is the draw ratio.

3. Normalization

Below we put

$$S(x) = \frac{1}{1 + 10^{-x/400}}$$

This is the function which converts ordinary ("logistic") Elo into an expected score. It is convenient to write

$$S(x) = L(\beta x)$$

where $\beta = \log(10)/400$ and L is the usual logistic function

$$L(x) = \frac{1}{1+e^{-x}}$$

L satisfies the functional equation

$$L'(x) = L(x)(1 - L(x))$$

Let

$$\boxed{C_{e/t} := \frac{2}{\beta} = \frac{800}{\log(10)} \cong 347.43}$$

We claim that for small Elo differences we have

$$t_n \cong \frac{1}{C_{e/t}} \frac{e_l}{2\sigma_{\text{pg}}} \tag{3.1}$$

where e_l is the logistic Elo difference between two engines. To see this note

$$e_l \cong (s - S(0))/S'(0) = (s - 1/2)/(\beta L'(0)) = (s - 1/2)/(1/2(1 - 1/2)\beta) = 4(s - 1/2)/\beta$$

Hence

$$\frac{1}{C_{e/t}} \frac{e_l}{2\sigma_{\text{pg}}} \cong \frac{\beta}{2} \frac{4}{\beta} \frac{s - 1/2}{2\sigma_{\text{pg}}} = \frac{s - 1/2}{\sigma_{\text{pg}}} = t_n$$

We define the normalized Elo difference between two engines as

$$\boxed{e_n := C_{e/t} t_n} \tag{3.2}$$

In case of a perfectly balanced book it follows from (2.1) and (3.1) that

$$\boxed{e_n \cong \frac{e_l}{\sqrt{1-d}}} \tag{3.3}$$

This simple formula is the motivation for the normalization introduced in (3.2). We see in particular that for $d = 0$ normalized Elo and logistic Elo coincide. For other draw ratios we have the following conversion table.

Draw ratio	0.00	0.30	0.50	0.60	0.70	0.80	0.90
Normalized Elo	5.00	5.00	5.00	5.00	5.00	5.00	5.00
Logistic Elo	5.00	4.18	3.54	3.16	2.74	2.24	1.58

Let us now discuss the duration of an SPRT test for H0:$e_n = e_{n,0}$ versus $H1 : e_n = e_{n,1}$. Under the assumption that the the Type I/II error probabilities are given by $\alpha = \beta = 0.05$ we get that the worst case expectation duration (which corresponds to the actual Elo being half way between H0 and H1) of the test is given by

$$\boxed{T = \frac{D}{(e_{n,1} - e_{n,0})^2}} \tag{3.4}$$

where

$$\boxed{D := C_{e/t}^2 \log(19)^2 \cong 1046535}$$

This leads to the following table

Normalized Elo difference	1	2	3	4	5	6
Expected duration	1046535	261634	116282	65408	41861	29070

Note that D is close to 1000000 which is sufficiently accurate for back of the envelope calculations.

Let us derive the formula (3.4). We may equivalently consider an SPRT of $t_n = t_{n,0}$ versus $t_n = t_{n,1}$. Let us suppose that σ_{pg} is known (see §4 below for a discussion) so it is sufficient to consider an SPRT test for $s = s_0$ versus $s = s_1$. According to [6] the expected duration of such a test, when the actual score is s is equal to

$$T = \frac{T(h_s)}{w^2}$$

where

$$\begin{aligned} w &= \frac{s_1 - s_0}{\sigma_{\text{pg}}} = t_{n,1} - t_{0,n} \\ h_s &= \frac{2s - (s_0 + s_1)}{s_1 - s_0} = \frac{2t_n - (t_{n,0} + t_{n,1})}{t_{n,1} - t_{n,0}} \\ T(h) &= \frac{2b}{h}\frac{1 - e^{-hb}}{1 + e^{-hb}} \\ b &= \log\left(\frac{1-\alpha}{\alpha}\right) \end{aligned}$$

when the Type I/II error probabilities are both equal to α.

The worst case is given when $t_n = (t_{n,0} + t_{n,1})/2$. In that case $h_{t_n} := h_s = 0$. Applying l'Hôpital's rule we find

$$T = \frac{b^2}{(t_{n,1} - t_{n,0})^2} = \frac{C_{e/t}^2 b^2}{(e_{n,1} - e_{n,0})^2}$$

If $\alpha = 0.05$ then $b = \log(19)$ and we obtain (3.4).

For completeness we note that in case $t_n = t_{n,0}$ or $t_n = t_{n,1}$ the expected duration is given by a similar formula as (3.4) where the numerator is replaced by

$$\boxed{D' := \frac{9C_{e/t}^2 \log(19)}{5} \cong 639770}$$

It is probably more meaningful to consider the ratio D'/D.

$$\boxed{\frac{D'}{D} \cong 0.61132}$$

4. LLR computation

What we call an SPRT is strictly speaking a GSPRT [7] which is based on monitoring the Generalized Log Likelihood Ratio (which we denote by LLR below) of H1 versus H0. See [2] for an introduction.

In [5] a relatively elegant method was given to compute the LLR for an SPRT for the mean of a multinomial distribution. Unfortunately I did not succeed in finding a similarly elegant approach for the normalized t-value. Adhoc numerical approaches are not very attractive mathematically and in addition provide little insight.

In [5] an approximation for the LLR was derived and in [3] this was compared to the exact one. It seems a good strategy to use this approximation if in addition we estimate σ_{pg} (necessary to convert the mean into a t-value) from the test itself.

We now describe the procedure more precisely. Assume given real numbers

$$a_1 < a_2 < \cdots < a_l$$

and a discrete probability distribution

$$P : \{a_1, \ldots, a_l\} \to \mathbb{R} : a_i \mapsto p_i \,.$$

with mean s and standard deviation σ. Assume a sample taken from $\{a_1, \ldots, a_N\}$ according to P has sample distribution $(\hat{p}_i)_{i=1,\ldots,N}$. Then by [5] the LLR for an SPRT of H0:$s = s_0$ versus H1:$s = s_1$ may be approximated by

$$\mathrm{LLR} \cong \frac{n}{2} \log\left(\frac{\sum_i \hat{p}_i (s_0 - a_i)^2}{\sum_i \hat{p}_i (s_1 - a_i)^2}\right) = \frac{n}{2} \log\left(\frac{(\hat{s} - s_0)^2 + \hat{\sigma}^2}{(\hat{s} - s_1)^2 + \hat{\sigma}^2}\right)$$

where n is the sample size. Let s_{ref} be some reference value and put

$$\hat{t} = \frac{\hat{s} - s_{\mathrm{ref}}}{\hat{\sigma}}, \qquad t_0 = \frac{s_0 - s_{\mathrm{ref}}}{\sigma}, \qquad t_1 = \frac{s_0 - s_{\mathrm{ref}}}{\sigma} \tag{4.1}$$

In the context of the pentanomial model described above, $s_{\mathrm{ref}} = 1/2$, $n = N/2$ and the t-values in (4.1) are the normalized t-values multiplied by $\sqrt{2}$. Assuming that $\hat{\sigma}$ is a good approximation for σ we find

$$\boxed{\mathrm{LLR} \cong \frac{n}{2} \log\left(\frac{1 + (\hat{t} - t_0)^2}{1 + (\hat{t} - t_1)^2}\right)} \tag{4.2}$$

That (4.2) performs entirely satisfactorily is confirmed by simulation. See [1].

Remark 4.1. In practice $\hat{t}$, t_0 and t_1 will be small compared to 1 so that (4.2) can be further approximated by

$$\mathrm{LLR} \cong \frac{n}{2}((\hat{t} - t_0)^2 - (\hat{t} - t_1)^2) = \frac{n}{2}(t_1 - t_0)(2\hat{t} - t_0 - t_1) \tag{4.3}$$

which in case of the trinomial or pentanomial model becomes

$$\mathrm{LLR} \cong \frac{N}{2}(t_{n,1} - t_{n,0})(2\hat{t}_n - t_{n,0} - t_{n,1})$$

This formula works just as well but it has a small technical inconvenience. At the beginning of a test (say after a few game pairs with identical outcomes) $\hat{\sigma}$ will still be very small[1] and hence $\hat{t}$ may be spuriously large. Then the same will be true for (4.3). This is not a serious problem if one waits with reporting the LLR until there are some game pairs with different outcomes. This is enough to bring σ into safe territory.

However (4.2) does not suffer from this issue so when using this approximation no special arrangements need to be made for the beginning of the test.

[1]In fact $\hat{\sigma} = 0$. However for simplicity we replace zero outcome frequencies with a small $\epsilon > 0$ to avoid division by zero.

References

1. Michel Van den Bergh, *A multi-threaded pentanomial simulator in C*, `https://github.com/vdbergh/simul`.
2. ———, *A practical introduction to the GSPRT*, `http://hardy.uhasselt.be/Fishtest/GSPRT_approximation.pdf`.
3. ———, *Comparing the approximations for the Generalized Log Likelihood Ratio of a multinomial distribution*, `http://hardy.uhasselt.be/Fishtest/comparing_approximations.pdf`.
4. ———, *Normalized Elo*, `http://hardy.uhasselt.be/Fishtest/normalized_elo.pdf`.
5. ———, *The Generalized Likelihood Ratio for the expectation value of a distribution*, `http://hardy.uhasselt.be/Fishtest/support_MLE_multinomial.pdf`.
6. ———, *The SPRT for Brownian motion*, `http://hardy.uhasselt.be/Fishtest/sprta.pdf`.
7. Xiaoou Li, Jingchen Liu, and Zhiliang Ying, *Generalized Sequential Probability Ratio Test for Separate Families of Hypotheses*, `http://stat.columbia.edu/~jcliu/paper/GSPRT_SQA3.pdf`.