NORMALIZED ELO

MICHEL VAN DEN BERGH

1. Introduction

It is often discussed how one should compare the suitability for engine testing of opening books with different characteristics (balanced versus unbalanced, drawishness). Here we answer this question. A lot of inspiration came from Kai Laskos' empirical results posted on talkchess as well as from his comments. Credit goes also to him for first noting that the trinomial model is insufficient for working with unbalanced openings.

2. Normalized Elo

The primary purpose of an engine test is to show that one engine is stronger than another. The amount of evidence gathered from a test can be expressed by its "t-value".

$$t = \frac{\hat{s} - 1/2}{\hat{\sigma}(\hat{s})} \tag{2.1}$$

where $\hat{s} = \hat{w} + (1/2)\hat{d}$ is the score of a match (with $\hat{w}, \hat{d}, \hat{l}$ being the empirical WDL-ratios) and $\hat{\sigma}(\hat{s})$ is empirical standard deviation of this score. Note that the null-hypothesis is $s = 1/2$ and hence $t = 0$. The relation between the t-value and the more commonly used Likelihood Of Superiority is

$$\mathrm{LOS} = \Phi(t)$$

where Φ is the cumulative distribution function of a normal distribution with unit variance and zero mean. The corresponding p-value is

$$p = 1 - \mathrm{LOS}$$

The disadvantage of (2.1) is that $\hat{\sigma}(\hat{s})$ depends on the number of games N in the test via

$$\hat{\sigma}(\hat{s}) = \frac{\hat{\sigma}_0}{\sqrt{N}}$$

where σ_0 refers to the standard deviation of a single game. So

$$t = \sqrt{N}\frac{\hat{s} - 1/2}{\hat{\sigma}_0}$$

So to compare opening books we should really use the normalized quantity.

$$\boxed{t_0 = \frac{\hat{s} - 1/2}{\hat{\sigma}_0}} \tag{2.2}$$

In the case of balanced openings we find

$$\boxed{t_0 = \frac{\hat{s} - 1/2}{\sqrt{\hat{w} + \frac{1}{4}\hat{d} - \hat{s}^2}}} \tag{2.3}$$

Remark 2.1. In the case of unbalanced openings with paired games with reversed colors the formula $\hat{\sigma}_0 = \sqrt{\hat{w} + \frac{1}{4}\hat{d} - \hat{s}^2}$ over estimates $\hat{\sigma}_0$ and we should use the pentanomial model instead [2]. See §4.

Remark 2.2. Note

$$\boxed{\sigma(t_0) = \frac{1}{\sqrt{N}}} \tag{2.4}$$

so a 95%-confidence interval for t_0 is given by

$$\boxed{\left[t_0 - \frac{1.96}{\sqrt{N}}, t_0 + \frac{1.96}{\sqrt{N}}\right]} \tag{2.5}$$

It makes sense to call t_0 "normalized elo" since the result of an engine match is usually expressed as elo ± error where a 95%-confidence interval is assumed and then in good approximation

$$\boxed{t_0 \cong \frac{1.96\,\mathrm{elo}}{\mathrm{error}\sqrt{N}}} \tag{2.6}$$

Other reasonable names for t_0 are "sensitivity" (when measured under a standard set of conditions) or "signal-to-noise ratio" (see §3).

Remark 2.3. It is easy to see that t_0 is independent of linear reparametrizations of the outcome scoring. Hence we may score wins, draws, losses as $-1, 0, 1$. Then we find

$$t_0 = \frac{\hat{w} - \hat{l}}{\sqrt{\hat{w} + \hat{l} - (\hat{w} - \hat{l})^2}}$$

which yields the approximation

$$\boxed{t_0 = \frac{t_0'}{\sqrt{1 - t_0'^2}}} \tag{2.7}$$

with

$$\boxed{t_0' = \frac{\hat{w} - \hat{l}}{\sqrt{\hat{w} + \hat{l}}}} \tag{2.8}$$

For small elo differences we obtain

$$\boxed{t_0 = t_0' + \frac{1}{2}t_0'^3 + \cdots} \tag{2.9}$$

so that we may use t_0 and t_0' interchangeably.

Remark 2.4. We have

$$t_0' = \frac{\hat{w} - \hat{l}}{\sqrt{1 - \hat{d}}} = 2\frac{\hat{s} - 1/2}{\sqrt{1 - \hat{d}}}$$

For small elo differences we have

$$\hat{s} - 1/2 \cong \frac{\log(10)}{1600}\text{elo}$$

so that we also find

$$t_0' \cong \frac{\log(10)}{800} \frac{\text{elo}}{\sqrt{1 - \hat{d}}}$$

which gives the approximation

$$\boxed{t_0 \cong t_0' \cong 0.002878\frac{\text{elo}}{\sqrt{1 - \hat{d}}}} \tag{2.10}$$

Remark 2.5. A typical setup to measure sensitivity is a self match with time odds using a well known engine. This reduces the number of parameters involved in setting up the testing environment. Time doubling seems reasonable although it is unknown if results for large elo differences truly extrapolate to small elo differences.

3. Signal-to-noise ratio

Here we present a different point of view on “normalized elo”. In engineering a signal is decomposed[1] as

$$X = X_s + X_n$$

where X_s is the true signal and X_n is the noise added to it. The power ratio between the two components of X is the signal-to-noise ratio (SNR)

$$\text{SNR} = \frac{\int X_s(t)^2\,dt}{\int X_n(t)^2\,dt}$$

As noise will typically have zero mean we may also write

$$\text{SNR} = \left(\frac{\text{RMS}(X_s)}{\sigma(X_n)}\right)^2$$

where RMS stands for “root mean square”:

$$\text{RMS}(X_s) = \sqrt{\frac{1}{T}\int_0^T X_s(t)^2\,dt}$$

We can view an engine match as a noisy DC-signal where X_s is the expected score s minus $1/2$ and X_n is the difference between the game outcome $\in \{0, 1/2, 1\}$ and s. We then find

$$\text{SNR} \cong t_0^2$$

[1] A signal will typically be sent over the wire by modulating the properties of a carrier signal, which has its own power content. Here we assume that the signal has already been demodulated.

4. The pentanomial model

Let $O := \{0, 1/2, 1, 3/2, 2\}$ be the set of possible outcomes of game pairs and let $(\hat{p}_i)_{i\in O}$ be the corresponding sample distribution. Let

$$\hat{s}_2 := \sum_{i\in O} \hat{p}_i i = 2\hat{s}$$

be the average outcome of a game pair. Then we get

$$\hat{\sigma}(\hat{s}_2) = \frac{1}{\sqrt{N/2}}\sqrt{\sum_{i\in O} \hat{p}_i i^2 - \hat{s}_2^2}$$

So for the t-value we obtain

$$t = \frac{\hat{s}_2 - 1}{\frac{1}{\sqrt{N/2}}\sqrt{\sum_{i\in O} \hat{p}_i i^2 - \hat{s}_2^2}}$$

Normalizing (dividing by $\sqrt{N}$) we get

$$\boxed{t_0 = \sqrt{2}\frac{\hat{s} - 1/2}{\sqrt{\sum_{i\in O} \hat{p}_i i^2 - 4\hat{s}^2}}} \tag{4.1}$$

5. Normalized elo and engine testing

5.1. **Introduction.** Below we will refer to a *context* as the conditions under which an individual game (or perhaps game pair) takes place in an engine match. Traditionally a context consists of a book and a time control. Other ingredients of a context could be certain engine settings (like contempt or hash size) and adjudication rules. As a general principle we will decorate quantities with a subscript to indicate the context under which they are measured.

For a context D and engines X and Y we denote in this section the normalized elo difference by $\mathrm{NE}_D(X, Y)$ (rather than the adhoc notation t_0 which we used above).

Remark 5.1.1. By (2.5) $\mathrm{NE}_D(X, Y)$ is a well defined quantity. It can be measured with precision $O(1/\sqrt{N})$ using a fixed length match with N games using the context D.

Now let us define for contexts C, D:

$$\mathrm{NE}_{D/C}(X, Y) := \frac{\mathrm{NE}_D(X, Y)}{\mathrm{NE}_C(X, Y)} \tag{5.1}$$

It follows from Remark 5.1.1 that $\mathrm{NE}_{D/C}(X, Y)$ is also a well defined quantity that can be measured to any desired degree of precision.

Now the idea is that $\mathrm{NE}_{D/C}(X, Y)$ should only be weakly dependent on X, Y so that in fact we may write

$$\mathrm{NE}_{D/C} = \mathrm{NE}_{D/C}(X, Y) \tag{5.2}$$

We will call this the *weak dependency hypothesis*. We may call $\mathrm{NE}_{D/C}$ the *relative sensitivity* of context D compared to context C.

Remark 5.1.2. Of course in the generality we have stated it, the weak dependency hypothesis will be trivially false. It is more likely to be somewhat correct when X is a patched version of Y and Y is a patched version of some base engine Z. But even then the hypothesis should probably be interpreted statistically since for example in the case of an increase in time control there could be "well scaling patches" and "badly scaling patches". Below we will ignore such subtleties to simplify the exposition.

Remark 5.1.3. Under the trinonial model we have by (2.10)

$$\frac{\mathrm{NE}_D(X,Y)}{\mathrm{NE}_C(X,Y)} \cong \sqrt{\frac{1-d_C(X,Y)}{1-d_D(X,Y)}}\frac{\mathrm{elo}_D(X,Y)}{\mathrm{elo}_C(X,Y)}$$

where $\mathrm{elo}_D(-)$ is the standard logistic elo difference for D and $d_D(-)$ is the draw ratio. It is well known that for closely related engines X, Y the draw ratio is already quite independent of X and Y. So the weak dependency hypothesis is more or less the same as the hypothesis that there is is a scaling factor for logistic elo measured under the contexts C and D which is weakly dependent on X, Y.

We will prove (somewhat heuristically)

Theorem 5.1.4. *Assume the weak dependency hypothesis holds. Let C, D and let $S = \mathrm{SPRT}_C(0,e)$, $T = \mathrm{SPRT}_D(0,f)$ be* SPRT*'s using the indicated contexts C, D which have the same power to separate the engines X, Y if their elo difference is e under context C. More precisely we assume.*

(5.3)
$$\begin{aligned} &P_S(\mathrm{H1}\ \textit{accepted} \mid \mathrm{elo}_C(X,Y)=0) = P_T(\mathrm{H1}\ \textit{accepted} \mid \mathrm{elo}_C(X,Y)=0) = \alpha \\ &P_S(\mathrm{H0}\ \textit{accepted} \mid \mathrm{elo}_C(X,Y)=e) = P_T(\mathrm{H0}\ \textit{accepted} \mid \mathrm{elo}_C(X,Y)=e) = \beta \end{aligned}$$

For $R \in \{S,T\}$ let $N_R(X,Y)$ be the expected number of games for a test R to finish. Then

$$\boxed{\frac{N_S(X,Y)}{N_T(X,Y)} = \mathrm{NE}^2_{D/C}} \tag{5.4}$$

Remark 5.1.5. We have seen in §3 that $\mathrm{NE}_C(X,Y)$ is the square root of the signal-to-noise ratio for an engine match. So (5.4) shows that the relative effort required to separate two engines using contexts C, D is proportional to the relative SNR. I consider this to be a very intuitively plausible statement.

Remark 5.1.6. The analogue of Theorem 5.1.4 for fixed length tests is trivial to prove.

5.2. **Proof of Theorem 5.1.4.** We will work somewhat more generaly, using the results in [1]. Assume that ϕ is some quantity we are measuring. An SPRT (or GSPRT) of $\mathrm{H0} : \phi = \phi_0$ against $\mathrm{H1} : \phi = \phi_1$ is (asymptotically) a sequential test based on monitoring

$$\mathrm{LLR}_M = \frac{1}{2}\frac{(\phi_1-\phi_0)(2\hat{\phi}_M-\phi_0-\phi_1)}{V(\hat{\phi}_M)}$$

with continuation region

$$A \le \mathrm{LLR}_M \le B$$

where

(1) A, B depend on the error probablities α, β in the usual way.
(2) $\hat{\phi}_M$ is the maximum likelihood estimator of ϕ after M observations.
(3) $V(\hat{\phi}_M)$ is an estimate for the variance of $\hat{\phi}_M$.

We will consider the special case that

$$\hat{\phi}_M = \frac{X_1 + \cdots + X_M}{M} \tag{5.5}$$

where X_i are identically distributed independent random variables with mean ϕ and variance σ^2. Then we get

$$\mathrm{LLR}_M = \frac{(\phi_1 - \phi_0)(\sum_{i=1}^{M}(X_i - (\phi_0 + \phi_1)/2))}{\sigma^2}$$

So $(\mathrm{LLR}_M)_M$ may be approximated by a Brownian motion with drift m and infinitesimal variance s^2 given by

$$\begin{aligned} m &= \frac{(\phi_1 - \phi_0)(\sum_{i=1}^{M}(\phi - (\phi_0 + \phi_1)/2))}{\sigma^2} = s^2\bar{\phi} \\ s^2 &= \frac{(\phi_1 - \phi_0)^2}{\sigma^2} \end{aligned}$$

where $\bar{\phi}$ measures the relative position of ϕ in the interval $[\phi_0, \phi_1]$ in such a way that $\bar{\phi} = -1$ corresponds to $\phi = \phi_0$ and $\bar{\phi} = 1$ corresponds to $\phi = \phi_1$. The formula for the expected boundary crossing time of such a Brownian motion is

$$N = -m^{-1}\frac{-Ae^{-\gamma B} + Be^{-\gamma A} - (B - A)}{e^{-\gamma B} - e^{-\gamma A}}$$

where

$$\gamma = 2\frac{m}{s^2} = 2\bar{\phi}$$

So

$$N = \left(\frac{\sigma}{\phi_1 - \phi_0}\right)^2 \times [\cdots] \tag{5.6}$$

where $[\cdots]$ depends only on $\bar{\phi}$.

Now we assume that we are in the situation of Theorem 5.1.4 so that[2] $\phi = \phi_C = \mathrm{elo}_C(X, Y)$. By the weak dependency hypothesis we have

$$\frac{\mathrm{elo}_D(X, Y)}{\sigma_D} = \mathrm{NE}_{D/C}\,\frac{\mathrm{elo}_C(X, Y)}{\sigma_C}$$

It follows that to satisfy the conditions (5.3) we should put

$$f = e\,\mathrm{NE}_{D/C}\,\frac{\sigma_D}{\sigma_C}$$

Since all we are doing is a rescaling it is clear that $\phi_D = \mathrm{elo}_D(X, Y)$ satisfies $\bar{\phi}_D = \bar{\phi}_C$ (note that $\bar{\phi}_C$ is computed with respect to the bounds $[0, e]$ and $\bar{\phi}_D$ is computed with respect to the bounds $[0, f]$). From (5.6) we get

$$\begin{aligned} \frac{N_S}{N_T} &= \left(\frac{\sigma_C}{e}\right)^2 \Big/ \left(\frac{\sigma_D}{f}\right)^2 \\ &= \mathrm{NE}_{D/C}^2 \end{aligned}$$

[2] Small elo differences are proportional to the difference between the win and the loss ratio. So (5.5) holds in good approximation.

which finishes the proof.

References

[1] Michel Van den Bergh, *A practical introduction to the GSPRT*, `http://hardy.uhasselt.be/Toga/GSPRT_approximation.pdf`.

[2] Chess Programming WIKI, *Match statistics*, `https://www.chessprogramming.org/Match_Statistics`.