The summary below was written by Paul Matthews, author of the rating system. It comes as part of the software for the "Accelerated Pairing System" which is a practical and equitable system for pairing players in tournaments.

INSIDE THE AGA RATINGS SYSTEM

7/28/90

Paul Matthews, Princeton Go Society

INTRODUCTION

Questions about ranks and ratings, who's really stronger, and how one part of the world compares with another, probably have no once-and-for-all-time answers. Local, national and international traditions evolve, players enter and leave active competition, the general level of go knowledge increases, and new champions appear. Yet there is a persistent interest in having some kind of measurement and recognition of playing strength. The AGA approach for many years has been to publish ratings, numbers on a continuous scale that can be equated roughly to traditional amateur ranks, but that reflect the ups and downs of competitive play.

In 1988 and 1989, the AGA ratings system was extensively overhauled. Phil Straus, Paul Matthews, Bob High, Steve Fawthrop, Laurie Sweeney, Richard Cann, Bruce Ladendorf, Nick Patterson, and others, contributed mightily of their time and expertise to launch the new system. Although the initial goal was to correct logical inconsistencies that had crept into the old system, the bulk of the work turned out to be concerned with data integrity, tournament reporting practices, computer software development, and proving to each other that the new system really worked. The present article takes an inside look at the new system.

NUMERICAL SCALE

Ratings are expressed on a scale 100 and up for dan level players, and -100 and down for kyu level players. Dividing a rating by 100 yields the rank equivalent; thus, 276 is a 2 dan rating, and -432 is 4 kyu. Because there is no rank between 1 kyu and 1 dan, there are no ratings between -100 and 100, which can be confusing when doing ratings arithmetic.

When a player first enters the system, his or her self-declared rank is translated to a provisional rating. For example, 6 dan is translated to 650, and 1 kyu to -149. Ratings adjust quickly, so that a new player reaches the right level in just a few tournaments, and no player's rating gets stuck; this is one of the improvements over the old system.

CREDIBILITY

Your AGA rating does not tell you precisely how strong you are. What it does tell you is how you stand relative to other players based on your recent performance in tournaments and other rated events. Your perception of your strength is based on more games than are rated, and you may be more accurate, particularly if you have been playing at about the same level for several years. However, if your estimate differs radically from your AGA rating, say higher by as much as 200 points, then most players would agree that you have something to prove, and be quite willing to give you the chance! Discrepancies of up

to 100 points are within the range of statistical error, but if your rating were chronically 100 points below your claimed rank, then you ought to reassess the strength of your play.

Be aware that many of your opponents may exaggerate their rank. In tournaments, players often enter at a higher rank to gain experience. But the ratings system sees them as they are, and consequently, your victories may not gain as many rating points as you think they should, and your losses may be more serious. In the United States, about one third of the players who claim ranks between 6 kyu and 3 dan have ratings that are one or more ranks lower. However, the ratings of players below 6 kyu and above 3 dan agree remarkably well with their claimed ranks.

STATISTICAL MODEL

A statistical model is indispensable to avoid logical inconsistencies and to do ratings arithmetic properly. In common with the Elo system used internationally in chess, the AGA model expresses the probability of winning a game as a function of rating difference. This so called "percentage expectancy" curve, PX, is represented as a normal probability distribution function with standard deviation px_sigma. Working backward from this assumption, it is possible to infer likely rating differences given actual game results.

One problem this approach must address is to estimate a rating difference based on a single game, or any set of games where one player always wins. The mathematics of simple maximum likelihood estimation would suggest that the winning player is likely to be infinitely stronger than the loser! Given that most games are approximately evenly matched, this inference is obviously unreasonable, and ignores the fact that we have some prior knowledge about the players. The AGA system uses Bayesian statistical methods to solve the problem. The essential idea is to capture the notion that players are probably about the strength they say they are; the technical device is a normal probability density function, called the "rating prior," RP, centered on the player's presumed rating and with standard deviation rp_sigma. For one game, the Bayesian likelihood is of the form,

*likelihood(outcome) = RP(rating1) * RP(rating2) PX(outcome | rating1 - rating2)*

At some point, the increase in PX likelihood as the estimated ratings of the two players spread apart is balanced by decreases in player RP likelihoods as ratings are stretched farther from the players' prior presumed strengths; new ratings are defined by the balance point where likelihood is at a maximum. The magnitude of the rating change is determined by rp_sigma, larger values allowing larger movements.

For multiple games, the RPs for all the players, and the PXs for all the games, are multiplied together to obtain the overall likelihood. This connects the ratings of all players together in a network of interlocking games, and improves the stability and accuracy of ratings compared with updating ratings one game at a time. The maximum Bayesian likelihood is found numerically by simultaneously adjusting all the ratings until the best (i.e., most likely) combination is found.

PARAMETER VALUES

The current values of the AGA ratings system parameters are shown in the table below. A px_sigma value of 104 implies that a player who is stronger by a full rank (i.e., 100 rating points) should win about 83% of the time; the percentage for two ranks is 97%. The value

of px_sigma was chosen, based on the analysis of thousands of games, to be consistent with the model that the rating point equivalent of an n stone handicap is 100n.

RATINGS SYSTEM PARAMETER VALUES

Rating System Parameters

px_sigma = 104

rp_sigma = 80

Rating point equivalents of handicaps:

*50 - 10 * komi, if stones = 0*

*100 * stones - 10 * komi if 2 <= stones <= 9*

where -20 <= komi <= 20

Rp_sigma expresses the uncertainty associated with old ratings; in practice, rp_sigma controls the volatility of ratings. The current default value of 80 was chosen so that the average rating point value of a single game is 30, which limits the expected maximum gain in a five round tournament to 150 rating points. Simulations showed that both large and very small values of rp_sigma work poorly, leading to severe fluctuations or stagnant ratings respectively.

The rating point equivalent of no komi, the so called "one stone" handicap, is significantly less than 100, a fact that was also recognized in the old ratings system. The rating point values of other komi handicaps is an interesting topic for future statistical investigation. The data that is currently available, much of it provided by Wayne Nelson, suggests that every point of a komi compensates for about 10 rating points. Thus, since the value of the first move (i.e., taking Black) is about 50 rating points, a reverse komi of 5 1/2 points should come close to compensating for a full rank difference.

IMPROVING PLAYERS

Many players believe that they are growing stronger, and are annoyed if their rating lags behind their self assessment. The default value of rp_sigma seems sufficient for routine rating adjustments; however, a rapidly improving player may play at a rank several hundred points above his or her old rating, and a boost is needed. Players who declare a rank more than 50 points higher than their rating, have the mean and standard deviation parameters of their RP function increased. By adding points to the RP mean, points are added to the whole system, helping to counteract the tendency for the ratings of stable players to deflate as other players improve. The larger standard deviation allows an improving player's rating to float more freely, upward or downward, and have less effect on the ratings of opponents. Note that a player who performs poorly when playing above his or her rating risks a larger loss of rating points.

SOFTWARE

The AGA ratings system is a suite of programs implemented (in C) for IBM PC compatible machines running DOS. The ratings system software has been extended to provide on-site support for a wide variety of handicap and championship tournaments, both small and large. Now tournament directors can generate on-the-spot ratings based on entry ranks

and tournament games, and can even use the ratings to do pairings and figure out the tournament winners! These extensions are called the "Accelerated System." Significant effort also is being devoted to software that supports the verification and correction of AGA ID#s and names, preferrably at the tournament site.

FUTURE WORK

The revitalized AGA ratings system is a world class system that is a credit to the AGA and the go world. But it will never be perfect, and work continues. Phil Straus, the AGA Ratings Commission chairperson, is doing a super job in coordinating and motivating many activities relating to ratings. Some of the areas that are currently being addressed are: a comparison of ranks in foreign countries with AGA ratings; rating the games of professional players; and better tournament practices to improve data integrity.