Probability Matching, the Magnitude of Reinforcement, and Classifier System Bidding

DAVID E. GOLDBERG   (dgoldber@ualvm.ua.edu)
The University of Alabama, Tuscaloosa, AL 35487

Abstract. This paper juxtaposes the probability matching paradox of decision theory and the magnitude of reinforcement problem of animal learning theory to show that simple classifier system bidding structures are unable to match the range of behaviors required in the deterministic and probabilistic problems faced by real cognitive systems. The inclusion of a *variance-sensitive bidding* (VSB) mechanism is suggested, analyzed, and simulated to enable good bidding performance over a wide range of nonstationary probabilistic and deterministic environments.

Keywords. Classifier systems, genetics-based machine learning, credit assignment, conflict resolution.

1. Introduction

Classifier systems (CSs) are genetics-based machine learning systems that combine syntactically simple rules called *classifiers*, parallel rule activation, rule rating and conflict resolution by analogy to a competitive service economy, and genetic algorithms (GAs). Although classifier systems were originally suggested quite some time ago (Holland, 1971), and despite their existence for over a decade (Holland & Reitman, 1978), the development of usable theory of classifier system operation has lagged behind CS simulation efforts, especially when compared to the development of applicable theory for genetic algorithms in search and optimization. In this paper, I connect two important questions of learning and decision theory—the *paradox of probability matching* and the *magnitude of reinforcement problem*—to an important question of classifier system bidding structure. The connection suggests a new direction for classifier system design.[1]

In the remainder of this paper, I restate the paradox of probability matching, suggest a plausible explanation for its occurrence, and discuss its connection to the two most commonly used classifier system bidding structures: the noisy auction and the weighted roulette wheel. The magnitude of reinforcement problem is then discussed in the context of these same two bidding structures. The juxtaposition of the two problems immediately suggests the need for a modified classifier system design that incorporates *variance-sensitive bidding* (VSB). An equilibrium analysis, a cursory analysis of transitions, and some computer simulations are presented to demonstrate the utility of this mechanism in mixed probabilistic-deterministic environments.

2. The paradox of probability matching

Suppose you are faced with a decision between two mutually exclusive alternatives. For example, suppose you must guess whether a friend is going to wear a red necktie or a blue necktie, and further suppose the friend chooses red with probability p and blue with

probability $1 - p$. At first you are uncertain as to the value of the red-tie probability p. Over time, you gain experience with your friend's preferences in cravat color and determine with reasonable certainty an estimate of his preference probability. Thereafter, you are faced with a question: what strategy should you adopt if you want to maximize the expected number of correct predictions of tie color? According to decision theorists, the answer is straightforward: simply select the tie with maximum probability of occurrence:

if $p = \max\{p, 1 - p\}$ **then** decision = red

else decision = blue.   (1)

That this is the optimal strategy once p is known can be shown by assuming that you choose the red tie alternative with probability p' and the blue tie with probability $1 - p'$. Maximization of the expected value of the proportion of correct decisions

$$D = pp' + (1 - p)(1 - p')$$ (2)

by elementary means yields a maximum when either p' or $1 - p'$ is unity, depending whether the red or blue tie has maximum probability of being selected. Of course in choosing this way, you will be correct a proportion $\max\{p, 1 - p\}$ of the time. In the remainder of this discussion, we assume that the red tie is preferred ($p > 0.5$). We will also call this all-or-nothing strategy the *decision theory* (DT) solution.

The preceding calculation seems straightforward and hardly open to question, but when human subjects are placed in a similar decision-making situation they do not follow the all-or-nothing allocation strategy of DT. When faced with a similar binary prediction task, numerous experiments have repeatedly shown that human (Lee, 1971) and non-human (Mackintosh, 1974) decision makers seem to choose randomly between the two alternatives using a probability of selection roughly equal to the probability of presentation ($p' = p$). This behavior has been called *probability matching*, and a simple calculation shows that the expected payoff under this strategy is less then the amount that can be obtained by selecting the better alternative. If a subject adapts his probabilities to match the reinforcement probability, the expected number of correct decisions under probability matching, D_{pm}, may be calculated as

$$D_{pm} = p^2 + (1 - p)^2.$$ (3)

Figure 1 shows a plot of the ratio of payoff under probability matching to that expected under the decision theory approach. For the worst case (when $p = \sqrt{2}/2$), probability matching is at a 17% disadvantage compared to the decision theory approach. That humans and other animals should select such *sub-optimal* behavior has bothered psychologists and decision theorists for some time, and numerous theories have been developed to explain these supposed *irrationalities* of animal and human behavior.

For example, Flood (1954) suggested that human subjects do not recognize the sequence as random, and further suggested that had they known this, they would have picked the decision theory (DT) solution. Unfortunately, subsequent experiments have lent only limited
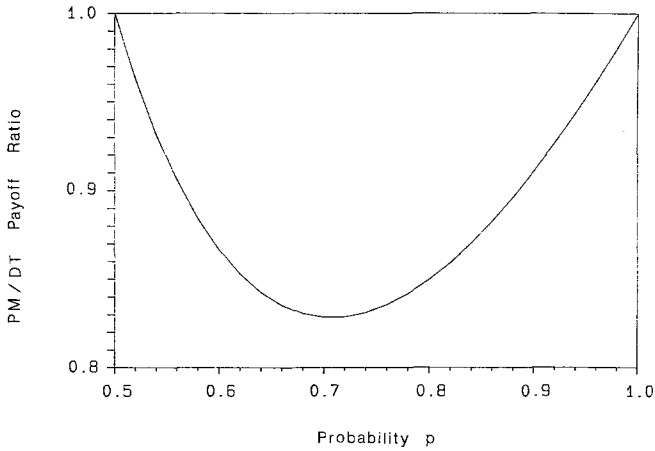
Figure 1. The ratio of payoff under probability matching to payoff under decision theory is plotted as a function of environmental probability *p*.

support to this hypothesis. DT-like results (with varying rates of error) were obtained with humans only after supplying subjects with specific instructions to perform in a manner very close to the DT solution (Goodnow, 1955; McCracken, Osterhout, & Voss, 1962). Others have tried to explain PM by inventing utilities for activities that would justify the behavior. For example, Siegel (1959) assumed that a subject is motivated to be both correct and not bored. By introducing a utility value for boredom-alleviating caprice, Siegel calculated a utility function that justified the PM response somewhat. Though he had some success, certain parameters in this composite utility function were varied over the range of probabilities *p*, leaving the model open to question.

3. A different view

It is strange—at least, interesting—that a number of authors have sought to explain PM as a capricious, if not improper, act. What happens instead if we take the opposite tack and view the behavior as correct? Can we uncover any holes or loose assumptions within the problem definition that might allow us to understand the circumstances under which PM behavior should be preferred to its DT counterpart? After all, humans (and many other organisms) are a competitive lot. It is difficult to believe that such exploitative beings would leave a large window of payoff open for others with no good reason. Furthermore, the selective pressures that developed *Homo sapiens* over 3.5 to 4.5 billion years of evolution could hardly have failed to fill such gaping competitive holes with more cost-conscious beings.

Unfortunately, there do not appear to be too many places we can look for flawed assumptions in the PM problem statement. We can question our unquestioned use of knowledge regarding the probability *p* when we are, at first, uncertain of its value. After all, we must estimate *p* and then make our decision, and in a finite number of trials we will never know

p with certainty. This, of course, is the tradeoff between *exploration* and *exploitation* that is faced whenever the results of a search must be used to some practical end; however, in the present case, this hole in the problem statement is little help in explaining the choice of probability matching over the decision theory approach: it can be shown that the optimal decision under uncertainty allocates greater than exponentially increasing numbers of trials to the observed better alternative as the number of trials increases (Holland, 1973, 1975). After sufficient experimentation, this calculation agrees with the decision theorists: simply give the preponderance of the trials to the better alternative.[2]

Fortunately, there is one other place we can turn to combat the label of impropriety, so unceremoniously pinned upon ours and other species by the paradox of probability matching. Returning to our necktie example, what would happen if our friend changed his mind and started wearing blue ties more frequently? Put another way, what if we no longer accept as reasonable the previously hidden assumption of environmental stationarity? Instead, we consider the possibility that the preference probability varies as an unknown function of time $p = p(t)$.

For the sake of concrete argument, let's assume that our friend simply switches his red and blue tie preference probabilities. At the moment of this switch, we are not aware of the change and through careful observations are able to readjust our preference probability estimate. During this readjustment period we assume that the environmental switch has occurred, but we have not corrected yet for the change. Thus regardless of whether we are using a DT or PM approach, during the readjustment period, we make a higher level of errors than we were making prior to the switch. In the case of the decision theory approach, we are correct during this readjustment period a proportion D'_{dt} as given by the equation

$$D'_{dt} = (1 - p). \tag{4}$$

If we are using probability matching, the proportion of correct decisions would be given by the equation

$$D'_{pm} = 2p(1 - p), \tag{5}$$

which may be obtained by recognizing that there are two equally probable ways of getting the right answer. Thus, during this readjustment phase, the probability matching approach has the upper hand because $D'_{pm} > D'_{dt}$ or

$$2p(1 - p) > 1 - p, \tag{6}$$

whenever $p > 0.5$ (as was previously assumed). This draws us toward an interesting conclusion. If the environment of decision is changing such that it is possible for the environment to go against our current thinking, then probability matching can hold the upper hand during the readjustment phase. This possibility permits the calculation of the circumstances under which PM has an advantage over DT. This is precisely the calculation performed in the next section.

4. A simple analysis

In this section, we consider a symmetric nonstationary environment (a switching environment) where preferred becomes less preferred during alternating periods of time. It is important to recognize that there is no need to assume such a drastic change, however. A simple calculation shows that a cost advantage is maintained for PM during the readjustment phase whenever the environment goes against the current preference (whenever p falls below 0.5). The switching environment is a simple abstraction of a capricious environment that changes its mind, however, and our simple model will demonstrate the important points of any such assumed changes. Notice that we have shifted the presence of caprice from the learning subject (as was assumed by Siegel) to the subject's environment. This simple shift places the burden of whimsical behavior where it belongs and allows us to understand more clearly why subjects might be willing to pay insurance premiums as a hedge against the uncertainty of their environments.

To begin the analysis, consider the situation posed in the time line of Figure 2. Here we assume the environment will switch every T_r units of time, and the decision theory approach takes T_c units of time to correct itself. We further assume that the probability matching approach takes αT_c units of time to correct with $0 < \alpha \leq 1$. Defining ξ as the ratio of DT correction time to environmental reversal time $\xi = T_c/T_r$, we may calculate the breakeven point for the probability matching approach versus the decision theory approach as follows:

$$(1 - \alpha\xi)[p^2 + (1 - p)^2] + 2\alpha\xi p(1 - p) \geq (1 - \xi)p + \xi(1 - p). \qquad (7)$$

Letting $q = 1 - p$ and solving for the breakeven ξ value, we obtain the following inequality:

$$\xi \geq \frac{p - (p^2 + q^2)}{\alpha[2pq - (p^2 + q^2)] + 2p - 1} . \qquad (8)$$
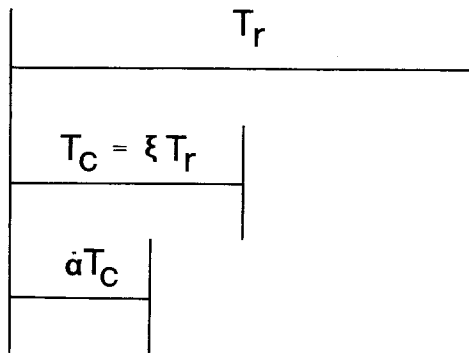

Figure 2. A time line shows the relationship between the reversal (switching) time T_r, the DT correction (readjustment) time $T_c = \xi T_r$, and the PM correction time αT_c in a symmetrical switching environment.

We may solve for some important special cases quite directly. Examining the case when correction times for probability matching and decision theory approaches are equal (when $\alpha = 1$), we obtain the result that $\xi > 0.5$. In other words, if DT and PM approaches require comparable times to correction, then that correction time must be at least one half of the switching time in order for probability matching to have an overall cost advantage over the decision theory approach.

It may also be reasonable to assume that probability matching might have an advantage in readjusting to an environmental switch because it has a head start (its bet is already hedged). Taking this argument to the extreme (setting $\alpha = 0$) and solving for the breakeven value of ξ yields the computation

$$\xi > 1 - p. \tag{9}$$

Figure 3 shows the two limiting curves along with curves for intermediate values of α. It is clear from this curve that only reasonable speed increases need be assumed to get ξ values that make probability matching the more profitable approach.

Additionally, a hidden assumption in the above calculations may further understate the case for probability matching. The calculations assume that both approaches readjust and eventually correct themselves, but this may not be an appropriate assumption for a DT decision maker or automaton. Recall that the idealized DT decision maker estimates the preference probabilities and then gives all additional trials to the more frequent alternative. For many implementations of reward estimator (such as those found in classifier systems,
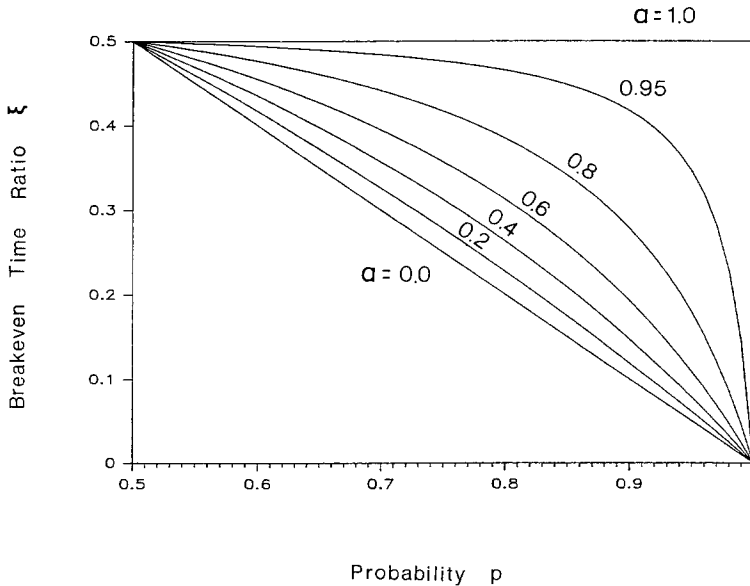

Figure 3. The ratio of DT readjustment time to the environmental switching time ($\xi = T_c/T_r$) is plotted versus the environmental probability p at different ratios of PM-to-DT readjustment time (α).

learning automata, and neural networks), it is not difficult to imagine switching environments that fool an overly decisive automaton into making the wrong decision indefinitely (even if the automaton continues to estimate its alternative probabilities). Under such circumstances, the calculations of this section are overly conservative, and the payment of the relatively small insurance premium of PM to avoid improper convergence is essential if the system is not to get stuck after an adverse environmental shift.[3]

5. Do CS bidding structures permit probability matching?

The previous section demonstrates how the indecision of probability matching may be viewed as a reasonable strategy when the environment of decision is itself uncertain. In a sense, the PM decision maker is willing to pay a small insurance premium (up to 17% of the potential reinforcement) to guard against either slowed recovery resulting from an environmental shift or, worse yet, possible long term convergence to the worse alternative following an adverse environmental shift. This is a crucial point in all reinforcement learning systems, because the system's decision and its ability to get feedback are completely coupled. This contrasts sharply to supervised learning systems, where multiple decisions may be made simultaneously, knowledge of the correct decision is assumed to exist, and this knowledge may be used to correct all errors. In this light, it is important to compare different classifier systems to see whether they can exhibit PM behavior.

Different classifier systems vary in their implementation details, but most have three main components (Goldberg, 1989):

1. rule and message system;
2. genetic algorithm;
3. apportionment of credit system.

The rule and message system implements the raw syntactical capability of the system to process information. Although a number of rule-and-message variants have been used and many more are possible, rule syntax is not an issue here. The genetic algorithm is the primary rule discovery heuristic used in classifier systems, and the details of GA operation are covered in standard references. The GA is also not an issue here, because we assume that the competing rules have already been discovered. The primary object of attention in this study is the apportionment of credit system. More specifically, the combination of strength adjustment and bidding mechanism sometimes called the *bidding structure* is the focus of concern. To study this, we consider the following simple situation, where two rules, R1 and R2, exist in a classifer store:

$R1$: **if** [prediction required] **then** [predict red tie];

$R2$: **if** [prediction required] **then** [predict blue tie]; (10)

As per usual (Goldberg, 1989), classifiers are assumed to bid in proportion to their strength, and strength is incremented in the amount of their receipts. Mathematically, bid B_i and the strength S_i of the ith classifier vary as follows:

$$B_i(t) = cS_i(t), \tag{11}$$

$$S_i(t + 1) = (1 - c)S_i(t) + R_i(t), \tag{12}$$

where c is the bidding coefficient, R_i is the total reward received, and t is an iteration index (the strength calculation assumes that the ith classifier has won the bid and has been permitted to post its message; otherwise the strength value remains unmodified or is reduced by a small tax). Analysis of the steady state strength and bid values yields the following calculation:

$$B_i(ss) = R_i(ss) \tag{13}$$

$$S_i(ss) = R_i(ss)/c, \tag{14}$$

where ss is a sufficiently high time index value where steady state may be assumed to have been reached. In the present case, the reward at steady state is simply the expected value of reward for the ith arm (simply the preference probabilities p and $1 - p$).

This description of classifier system operation applies to different classifier systems, with minor adjustments in notation; however, the selection of decisions from the bids—the CS's bidding structure—has been a source of diversity in different classifier systems. At two ends of the spectrum are the noisy auction suggested in my work (Goldberg, 1983, 1989) and the roulette wheel section suggested elsewhere (see Holland & Reitman, 1978; Wilson, 1987).

In the noisy auction an effective bid EB_i is made by the ith classifier by simply adding zero-mean Gaussian noise to the classifier's bid B_i:

$$EB_i = B_i + G(\sigma_n), \tag{15}$$

where G is a Gaussian (normal) noise generator and σ_n is the standard deviation of the noise (a system parameter).

In roulette wheel selection, a probability of selection for the ith rule is calculated as follows:

$$P_i = B_i/\Sigma \, B_j, \tag{16}$$

where the sum is taken over all bidding classifiers j. Thereafter, the probability distribution determined by the P_i values is used to determine a winner.

In the red-blue tie problem, the roulette wheel mechanism is particularly easy to analyze. The expected steady-state bid of the two classifiers is simply

$$\langle B_1 \rangle = p, \tag{17}$$

$$\langle B_2 \rangle = 1 - p, \tag{18}$$

and since roulette wheel selection chooses according to bid, this bidding mechanism must follow probability matching behavior. A more detailed analysis treats the bids as random

variables, but, as we shall soon see, the bid variance may be made arbitrarily small by judicious choice of the bidding coefficient c.

The noisy auction may also be analyzed. The expected bid values are the same as for the roulette wheel case. The variance values for the bids may be calculated directly by summing the squares of the coefficients

$$\text{var}(B) = c^2\sigma^2\sum_{j=0}^{\infty}(1 - c)^{2j}, \tag{19}$$

where c is the bidding coefficient and σ^2 is the variance of the receipt ($p[1 - p]$ for a Bernoulli trial). The infinite sum may be calculated since if $I = \Sigma(1 - c)^{2j}$, then $I(1 - c)^2 = I - 1$. Thus $I = 1/[c(2 - c)]$. Thereafter, the variance of the bid may be calculated as

$$\text{var}(B) = \frac{c}{2 - c}\sigma^2. \tag{20}$$

The variance of the bid decreases with decreasing c, an intuitive result if we recognize the bidding coefficient as being responsible for the number of time steps over which the memory of the system is averaged.

Since we may make the variance of the bid as small as we wish, we analyze the noisy auction assuming that the bids are perfect estimators of the expected payoff. Whether rule one or two wins the noisy auction depends strictly on the expected values for the bids and the standard deviation of the bidding noise. Since Gaussian noise has been added to both signals, this probability may be calculated as the convolution of the two distributions such that the alternative with higher mean has lower value than the alternative with lower mean. Recognizing that the difference of two Gaussian distributions is Gaussian, we obtain the result that the probability of the signal with lower mean winning is the same as the probability of a normally distributed random variable with mean $2p - 1$ and variance $2\sigma_n^2$ having a value less than zero. Using this relationship to analyze the noisy auction at its extremes, we recognize that a range of behavior from DT-like to random guessing can be obtained depending upon how the noise parameter σ_n is set. With $\sigma_n = 0$, the noisy auction simply picks the better rule; in other words, it executes DT behavior. With σ_n very large, the alternatives are indistinguishable and the noisy auction essentially tosses a coin. At an intermediate value of σ_n, the noisy auction can be made to match probabilities. Setting $z = (2p - 1)/\sqrt{2\sigma_n}$, and setting the value of the cumulative unit Gaussian distribution at z equal to the environmental probability p as

$$p = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}e^{-\zeta^2/2}d\zeta, \tag{21}$$

we may then solve for the values of σ_n that yield PM at different p values (Table 1). Thus, we see that the noisy auction can emulate the probability matching solution if the noise is made sufficiently large. Note that over the range of probability values the noise values change only slightly, meaning that a PM-like solution can be obtained over the range of problem probabilities without much manipulation of the noise parameter.

Table 1. Bid noise values required for probability matching under the noisy auction

p value	σ_n value
0.6	0.559
0.7	0.540
0.8	0.504
0.9	0.441

Table 2. Noisy auction simulation results with PM σ_n values

p	Simulated	|Difference|	Std. Dev.
0.6	0.5904	0.0096	0.0069
0.7	0.6968	0.0032	0.0065
0.8	0.7990	0.0010	0.0057
0.9	0.9052	0.0052	0.0042

To test this suggestion that the noisy auction can match probabilities, computer simulations of the noisy auction have been run using the σ_n values of Table 1. A bidding coefficient $c = 0.001$ has been chosen to minimize the variance contributed by fluctuations in the bid values B_i. Initial strength values have been set at their steady values, $S_i(0) = p_i/c$. The time-averaged proportion of red trials at $t = 5000$, their absolute deviation from the probability matching value, and their expected standard deviation values are shown in Table 2. Here the standard deviation in proportion is conservatively calculated as that we should expect from a sequence of 5000 Bernoulli trials, Std. Dev. $= \sqrt{p(1 - p)}/5000$. All runs are within two standard deviations of probability matching as expected.

6. Counterpoint: The magnitude of reinforcement problem

We have just seen how the two major types of bidding structure can model PM decision-making behavior in a probability learning problem. In the case of roulette wheel section, probability matching behavior is inevitable. For the noisy auction, PM-like behavior can be obtained over a range of problems through the judicious choice of bidding noise, but this behavior is not hard-wired into the auction procedure. This immediately raises the question whether there is ever any motivation to depart from PM-behavior and move toward the more decisive approach of decision theory.

The answer from psychological experiments, as has already been hinted, is that PM-behavior in humans is no absolute state of affairs. Real payoff and assurances that the environment is stationary can cause human subjects to move toward DT-like behavior (Lee, 1971). If real subjects can adjust the degree to which they hedge their bets, perhaps there is some motivation to have a similarly adaptive mechanism within a classifier system.

To put this in closer perspective, consider a deterministic binary decision problem where the *magnitude of reinforcement* assigned to the two alternatives differs. For example, suppose two paths are presented to a subject in a T-maze. When the subject chooses path one, he receives payoff r_1; when he chooses path two, he receives a payoff r_2. If we normalize

the payoff values so that $r_1 + r_2 = 1$, we have a problem that connects directly with the probability learning problem discussed earlier in that both problems can be made to have the same expected reward, but in the magnitude of reinforcement problem all uncertainty is removed from the picture. It might be tempting to assume that this difference between the two problems is negligible, and we might reason by analogy that the PM solution to the probability learning problem carries over to the magnitude of reinforcement problem, but when real decision makers are actually observed, it is somewhat surprising to discover that this originally surprising result is not observed in magnitude of reinforcement experiments: in the magnitude of reinforcement problem natural subjects *do* follow a DT strategy. Specifically, it has been observed for varying ratios of reinforcement that animals presented with a binary T-maze learn to allocate all of their trials to the observed better path (Mackintosh, 1974). It is true that it takes animals longer to learn to choose the better alternative when the difference in payoff is small, but eventually they allocate the preponderance of trials to the observed best.

As before, we need to ask whether our two prototypical bidding structures can emulate the observed behavior of natural decision makers in the magnitude of reinforcement problem. Since uncertainty has disappeared from consideration, the performance of roulette wheel selection and the noisy auction may be derived directly. Roulette wheel selection achieves steady state bids for the two rules equal to the reinforcement:

$$B_i = r_i. \tag{22}$$

Thereafter, winners are chosen as the proportion of an individual rule's bid to the total bid: roulette wheel selection follows a deterministic equivalent of the probability matching solution.

On the other hand, the noisy auction can be made to follow the behavior of natural decision makers in the magnitude of reinforcement problem as closely as desired simply by setting the bidding noise σ_n close to zero. The expected bid of a rule is again simply the reinforcement it receives. As soon as the reinforcement is estimated with reasonable accuracy, the not-too-noisy auction chooses the better of the two alternatives. Of course, there may be some motivation for retaining some small amount of noise to insure that there is some non-zero probability of occasionally trying the worse alternative.

7. A new direction: variance-sensitive bidding

At first, the juxtaposition of probability learning and the magnitude of reinforcement problem is puzzling. Natural subjects seem to adapt the certainty of their decision-making and learning to the level of certainty in their environment. This results in the selection of a range of behavior varying from DT-like to PM-like. Yet, in examining different classifier system bidding structures, we have seen how these do not adapt to environmental uncertainty. It is true that both of the classifier system bidding structures examined herein can emulate probability matching behavior. In the case of the roulette wheel, this occurs automatically, and in the case of the noisy auction, it can occur with the addition of a fair amount of noise to each bid. On the other hand, of the two bidding structures we have examined,

only the noisy auction can easily emulate the behavior of natural decision makers in the magnitude of reinforcement problem. The noisy auction can give the preponderance of its trials to the observed best after sufficient learning if the bid noise is set close to zero, whereas the roulette wheel must stick with an allocation of trials in proportion to rule rewards. Thus, if adapting to a *range of environmental uncertainty* is important—as it often is—the noisy auction appears to be the more flexible of the two procedures,[4] and in the remainder of the paper it is considered exclusively.

In normal practice, this flexibility of the noisy auction has not been realized. For example, in my dissertation work (Goldberg, 1983), I chose a compromise value for the bid noise and held that value constant for all time across all rules in the population. If, as the juxtaposition of the probability learning and magnitude of reinforcement problems suggests, we require a range of behaviors from PM-like to DT-like, then setting the noise at any fixed value is incorrect. No single noise setting can model the observed behavior of natural decision makers over a range of problems. In order to achieve the necessary flexibility, I propose the addition of *variance-sensitive bidding* (VSB) to future classifier systems.

In one possible implementation of VSB, let each classifier maintain one other statistic in addition to its strength: a variance estimate V. After a reward cycle, calculate the squared difference between a rule's bid and its receipts and update the variance estimate as follows:

$$V_i(t + 1) = (1 - b)V_i(t) + b[R_i(t) - B_i(t)]^2, \; 0 < b \leq 1. \tag{23}$$

In this way, the variance estimate maintains a geometrically weighted average of the squared difference between a rule's bid and its actual reinforcement. In turn, the variance estimate may be used in the noisy auction to calculate each rule's effective bid:

$$EB_i(t) = cS_i(t) + G[\beta\sqrt{V_i(t)}], \tag{24}$$

where the effective bid EB is calculated as the sum of a rule's usual bid, $cS(t)$, and the zero-mean Gaussian noise with standard deviation $\beta\sqrt{V_i(t)}$, where β is a system constant. In this way, the noisiness of the auction is controlled by the variation of the actual-to-expected receipts of participating rules.

7.1. An equilibrium analysis of variance-sensitive bidding

At equilibrium in a probability learning problem, the variance estimate V approaches the variance of the receipts, $V \to p(1 - p)$. Taking the difference between the expected bid of the two rules and dividing by their β-adjusted standard deviation value, the unit normal random variable z may be calculated as follows:

$$z = \frac{2p - 1}{\beta\sqrt{2p(1 - p)}} \; . \tag{25}$$

Evaluating the proportion of times the better arm is selected—evaluating p_1'—is simply a matter of calculating the cumulative probability of the unit normal distribution at z:

$$p_1' = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\zeta^2/2} d\zeta. \qquad (26)$$

Figure 4 displays the proportion of better trials as a function of environmental probability p for $\beta = 1.35$. Actually, probability-matching-like behavior is predicted over a fairly wide range of β values, but $\beta = 1.35$ is adopted for the remainder of the study.

An equilibrium analysis of a VSB-augmented auction in a deterministic problem is straightforward. At equilibrium the variance estimates V approach zero, the bids approach the deterministic reward values, and the system gives all of its trials to the better alternative. In a moment, we will examine some simulations to observe DT-like and PM-like behavior in the face of environmental shifts. Before we do this, we need to consider whether the mechanism can become stuck in transitions.

7.2. A cursory analysis of transitions

A complete analysis of the transient behavior of the VSB-augmented noisy auction is beyond the scope of this paper; however some simple reasoning suggests that VSB and the noisy auction should be able to react to the four possible types of shifts between probability learning problems (*probabilistic*) and magnitude of reinforcement problems (*deterministic*):

1. Probabilistic to probabilistic
2. Probabilistic to deterministic
3. Deterministic to probabilistic
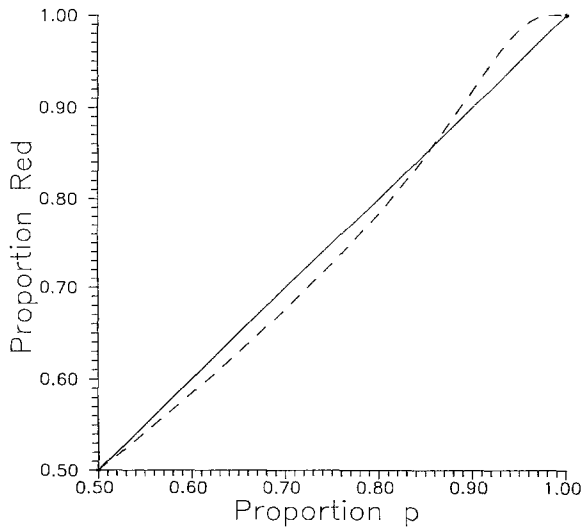4. Deterministic to deterministic

Figure 4. Near-PM performance is predicted by an equilibrium analysis over the range of probabilistic environments. The solid line is the perfect PM solution and the dashed line is the equilibrium solution for the VSB-augmented auction when $\beta = 1.35$.

In all cases, we make the conservative assumption of an adverse shift, where preferred becomes less preferred on alternating cycles.

In an adverse, probabilistic-to-probabilistic environmental shift the VSB-augmented classifier system should adapt easily. Near-PM behavior can be maintained if the β coefficient is sized to achieve the noise standard deviation values close to those of Table 1.

Similarly a shift from probabilistic to deterministic should cause no adaptation difficulty, and the existence of VSB will insure that the system moves from PM-like behavior on the probabilistic problem to DT-like behavior on the deterministic problem.

Shifts from deterministic to probabilistic should also pose no particular difficulty. At the time of the problem shift, the noise of the probability learning problem will show up in the variance estimate V. This in turn will increase the noisiness of the bid, which will encourage trial of the other alternative, thereby allowing discovery of the correct answer.

Shifts from deterministic problem to deterministic problem are the most worrisome of the lot. Under VSB, when such shifts occur, rules have stable strength values and low variance estimates. At the time of the shift, however, the previously correct rule starts to make errors with respect to its expected value. This *surprise* increases the variance estimate, which introduces enough noise to cause the required trials of the other rule. This requirement of being sensitive to variance shifts suggests that the variance averaging parameter b should be of the same magnitude or larger than the bidding coefficient c. It may also be wise to prevent the bid noise from falling below a specified value. This will ensure an occasional trial of an out-of-favor rule.

Thus, we see how a VSB-modified classifier system is expected to have little difficulty in matching the types of environmental shifts it might encounter. In the next section, some simulations verify the equilibrium performance of a VSB-augmented auction. Other simulations demonstrate typical transition performance of the algorithm.

7.3. Simulation results

In a probabilistic problem, the VSB-augmented auction achieves PM-like behavior as predicted by the equilibrium analysis of Section 7.1. Simulation results over 5000 time steps using a bidding coefficient $c = 0.001$, VSB update coefficient $b = 0.001$, VSB auction coefficient $\beta = 1.35$, initial strength values $S_i(0) = p_i/c$, and initial variance estimates $V_i(0) = p_i(1 - p_i)$ confirm the expected performance as shown in Table 3. The standard deviation of the proportion is again calculated conservatively assuming the VSB-augmented auction behaves as a sequence of Bernoulli trials, except here the predicted proportion is used in the computation instead of the probability matching value.

Table 3. Comparison of VSB analysis and simulation in a probabilistic environment ($\beta = 1.35$)

Probability p	Predicted	Simulated	|Difference|	Std. Dev.
0.6	0.5847	0.5760	0.0087	0.0070
0.7	0.6762	0.6730	0.0032	0.0066
0.8	0.7840	0.7808	0.0032	0.0058
0.9	0.9188	0.9252	0.0064	0.0039

To test the performance of the VSB-augmented auction in switching environments, we examine simulations of a probabilistic-deterministic shift and a deterministic-deterministic shift.

The VSB-augmented auction is simulated in a 0.7–0.3 probabilistic-deterministic environment with fixed shift half-period of 1000 time steps. Bid coefficient is set to $c = 0.02$, the variance coefficient is set to $b = 0.04$, the auction coefficient is set of $\beta = 1.35$, and the strength and variance values are set to the steady state values appropriate to the first environment. Larger values of c and b than those used in earlier tests are used, because here we are less concerned with keeping system variance low. Instead, the main concern in these tests is that response be fairly rapid, which requires that the half-life of the auction be less than the half-period of the shift, $-\ln2/\ln(1 - c) \ll t_{shift}$. The proportion of red trials is shown as a 50-step moving average in Figure 5. As expected, near-PM performance is achieved in probabilistic half cycles and near-DT performance is achieved in deterministic half cycles. During probabilistic half cycles, the fluctuations in moving average value are reasonable if they are compared to the magnitude of the standard deviation of the average proportion of a 50-step Bernoulli seuqence. Although it is not shown, the variance estimate during probabilistic half cycles tends to fluctuate about a value near the Bernoulli value. Deterministic variance values approach zero in a manner consistent with the results of the deterministic-deterministic shift to be shown next.

Perhaps more interesting is the simulation of the VSB-augmented auction in the deterministic-deterministic environment shown in Figure 6. Here a 50-point moving average of the proportion of red trials is shown for the 0.7–0.3 oscillating magnitude of reinforcement environment. On each half-cycle, initial transients are followed by near-DT performance. The automaton does not become stuck because the variance of surprise is measured after each shift and contributes to an increase in the bidding noise which enables the decision to shift to the now-better alternative as shown in Figure 7, a graph of the variance estimate
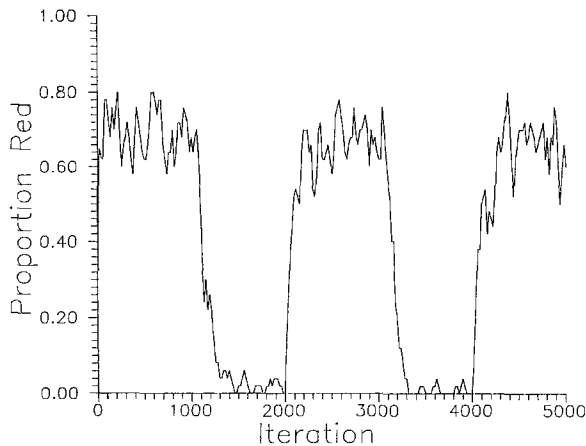

Figure 5. A 50-step moving average of the proportion of red trials shows that the VSB-augmented auction shifts appropriately in a 0.7–0.3 probabilistic-deterministic environment, approaching near-PM performance in each probabilistic half-cycle and near-DT performance in each deterministic half-cycle.
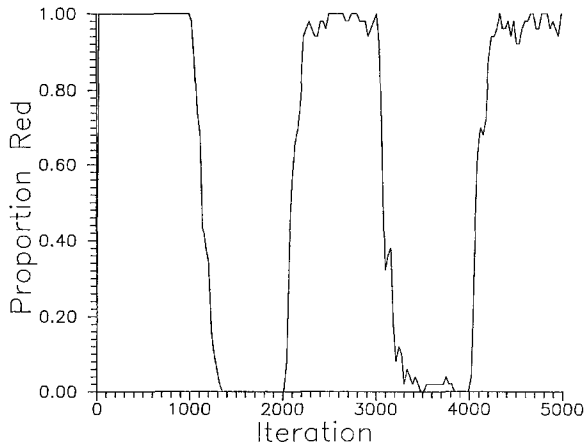
Figure 6. A 50-step moving average of the proportion of red trials shows that the VSB-augmented auction ($\beta = 1.35$) shifts appropriately in a 0.7–0.3 deterministic-deterministic environment, approaching near-DT performance during each half cycle.
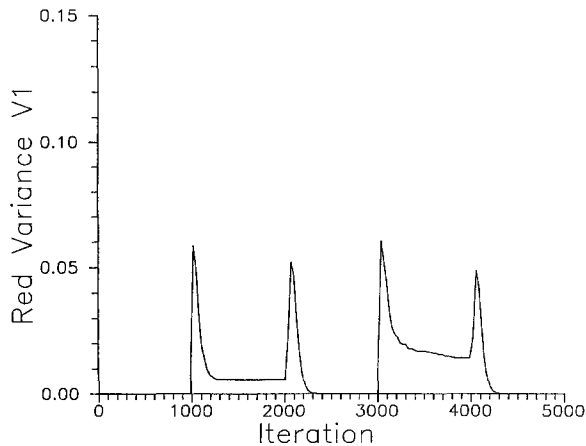

Figure 7. A time history of the variance estimate of the red rule, V_1, shows the variance of surprise that occurs at each shift. The relatively slow decay of the variance estimate is a result of the infrequent trial of the out-of-favor rule.

of the red rule, V_1 versus time. Note that the currently out-of-favor alternative maintains a fairly high level of residual variance which prevents less-than-perfect DT performance. The variance does not go to zero immediately, because the worse alternative is sampled infrequently, delaying the decay of the residual. This effect could be minimized by setting the variance coefficient much greater than the bidding coefficient—by setting $b \gg c$—but the small amount of indecision left as a result of this effect may not be particularly harmful in practical problems.

The other types of shifts have been simulated, but are not presented here, becuase they are a straightforward blend of the types of behavior demonstrated already.

7.4. Other considerations

VSB is primarily designed to permit a bidding strategy that adapts to the noise or its lack in a particular environment. It may also be useful in overcoming a knotty problem that arises in classifier systems when the genetic algorithm is used, *new rule insertion*. At present in a classifier system, when a genetic algorithm is activated, offspring rules are inserted into the population with strength values taken as a function of their parents' strength values. This has the advantage that the rules will usually get an opportunity to bid (offspring usually come from highly fit parents), but it has the disadvantage that bad rules may dominate system performance until they can be cleaned out through a number of bad bids. From a genetic standpoint this is also disadvantageous, because unused, high strength offspring can clutter the population and may dominate reproductive activity even though their strength values do not reflect their expected reward. VSB can help overcome this problem, because rules may be inserted with low strength values and high variance estimate values. Doing so will permit all offspring to win an occasional auction, but bad offspring will not dominate bidding or reproductive performance. In this way, the assignment of low strength and high variance better reflects our prior knowledge concerning a new rule's expected value to the system.

Although VSB has been proposed and has been tested in the context of a stimulus-response classifier system with a small number of rules, the mechanism should scale up to larger systems, including those with bucket-brigades or other rule-to-rule credit assignment mechanisms. Notice that the VSB mechanism is strictly local: it only considers what it expects to receive in relation to what it gets. Thus, payments between classifiers are unaffected, making the installation of VSB a relatively simple task.

8. Conclusions

In this paper, a problem of decision theory and a problem of learning theory have been juxtaposed and considered in the light of classifier system bidding structure. The paradox of probability matching and the magnitude of reinforcement problem have both been examined with the resulting conclusion that no existing classifier system bidding structure can match the range of behaviors required in the deterministic and probabilistic problems faced by most cognitive systems.

These thoughts have led to the development of the notion of variance-sensitive bidding (VSB), which allows matching of PM-like and DT-like behaviors when one or the other is appropriate. A simple implementation of VSB has been proposed and simulated, and early computational results, an equilibrium analysis, and some simple reasoning suggest that the method is able to track the correct decision in each of the four categories of environmental shift. More detailed simulations and analysis are required to confirm these initial results and to determine whether the method scales to systems with multiple rules, default hierarchies, and rule chains.

Acknowledgments

Notes

1. Although the paper addresses classifier systems most directly, the same notions may be useful in learning theory, learning automata, and neural networks. In the interest of concrete exposition, however, I stick to my knitting and develop the ideas in the context of the classifier system paradigm.
2. A slight modification of Holland's k-armed bandit formulation can achieve a PM-like allocation of trials through the formation of niches via forced sharing (Goldberg & Richardson, 1987; Holland, 1975). Simply stated, when organisms exist in a resource-limited environment they are forced to share food and other resources; this in turn limits subpopulation size in proportion to the fitness of each organism. In the present context, if we think of different organisms as different solutions, we see that niche formation advocates a PM-like allocation of trials among alternative organisms. The arguments we shall soon make concerning environmental uncertainty can be applied to the utility of niche formation without modification.
3. This argument can be made more rigorously (and more elegantly) by considering the minimization of *regret* (Simon, 1956). The argument of this section is fairly intuitive, however, and instead of pursuing the regret computation, we consider the more practical matter of whether probability matching behavior can be emulated by the various classifier bidding and payment mechanisms.
4. This apparent advantage of the noisy auction does not entirely rule out the roulette wheel, however. Wilson (1987) has observed a drive-out effect when roulette wheel bid selection is used in combination with the reproductive pressure of a genetic algorithm. In this way, the genetic algorithm cleans house, permitting more decisive, DT-like behavior than would be possible without the GA enabled. If it is desirable to separate credit assignment and discovery, there is some question whether such effects should be used. Additionally, it is unclear how any such population-based control mechanism can regulate the randomization of the bidding quickly enough to permit rapid shifting between problems.

References

Flood, M.M. (1954). Environmental non-stationarity in a sequential decision-making experiment. In R.M. Thrall, C.H. Coombs, & R.L. Davis (Eds.), *Decision processes*. New York: Wiley.

Goldberg, D.E. (1983). Computer-aided gas pipeline operation using genetic algorithms and rule learning. Doctoral dissertation, University of Michigan. *Dissertation Abstracts International, 44*, 3174B. (University Microfilms No. 8402282).

Goldberg D.E. (1989). *Genetic algorithms in search, optimization, and machine learning*. Reading, MA: Addison-Wesley.

Goldberg, D.E., & Richardson, J.J. (1987). Genetic algorithms with sharing for multimodal function optimization. *Genetic algorithms and their applications: Proceedings of the Second International Conference on Genetic Algorithms*. (pp. 41–49).

Goodnow, J.J. (1955). Determinants of choice-distribution in two-choice situations. *American Journal of Psychology, 68*, 106–116.

Holland, J.H. (1971). Processing and processors for schemata. In E.L. Jacks (Ed.), *Associative information processing*. New York: American Elsevier.

Holland, J.H. (1973). Genetic algorithms and the optimal allocation of trials. *SIAM Journal of Computing, 2*, 88–105.

Holland, J.H. (1975). *Adaptation in natural and artificial systems*. Ann Arbor, MI: University of Michigan Press.

Holland, J.H., & Reitman, J.S. (1978). Cognitive systems based on adaptive algorithms. In D.A. Waterman & F. Hayes-Roth (Eds.), *Pattern directed inference systems*. New York: Academic Press.

Lee, W. (1971). *Decision theory and human behavior*. New York: John Wiley & Sons.

McCracken, J., Osterhout, C., & Voss, J.F. (1962). Effects of instruction in probability learning. *Journal of Experimental Psychology, 64*, 267–271.

Mackintosh, N.J. (1974). *The psychology of animal learning*. London: Academic Press.

Siegel, S. (1959). Theoretical models of choice and strategy behavior: Stable state behavior in the two-choice uncertain outcome situation. *Psychometrika, 24*, 303–316.

Simon, H.A. (1956). A comparison of game theory and learning theory. *Psychometrika, 21*, 267–272.

Wilson, S.W. (1987). Classifier systems and the Animat problem. *Machine Learning, 2*, 199–228.