Journal of Experimental Psychology: General
1989, Vol. 118, No. 4, 417-421

0096-3445/89/$00.75

LMS Rules and the Inverse Base-Rate Effect: Comment on Gluck and Bower (1988)

Arthur B. Markman
University of Illinois

Gluck and Bower (1988) suggested that through the use of the Rescorla-Wagner learning rule, a connectionist network might be able to model the inverse base-rate phenomenon found by Medin and Edelson (1988). I prove that a network of the type that they proposed does not capture this effect. However, one can also prove that with additional assumptions about the encoding of features, the Rescorla-Wagner learning rule can be made to model the inverse base-rate effect. The importance of these assumptions and an outline of how they might be tested are then discussed.

Gluck and Bower (1988) recently showed how a version of the Rescorla-Wagner learning rule (a variant of the least mean squared [LMS] error correction rule) could be used to predict human behavior in a category-learning task. At the end of their article, they theorized that this rule might be able to account for subjects' incorrect use of base-rate information found in a set of experiments by Medin and Edelson (1988).

The inverse base-rate phenomenon (Medin & Edelson, 1988) is a surprising and counterintuitive effect shown by subjects in category-learning experiments in which the frequency of presentation of the categories is varied. Table 1 illustrates a simple case: Subjects are presented with exemplars for each of two categories. In the simple case, these categories have one common feature (A) and one distinctive feature (B or C). When one category is presented more frequently than the other, an interesting pattern or responses results. If subjects are given the common feature and asked which category is more likely to be described by this feature, subjects respond with the more frequently presented category. If either of the distinctive features is presented during testing, the appropriate category is selected. However, if both of the distinctive features are presented together, the subject is more likely to select the less frequently presented category, responding in opposition to the base-rate information.

Medin and Edelson (1988) discussed the possibility that the inverse base-rate effect stems from a competition between features. In a context in which the categories are diseases and the features are symptoms, they stated that

> Symptom A competes with symptom B to predict disease 1 and competes with symptom C to predict disease 2. Because symptom A is a better predictor of disease 1 than disease 2, it should compete more effectively with symptom B than symptom C. Because the total predictive strength is a constant, B will be weakened more by A than is C, and that therefore C will be stronger than B. Hence on the B,C test one would expect C to be more likely to control responding. (p. 74)

By this account, then, the changes in associative strength for each feature are not equal. Thus C has more associative strength with Category 2 than B does for Category 1.

LMS Rules and Base Rate Effects

Gluck and Bower (1988) state the belief that an LMS rule will capture the inverse base-rate effect found by Medin and Edelson (1988). They noted that

> In such circumstances, the LMS rule implies that Cue B will be relatively more blocked than Cue C in acquiring their respective associations, so that Cue C will dominate B in the BC conflict test. And this was the paradoxical reversal of base rate that was to be explained. (p. 242)

It was assumed that the differential associative strengths of the unambiguous features causes the inverse base-rate effect. If Gluck and Bower were correct, then some version of the Rescorla-Wagner learning rule should be able to capture the inverse base-rate effect. Furthermore, a connectionist system trained with the Rescorla-Wagner rule should exhibit a stronger connection between feature C and the low-frequency category than the connection between feature B and the high-frequency category. According to the Rescoral-Wagner rule,

$$\Delta W_{ij} = \beta(d_j - o_j) f_i, \quad (1)$$

where W_{ij} is the change in weight from input unit i to output unit j, β is the learning rate parameter, d_j is the training signal (expected output) for output unit j, o_j is the actual output of unit j, and f_i is the activation of input unit i.

If we apply this rule to a small connectionist system with three input units and two output units (see Figure 1), an interesting pattern of results emerges. There are two ways in which information can be coded in this system, and each method leads to a different outcome. The first is to code the presence of a feature with an activation of 1 at the corresponding input unit and the absence of a feature with 0 activation.

This work was supported in part by a University Predoctoral Fellowship from the University of Illinois in Urbana-Champaign given to the author and National Science Foundation Research Grant BNS-88-12193 given to Douglas Medin.

The author thanks Douglas Medin, Christopher Matheus, Edward Wisniewski, Dedre Gentner, Kenneth Forbus, Brian Ross, Kevin McReynolds, Mark Gluck, and Robert Goldstone for their helpful comments on earlier versions of this work.

Correspondence concerning this article should be addressed to Arthur B. Markman, Department of Psychology, University of Illinois, 603 East Daniel Street, Champaign, Illinois 61820.

Table 1
Category and Frequency Structure of a Simple Inverse Base Rate Effect Experiment

Feature	Response category	Frequency
Exemplar		
A, B	1	3
A, C	2	1
Test probe		
A	1	
B	1	
C	2	
A, B	1	
A, C	2	
B, C	2	
A, B, C	1	

The second method is to encode the presence of a feature with an activation level of 1 and the absence of a feature with an activation of −1. These two methods are treated separately because they make different assumptions concerning the processing of features. Gluck and Bower (1988) used only the first method in their model of category learning.

The Present 1, Absent 0 Case

Typically, the presence of a feature is encoded as a 1 and the absence of a feature is encoded as a 0. However, one can prove that with this coding, an LMS learning rule like the Rescorla-Wagner rule cannot capture the inverse base rate-phenomenon.

Initially, all weights in the system are set to 0. Thus the following equations hold:

$$a1 = b1 + c1; \tag{2}$$

$$a2 = b2 + c2. \tag{3}$$

During a learning trial, either Example 1 (A,B) or Example 2 (A,C) is presented to the system. One does this by setting the activation of each input node to 1 if its corresponding feature is present and to 0 if the feature is not present. This activation is then allowed to filter through the system.

One can find the output at any output node by multiplying the activation at each input node by the weight connecting

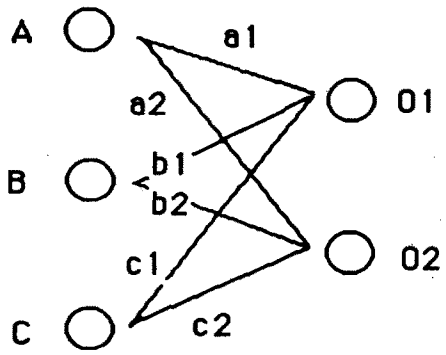

Figure 1. Simple network for modeling inverse base-rate experiments.

that node to a particular output node. Then, for each output node, the total input is added together in order to find the output of that unit; that is, for each output unit o_j,

$$o_j = \Sigma I_i w_{ij}, \tag{4}$$

where I_i is the activation of input unit i and w_{ij} is the strength of the connection between input unit i and output unit j.

Because the system is linear, no maximum or minimum activation level is set for any of the units, although a nonlinear activation function would not affect the results. The actual output obtained at the output nodes is then compared with the desired output, and the weights are changed according to the LMS learning rule described earlier.

I now examine the changes in weights that occur when Category 1 (A,B) is presented. Because the change in weights is dependent on the activations of the input nodes, the weights $c1$ and $c2$ will not change (C = 0). Furthermore, the weights $a1$ and $b1$ will change an equal amount because they are both equal to 1 and the other components of the LMS equation remain constant ($o1$, $d1$, and β). Similarly, $a2$ and $b2$ are altered by an equal amount. Thus Equations 2 and 3 still hold after a presentation of Category 1.

Similarly, presentation of Category 2 (A,C) will not change the value of $b1$ and $b2$ (B = 0). Furthermore, $a1$ and $c1$ will change equally, as will $a2$ and $c2$. Thus after the presentation of Category 2 in a learning trial, Equations 2 and 3 still hold.

In inverse base-rate experiments, presentation of A alone will lead to the selection of Category 1. Thus

$$a1 > a2. \tag{5a}$$

In order for the base rate phenomenon to be observed, presentation of B,C must lead to a response of Category 2; that is,

$$b1 + c1 < b2 + c2. \tag{5b}$$

However, substituting from Equations 2 and 3, we find that

$$b1 + c1 > b2 + c2. \tag{5c}$$

As a result, presenting B,C to the system results in selection of Category 1 as well, and the inverse base-rate effect is not observed.

This proof generalizes fairly well. Any change in the learning constant β will not affect the inequalities. Adding more concepts to the system will not affect this result because the connection among the nodes A, B, and C will not affect or be affected by other nodes in the system. Even if this task is implemented on a multilayer back propagation net (Rumelhart & McClelland, 1986), use of the standard LMS rule will not change this result.

One case that deserves more attention is the use of a distributed coding. For example, Gluck and Bower (1988) proposed a simple distributed coding in their Experiment 3. The presence and absence of a feature were noted explicitly with the use of separate nodes that may have had activation levels or 1 or 0. However, the proof given earlier still applies, and this distributed coding fails to capture the inverse base-rate effect. Under this representation, on any particular learning trial, if the connection weights between the node encoding

the presence of some feature and the categories change, the connection weights from the node encoding the absence of that feature will not change; that is, when a particular node (present/absent) is active, its opposite (absent/present) gets an activation of 0. Hence this coding will not predict inverse base-rate effects.

This proof shows that a representation scheme that entails activation values of 1 and 0 across the nodes of the system will not capture the inverse base-rate effect. The only case for which this proof does not generalize is the case in which the encoding of the presence and absence of features is not encoded as presence 1, absence 0.

The Present 1, Absent −1 Case

Changing the method of encoding stimuli for this task changes the results of the proof just given. This change in coding affects the underlying psychological interpretation of the model as well. First, a new proof is generated. Next, the assumptions made when this coding is used are discussed. Last, a discussion of how these assumptions may be empirically tested is outlined.

I return to the simple system of three input units and two output units described earlier (and pictured in Figure 1). The only change made to this system is that features present in the input are encoded as an activation level of 1 at the corresponding input node, whereas features not present in the input will be encoded as an activation level of −1.

Instead of using Equations 2 and 3 from above, I examine how the quantities $a1$ and $b1 + c1$ change in relation to each other. The same changes are examined for the pair $a2$ and $b2 + c2$.

When Category 1 (A,B) is presented to the system, input units A and B are set to 1 and unit C gets activation −1. According to Equation 1, this leads to the following changes in weights on an A,B trial:

$$a1 = \beta(1 - o1)(1), \tag{6a}$$

$$b1 = \beta(1 - o1)(1), \text{ and} \tag{6b}$$

$$c1 = \beta(1 - o1)(-1) \tag{6c}$$

and

$$a2 = \beta(-1 - o2)(1), \tag{7a}$$

$$b2 = \beta(-1 - o2)(1), \text{ and} \tag{7b}$$

$$c2 = \beta(-1 - o2)(-1). \tag{7c}$$

The quantities $b1 + c1$ and $b2 + c2$ do not change when Category 1 is presented. The changes in $b1$ and $c1$ are in equal but opposite directions. The same holds true for the weights $b2$ and $c2$. This same analysis can be carried out for the presentation of Category 2 (A,C), in which A and C are set to 1 and B is set to −1. Again, the quantities $b1 + c1$ and $b2 + c2$ undergo no change. This means that any transfer trial in which features B and C are both present or both absent, the selection of one category over the other will depend completely on other factors.

When the transfer tests for the inverse base-rate effect are applied to this system, an interesting result emerges. Examine the case in which presentation of feature A alone leads to a Category 1 response. When feature A is presented alone, input unit A is set to 1, and input units B and C are both set to −1. Because

$$b1 + c1 = 0 \tag{8a}$$

and

$$b2 + c2 = 0, \tag{8b}$$

it must be the case that

$$a1 > a2. \tag{8c}$$

Similarly, when features B and C are given to the system, input unit A is given a value of −1, and units B and C get the value of 1. Again, Equations 8a and 8b hold. In this case, however, input A has been set to −1. Therefore, the inequality in 8c is reversed. The output of unit 1 will be less than the output of unit 2. Thus Category 2 is chosen. The inverse base rate effect is indeed obtained.

Last, when A, B, and C are all presented simultaneously, input units A, B, and C are given values of 1. As in the previous two cases, Equations 8a and 8b hold because units B and C have equal values. As a result, Category 1, which is also selected when A is presented alone, will be selected in this case. These three results are all consistent with the results obtained by Medin and Edelson (1988).

However, unlike the proposals by Medin and Edelson (1988) and Gluck and Bower (1988), the LMS rule does not give differential strength to the unambiguous cues. Rather, the unambiguous cues are equal in strength and opposite in magnitude (see Equations 8a and 8b); that is, the presence of B increases the strength of Category 1 by the same amount that the presence of C decreases the strength of Category 1. Thus by this account, it is the greater association of the common cue with the high-frequency category that causes the inverse base-rate effect.

This proof generalizes as well. Changes in the encoding of the output, changes in the learning constant, and even distributing the encoding will not affect this result. However, when the system is made more complex by adding more input and output units, the results do change. Before one can examine how the system may be extended to include more features and classes, an examination of why the system works as it does is in order.

Encoding the Absence of Features

When the presence of a feature is encoded as 1 and the absence of a feature is encoded as −1, the system is actively encoding missing features as absent. Encoding an absent feature as an activation level of 0 is passive encoding. When the activation level of a missing feature is 0, no change is made to that weight during learning. When the absence of a feature in a Category 1 trial is encoded as −1, presence of that feature actively inhibits Category 1 in future trials.

Using this analysis, one may develop a three-tiered encoding scheme; that is, all nodes unrelated to the current learning trial can be given activation levels of 0. Thus only those features whose absence is noted explicitly, because of their association with the current input, will be given an activation of −1.

For an example in which this differential coding scheme becomes important, suppose that the system were extended to six input features and four classes (see Figure 2). There are now three new features (D,E,F) and two new categories (3,4). The study is now extended as well so that the experimental design described earlier is carried out twice for two sets of three input features and two classes. In this case, if all features not present were encoded as −1, the results obtained would be different from those obtained earlier. During learning, both A,B trials and A,C trials would include the information that D, E, and F were not present. This information would become more strongly associated with Category 1 than with Category 2 because there are more Category 1 trials than Category 2 trials. As a result, during A,B,C and B,C test trials, the encoding of D, E, and F as absent would cause a Category 1 response.

To regain the inverse base-rate effect, one may use a combination of the two encoding schemes. The missing feature on an A,B or an A,C trial (or a D,E or a D,F trial) would be encoded as −1, but the other missing features would be encoded as 0. Psychologically, this means that certain expectations are generated during learning trials. For example, on an A,C trial, seeing A sets up the expectation for B. When B is missing, it is actively encoded as not present. However, no such expectations are generated for D, E, or F with the presentation of A, B, or C, and so they are encoded as 0. This discussion is meant as an interpretation of the various activation levels that the input units can take. The specific implementation of an associative mechanism to notice correlations among subgroups of features in the input has not been developed.

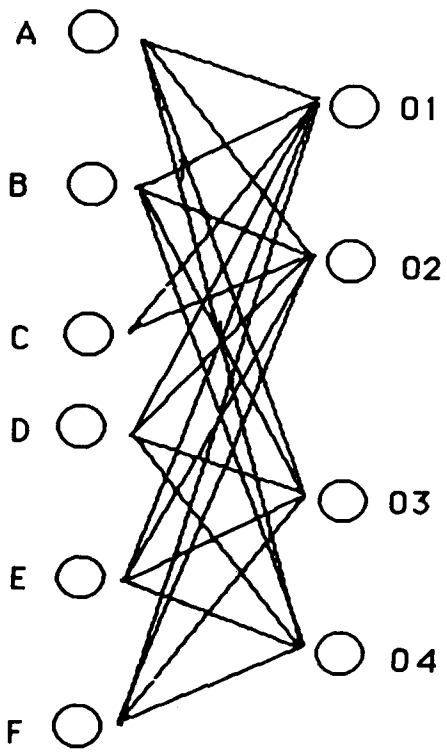

Figure 2. Extended version of the model.

Empirical Predictions

Interestingly, there are a number of empirically verifiable predictions made by this analysis. First, note that features B and C cancel each other out when they are both present or absent. Because the category selection is always dependent on feature A, feature A must be more strongly associated with Category 1 than with Category 2. As a result, the proportion of Category 1 responses on an A-alone trial or on an A,B,C trial should be greater than the proportion of Category 2 responses on a B,C trial. Furthermore, the proportion of Category 1 responses for A-alone and for A,B,C trials should be the same. Medin and Edelson's (1988) data seem to support this view.

There is one bit of evidence more directly in favor of the interpretation of the inverse base-rate phenomenon indicated by the Rescorla-Wagner rule. Medin and Robbins (1971) ran an experiment similar to those described earlier except that there was no common feature (A). The analysis just given would suggest that without a common feature, there would be no set of expectations, and no features would be encoded as missing. Thus no inverse base-rate effect should be found. In agreement with this analysis, Medin and Robbins found no inverse base-rate effect in this case.

Caveat

Before I conclude, one note of explanation is necessary. The use of an LMS rule to model the inverse base-rate effect is not necessarily being advocated here. Furthermore, there are a number of ways to implement the mechanism that generates expectations. Some of these schemes will probably fit the data better than others. However, as described earlier, the use of an LMS rule constrains the possible feature codings. Furthermore, these constraints make specific empirical predictions. It may turn out that these predictions are violated by empirical data. However, if these constraints are violated, then an LMS rule alone is probably not appropriate for modeling the inverse base-rate phenomenon.

Conclusions

As predicted by Gluck and Bower (1988), it is possible to use an LMS learning rule to model the inverse base-rate effect found by Medin and Edelson (1988). To model this phenomenon, however, one must assume that subjects actively encode certain features as missing. However, simply encoding the present features by using activation levels of 1 and 0, as Gluck and Bower did, will not work. An alternative representation, encoding activation levels of 1 and −1, will capture the inverse base-rate phenomenon. Under this coding, the inverse base-

rate effect stems from the fact that the contradicting features cancel each other out, whereas the absence of the common feature biases the subject's response to the low-frequency category. Because these coding assumptions can be empirically verified, it should be possible to determine the viability of LMS learning rules in simple category-learning experiments.

References

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117*, 227–247.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General, 117*, 68–85.

Medin, D. L., & Robbins, D. (1971). Effects of frequency on transfer performance after successive discrimination training. *Journal of Experimental Psychology, 87*, 434–436.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing.* Cambridge, MA: MIT Press.

Received October 25, 1988
Revision received February 23, 1989
Accepted February 28, 1989 ■