The Annals of Statistics
1996, Vol. 24, No. 6, 2319–2349

AN OVERTRAINING-RESISTANT STOCHASTIC MODELING METHOD FOR PATTERN RECOGNITION

BY E. M. KLEINBERG

State University of New York, Buffalo

We will introduce a generic approach for solving problems in pattern recognition based on the synthesis of accurate multiclass discriminators from large numbers of very inaccurate "weak" models through the use of discrete stochastic processes. Contrary to the standard expectation held for the many statistical and heuristic techniques normally associated with the field, a significant feature of this method of "stochastic modeling" is its resistance to so-called "overtraining." The drop in performance of any stochastic model in going from training to test data remains comparable to that of the component weak models from which it is synthesized; and since these component models are very simple, their performance drop is small, resulting in a stochastic model whose performance drop is also small despite its high level of accuracy.

1. Introduction. Traditionally, there are certain expectations in the area of data modeling concerning the interplay between size of training set, complexity of model, accuracy of model on training set and accuracy of model on test set. In somewhat simplistic form, conventional wisdom maintains that simpler models built from larger sets of training data, while usually less accurate on the training data, are better able to maintain their training data level of performance when subjected to new test data. This phenomenon is well known, appearing in things ranging from simple regression analysis [the linear function, while hitting none of the given (training) points, far better predicts the new points than some high-degree polynomial specifically designed to pass through the training points] to modern neural network analysis (where performance drop-off on test data due to complex, overtrained models is a major problem). (For example, see [7].)

One might be tempted to conclude that the field of pattern recognition is fundamentally limited by apparently conflicting procedural objectives: the desire to produce models which, on the one hand, become more and more accurate on training data, yet which, on the other hand, maintain such increased accuracy on test data.

As it turns out, however, this problem is not intrinsic to the field. In what follows, we will carry out a mathematical analysis of these issues and show how, under suitable theoretical conditions, the conflict does not exist.

This characterizes what is perhaps the most important feature of our approach to pattern recognition. The method we will introduce here is

Received October 1994; revised April 1996.
AMS 1991 *subject classifications.* 68T10, 68T05.
Key words and phrases. Pattern recognition, machine learning.

capable of producing, in general contexts, models which, for however long we decide to continue increasing their complexity (resulting in the expected improvement in performance on training data), *continue to improve their performance on test data*. This claim of resistance to overtraining is based on a mathematical analysis of "ideal" pattern recognition problems satisfying various natural solvability conditions (such as "representativeness" of the training set) to be described shortly. Of course, for "real-world" problems, the degree to which these conditions hold will vary, but the method appears to be robust in that small deviations from the ideal produce small degradations in performance. This has been confirmed by a number of results, both mathematical and anecdotal. Issues of "theory vs. practice" will be discussed as we proceed. And while there exist a few theorems on the subject, a full mathematical analysis of such sensitivity concerns has not yet been carried out. It appears to be a fruitful area for further research.

Our theoretical conclusions have been demonstrated experimentally in many different contexts. Perhaps a useful initial contrast may be drawn by considering a very simple, two-class experiment in handwritten digit recognition (later in this paper we will consider a more elaborate 10-class experiment). We trained both a stochastic model and a (feed-forward/back-propagation) neural network (see [9] and [11]) using identical data sets, and periodically tested each of them (as training was taking place) on a disjoint set of additional images. In keeping with virtually all other classical pattern recognition methods, the neural net showed a rise in performance on the training set as the decision surface was iteratively tailored to the specific characteristics of the training set, but after some point in time, performance on the test set began to decline.

We stopped the run after the first major drop in performance on the test set, and just prior to that, the neural network performed at a level of 93.33% correct on the training set and 87.58% correct on the test set (see Figure 1). By contrast (see Figure 2), the stochastic model performance continued to increase indefinitely on both training *and* test sets as model complexity increased, reaching the level of 99.77% correct on the training set and 99.24% correct on the test set at the point in time when we finally decided to terminate the training session.

In what follows we will initially introduce the ideas behind stochastic modeling in the context of a simple-to-visualize example in a two-dimensional feature space. We will then discuss a number of different experiments. In each of these, we use a standardized data set to contrast the performance characteristics of our method with those of standard pattern recognition techniques. Finally, we will develop the mathematical theory behind the technique of stochastic modeling including general results concerning the issues of computational feasibility and resistance to overtraining.

2. An example. In building discrimination models for pattern recognition problems, there are two fundamental goals one tries to achieve. The first, clearly, is to have the model perform well on the training data, for without

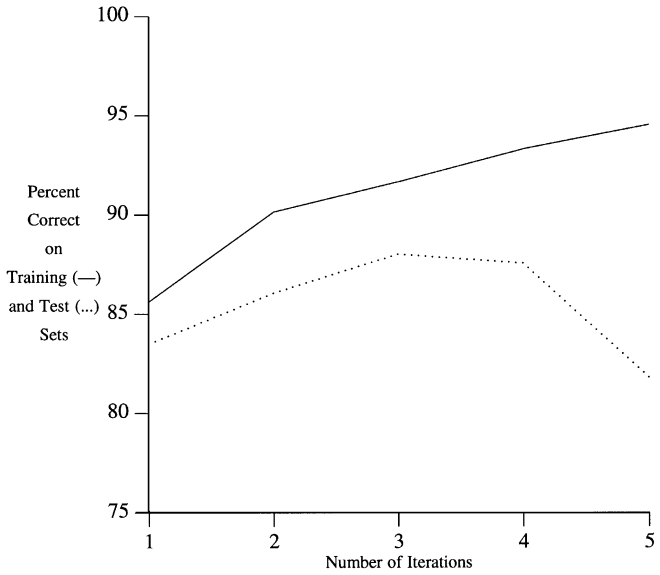

FIG. 1. *Feed-forward/back-propagation neural network.*

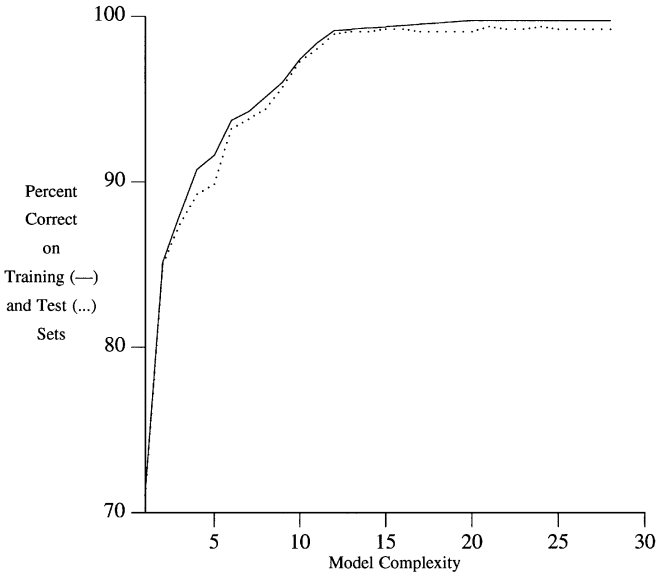

FIG. 2. *Stochastic modeling.*

this, there is no chance of decent performance in the real world. The second is to construct the model in such a way that its performance on training data "projects" to comparable performance in the real world. Of course, the "real world" in this context is usually some set of test data on which the completed model can be evaluated. One might simply define the "projectability" of a model to be 1 minus the difference in performance between training and test sets (so that the larger the projectability rating, the more projectable the model).

As we discussed above, these two goals are traditionally viewed as mutually conflicting. In common practice, it is usually the case that very simple models perform poorly on training data but have high projectability, while complex models perform well on training data but have low projectability.

The underlying idea behind our approach is this: given a problem, we will initially, in some routine, mechanical way, generate many very simple, so-called "weak models" for the problem which, although probably extremely inaccurate on the training set, will all be highly projectable. We will then take these weak models and combine them in a certain way to form a new, so-called "stochastic model." The key point is that the method of combining them will have the property that the projectability of the combined, stochastic model will be comparable to that of the component weak models from which it is built, yet, if enough weak models are used, the performance on the training set of the combined, stochastic model can be made arbitrarily good.

Before developing the formal theory associated with our approach, let us work through a simple example.

Let us assume that we are interested in a particular three-class pattern recognition problem, and that we are far enough along in the process so as to have settled on a mode of feature extraction whereby the objects among which we are trying to distinguish have been reduced to points sitting in a bounded region of some Euclidean space, the so-called feature space (see [5] and [15]). For example, the classes might be those associated with handwritten images of 1's, 2's and 3's, and the feature extraction process might extract 256 numeric features per image using a simple 16×16 gray-scale digitization. In this case, each such handwritten image would be reduced to a point sitting in Euclidean 256-space.

In order to be able to visualize things easily, let us assume that we are, in fact, only extracting two features per object so that the feature space is a subset of Euclidean 2-space, and that the regions of the feature space occupied by the (reductions of the) three classes are disjoint from one another. For example, things might appear as in Figure 3, where the regions in 2-space occupied by the three classes are indicated by different shades of gray.

As usual, when we are presented with this pattern recognition problem, we are not given the three classes in their totality, but are rather given finite, "representative," training subclasses which we will call TR_1, TR_2 and TR_3. We are also presented with finite, representative, test subclasses, TE_1, TE_2 and TE_3. In Figure 4, we show an example of what TR_1, TR_2 and TR_3 and TE_1, TE_2 and TE_3 might look like for the problem pictured in Figure 3. In

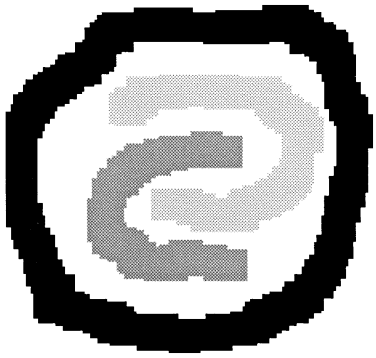

FIG. 3. *Three classes in two-dimensional feature space.*

this case, the training sets consist of 10%-dense pseudo-random samples of the underlying classes, and the test sets consist of 30%-dense pseudo-random samples. Corresponding training and test sets are disjoint from one another.

Our objective is to derive a discrimination model by "studying" the training data so that the (presumably) high degree of separation of TR_1, TR_2 and TR_3 achieved by the model projects to a comparable degree of separation of TE_1, TE_2 and TE_3.

Of course, it would be impossible to accomplish this without making some assumptions about the problem at hand. It is clear that in order for a given pattern recognition problem to have any hope of being solvable, there must exist models which, to at least some (however weak) degree, *discern* between points of different classes while not discerning between training and test points of the same class. In the context of our example above, for any pair of distinct indices i and j, there must exist a subset M of the feature space such that $P(M|TR_i)$ is (nontrivially) unequal to $P(M|TR_j)$ (i.e., M discerns, to

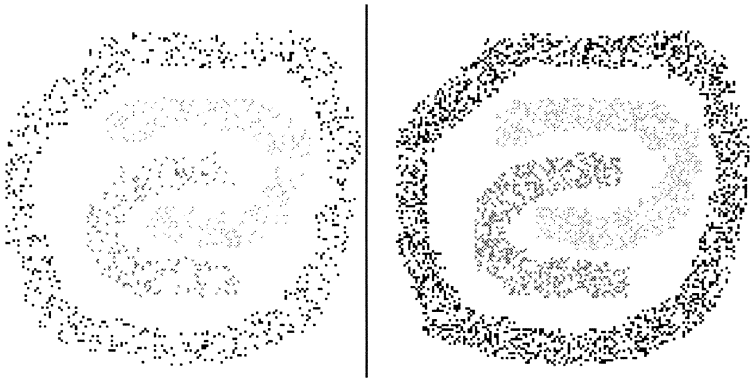

FIG. 4. *Training and test subsets.*

some extent, training set i from training set j), yet, for each index k, $P(M|TR_k)$ is (approximately) equal to $P(M|TE_k)$ (i.e., corresponding training and test sets are indiscernible by M).

Upon reflection, it turns out that our formalization of discernibility has a simple flaw, because it really only implies that TR_i and TR_j have some *local* discernibility from one another over that region of the feature space occupied by M; if this region is relatively small, then TR_i and TR_j might still be essentially indiscernible. Thus, in order to have the desired discernibility between TR_i and TR_j, we require the existence of a *collection* of sets M which is *spread over the feature space* such that for every M in the collection, $P(M|TR_i) \neq P(M|TR_j)$. We will leave the precise definition of the concept "spread over the feature space" for Section 4 of this paper, but informally, the sets in such a spread collection should treat training points of any given class equally so far as their degree of coverage by sets in the collection is concerned.

In practice, since representative training and test sets are usually assumed to be spatially distributed throughout the region of the feature space occupied by the underlying class (which is clearly the case for the example at hand), the requirement of *indiscernibility* between training and test sets with respect to a set M usually results, routinely, if we simply require that M be, in some sense, topologically "thick," that is, that it not consist of small regions of the feature space which might capture individual points from a training set without simultaneously capturing nearby points from the corresponding test set.

For the example under consideration, we used large rectangular regions, with area close to 50% that of the full feature space, as our thick subsets. In particular, for $i \neq j$, let $\mathbf{Q}_{i,j}$ denote the collection of those large rectangles, M, in the feature space such that $P(M|TR_i) \neq P(M|TR_j)$. Then, for each $i \neq j$, $\mathbf{Q}_{i,j}$ is (approximately) a spread collection of sets having the desired discernibility and indiscernibility properties. We will refer to regions in the union of the $\mathbf{Q}_{i,j}$ as *weak models* for the problem at hand.

Let us now use these weak models to actually construct a "stochastic discrimination model." We will proceed somewhat informally here—the reader who, at this time, desires more formal detail, or actual proof, should refer to Section 4.

Our model will be built by carrying out random sampling from the collections $\mathbf{Q}_{i,j}$. In fact, for any point q in our feature space and any $i \neq j$, $1 \leq i, j \leq 3$, we define a random variable $X^q_{(i,j)}$ on $\mathbf{Q}_{i,j}$ by

$$X^q_{(i,j)}(M) = 2\left(\frac{\chi_M(q) - P(M|TR_j)}{P(M|TR_i) - P(M|TR_j)}\right) - 1,$$

where χ_M denotes the characteristic function of M. Given the nature of $\mathbf{Q}_{i,j}$, it is not very difficult to prove that the expectation of $X^q_{(i,j)}$ is 1 if q is a member of TR_i, is -1 if q is a member of TR_j and is something close to 0 if q is in TR_k, $k \neq i$, $k \neq j$.

If we now define, for any point q in our feature space and any i, $1 \leq i \leq 3$, the random variable k_i^q to be $\frac{1}{2}\Sigma_{j \neq i} X_{(i,j)}^q$, then the expectation of k_i^q will be 1 if q is a member of TR_i, and less than 0 otherwise.

Consider now the following "first pass" at a stochastic discrimination model for our problem: given a point q in the feature space, in order to classify q simply evaluate each of k_1^q, k_2^q and k_3^q at a randomly chosen point in each of their respective sample spaces; now classify q as being of class i where it is the value of k_i^q which is the largest of the three values so calculated. Our rationale here is simply that if q were indeed of class i, then the expectation of k_i^q is 1, yet the expectation of k_j^q for j different from i is less than 0.

If instead of sampling only once, we do it many times (and take the average of the values computed), then, by the law of large numbers, the variance (of the "average" variable) goes to 0, and so the probability of being close to the expectation goes to 1. Thus, under our assumption that the classes $\mathbf{Q}_{i,j}$ are spread and have appropriate discernibility properties, the accuracy of such discrimination models on the training set goes to 100%.

Furthermore, because of our indiscernibility assumption concerning the $\mathbf{Q}_{i,j}$, the expectations (and the variances) of the k_j^q are approximately the same whether q is a training point or a test point. As a result, our discussion concerning the accuracy of our stochastic discrimination model on the training set applies to the test set as well. And so, as the size of our random sample from the $\mathbf{Q}_{i,j}$ increases, the accuracy of the stochastic model on *both* the training and the test sets goes to 100%.

Let us make a remark concerning the rate of convergence. Given the definition of $X_{(i,j)}^q$, it is clear that its variance can be driven up by small values of $|P(M|TR_i) - P(M|TR_j)|$. If we define the (i, j)-*enrichment degree* of a collection Q of weak models to be

$$\inf\left\{\left|P(M|TR_i) - P(M|TR_j)\right| \middle| M \in Q\right\},$$

then one can derive an estimate of the number of weak models needed in order that the stochastic model built from them achieves a given level of performance as a function of (a) that level of performance and (b) the enrichment degree of the set from which the weak models were selected. A simple analysis of this will be carried out in Section 4, but we might note here that for a two-class, i vs. j, pattern recognition problem, the size of the weak model sample needed so that the expected accuracy of the associated stochastic model is greater than $1 - 1/u$ is directly proportional to u and inversely proportional to the square of the (i, j)-enrichment degree.

Given this fact, one might be tempted to work exclusively with highly enriched collections of weak models. However, it is clearly more difficult to find such "stronger" weak models, so there is an immediate time trade-off involved. In addition, if weak models are allowed to become too strong, they may start to exhibit signs of overtraining leading to unacceptable degradation of our indiscernibility conditions.

Before considering the application of our method to the synthetic problem illustrated above, let us make several general comments in order to provide some historical perspective on this approach. The underlying ideas behind stochastic discrimination, the basis for our work here, were first introduced in [16]. Immediately following this initial work, we began experimenting with implementations of the method, and, simultaneously, began formulating the mathematical concepts needed to analyze this practical application of stochastic discrimination to statistical pattern recognition problems. Early drafts of the current paper, including the mathematical conditions needed to guarantee perfect solvability of problems derived from representative training sets, have been in circulation since 1991. Based on these drafts, others began research on our methods and on variations of our implementation. For example, one such variant, based on ideas by Ho for changing the random variable $X_{(i,j)}$ ([13]), was developed and studied extensively by Berlind [2]. And in [14], Ho studied the use of leaves from fully split decision trees, where each leaf was perfectly enriched for one class, and each point was covered by exactly one leaf of each tree. An objective here was the desire for good "spread" in the collection of weak models.

The general notion of improving classifier accuracy by combining a number of less accurate classifiers trained for the same task has been around for quite some time (see [10] and [12]). Even in the context of PAC learning (see, e.g., [6], [8] and [20]), there has been recent work concerned with "boosting" accuracy by using, in concert, numbers of different "weak" hypotheses. The weak hypotheses here are generated *sequentially* by training on different sets of examples, where the derivation of each such set of examples at a given stage is based on what took place at previous stages. The (relatively small) number of weak hypotheses generated in this way are then combined using majority vote. In practice, it has helped to improve performance of various classification approaches, as one would expect from any of the combination methods.

Our approach is fundamentally different in that it is based on random sampling from *the space of all possible classifiers*, that is, sampling from the power set of the feature space. And while accuracy increases with the use of larger and larger samples of such subsets, this is in no way a serial learning technique because the sampling can be carried out entirely in parallel. Thus, on a suitable machine, one of our stochastic models can be built basically in one step. In practice, we restrict our attention to subspaces satisfying certain required conditions (based on the notions of spread, indiscernibility and enrichment), and even here, the sampling/combining process maintains projectability independent of the number of weak models (samples) chosen, and hence, of the resulting complexity of the classifier. In effect, our notion of stochastic modeling is a technique, based on the laws of large numbers and the central limit theorem, to amplify small differences through the use of discrete stochastic processes.

As mentioned above, stochastic models are computationally feasible both to produce and to use. They are especially suitable to parallel implementation,

for both the generation and evaluation of component weak models are things best carried out in parallel. Indeed, the weak models can all be chosen (if building the stochastic model) or evaluated at a point in the feature space (if evaluating the stochastic model at a point in the feature space) *simultaneously*. Some parallel implementations have, in fact, been tested experimentally. This work will be reported on elsewhere.

Let us now take a look at our attempt to build a stochastic model for the problem illustrated in Figures 3 and 4 above. Just as described, we carried out random sampling from the collection of (large rectangular) weak models. During the course of our sampling, we periodically evaluated stochastic models based on the samples of weak models we had to that point in time. For example, with a total sample size of only two weak models, the associated stochastic model performed at a level of 44.63% on the training set and 43.94% on the test set. In Figure 5, we tabulate stochastic model performance as a function of sample size for a range of values during our run.

A more complete selection of stochastic model performances as a function of size of weak model sample for our run is presented graphically in Figure 6.

Note that the performance on training and test sets is not identical. This is because the indiscernibility of training and test sets by regions in the $\mathbf{Q}_{i,j}$ is just approximate, and so the similarity of the probability density functions of the random variables k_j^q for points q in TR_k and points q in TE_k, a similarity which serves as the basis for our results concerning the projectability of stochastic models, is also only approximate. In effect, the weak models themselves, right from the start, suffer from some (small) degree of overtraining. However, it is extremely interesting to see in Figure 6 that this small divergence between training and test set performance underlying the space of weak models does not appear to increase as more and more weak models are sampled for inclusion in the rather complex, highly accurate, stochastic

Size of Weak Model Sample	Training Set Performance	Test Set Performance
2	44.63%	43.94%
4	55.99%	55.77%
6	69.69%	69.23%
10	75.29%	73.69%
13	83.22%	82.18%
16	86.70%	85.96%
28	91.10%	90.07%
208	95.21%	94.25%
516	97.72%	96.95%

FIG. 5. *Stochastic model performances as function of size of weak model sample.*

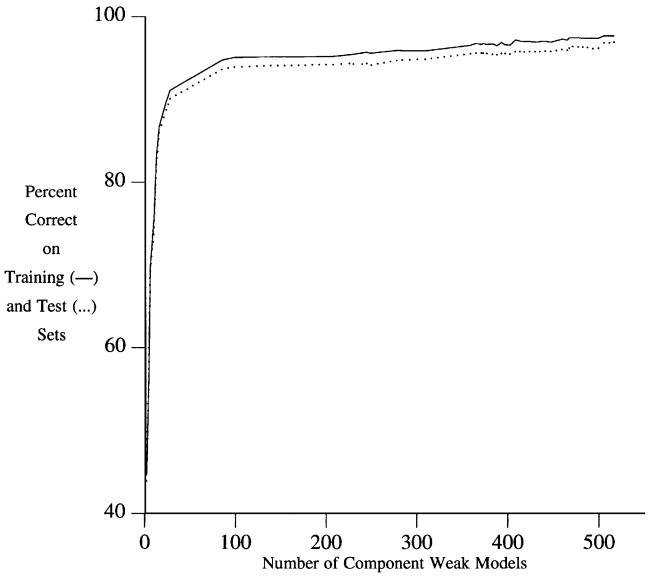

FIG. 6. *Stochastic model performances as function of size of weak model sample.*

model. "Weaker" weak models might well suffer less from overtraining and lead to stochastic models with an even smaller divergence between training and test set performance. For this particular problem, we have, in fact, built a stochastic model which performed at a level of 99.60% correct on the training set and 99.26% correct on the test set.

There is one additional intriguing point we might make here relating theory and practice. Given some degree of overtraining on the part of the weak models considered, there will be some lag of test set performance behind training set performance as the stochastic model is being built. However, given that the increasing accuracy of a stochastic model on either training or test set is based on certain variances approaching 0, as a practical matter, again supported by the theory, so long as the indiscernibility of training and test set is close enough to the theoretical requirement so that the expectations of the k_j^q for test set q of different classes are sufficiently separate, it could well happen that at some point in building the stochastic model, while training set performance maxes out at some level, test set performance could still continue to increase as more weak models are added. In effect, the longer the training goes on, the less overtrained the stochastic model becomes.

Finally, in Figure 7, we present a sequence of pictures (to be read from left to right starting with row 1) which show the partitions of the feature space created by each of the nine stochastic models reported on in Figure 5. Here we get some sense of the nature of stochastic models as discrimination tools based on decision regions rather than (conventional) decision surfaces. Note the increasing resolution over the sequence of the nine models.

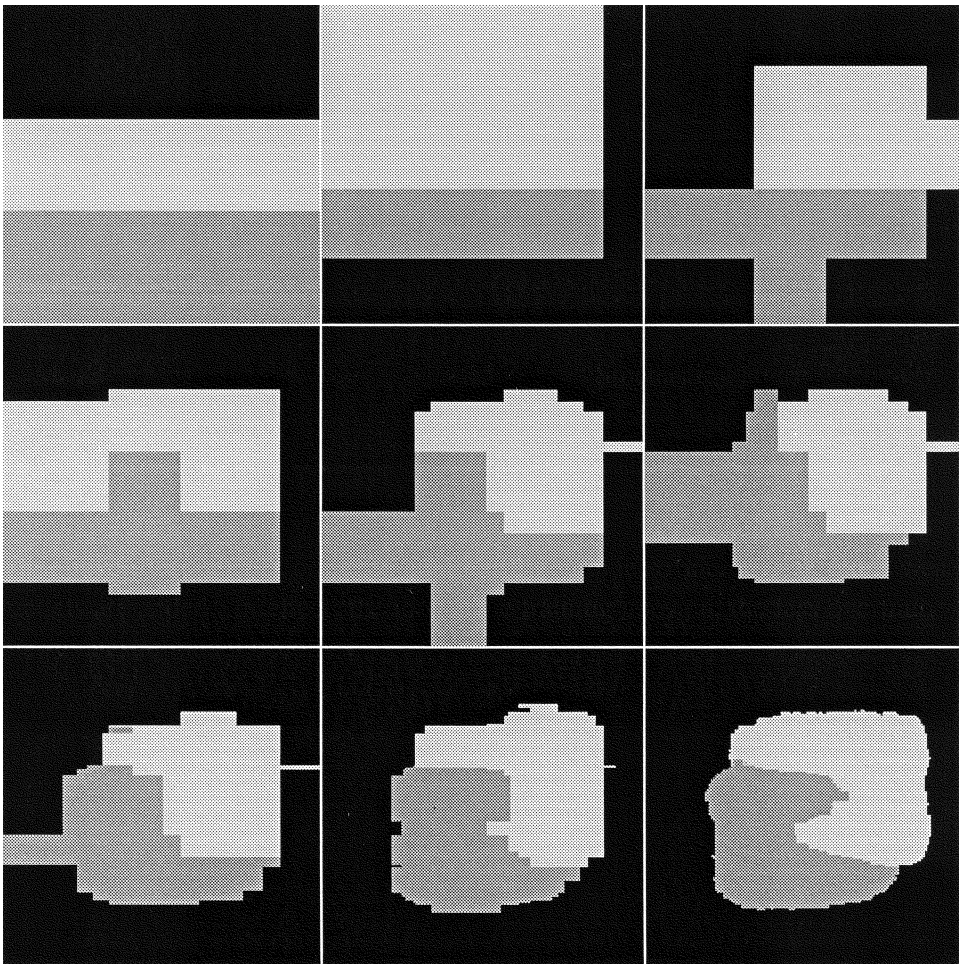

FIG. 7. *Stochastic partitions of feature space.*

3. Some experimental results. Let us now consider experimental results in handwritten digit recognition. The algorithmic implementation of our method which was used here was identical to that developed above for use in the two-dimensional feature space example. The only adjustment needed was a simple modification dictated by the fact that our underlying feature space was no longer two dimensional.

Our study was carried out on handwritten digits comprising a standard database supplied by the National Institute of Standards and Technology (N.I.S.T.) containing writing samples from thousands of different people. We selected sample images from the set by taking, for the first 1000 different people represented, the first example of each digit written by each of them. In this way, we were able to put together a set of almost 10,000 digits containing approximately 1000 examples of each digit.

The handwritten digits contained in this database were originally scanned and binarized at 300 dpi, but we preprocessed them prior to sending them to the stochastic modeling routine. The preprocessing reduced each scanned image to a fixed-length numeric record by first size-normalizing the image to 16×16, and then, within each of the 256 windows in the image associated with this size normalization, calculating the number of black pixels present as a fraction of the total number of pixels in the window. We then thresholded the fraction against 0.2 to decide whether to call the window "on" or "off." In other words, if the fraction of black pixels in the window was greater than 0.2, we assigned the field value 1 to that window for the record associated with the given image—otherwise we assigned the field value 0. In this way, each original scan was reduced to a fixed-length numeric record of 256 integer fields, each of whose values was either 0 or 1.

Once the image database was reduced in this way, we selected the first 4997 records as our training set and reserved the remaining 4975 records as our test set. In this way, we had about 500 examples of each digit type in each of the training and test sets. Given the way in which we selected our images at the start, this guaranteed that not only were the training and test sets disjoint from one another, but also that no person with a handwritten image appearing in the training set contributed any images to the test set.

We then fed the training set of records into our stochastic model building program and began the (weak model) sampling process. One should keep in mind that since our feature space was now a subset of Euclidean 256-space, our weak models were now constructed from large, 256-dimensional rectangular parallelopipeds, that is, large regions in the feature space $\{0,1\}^{256}$ which can be written in the form

$$\{(x_1, x_2, \ldots, x_{256}) | \text{for each } i,\ a_i \le x_i \le b_i\},$$

where $(a_1, a_2, \ldots, a_{256})$ and $(b_1, b_2, \ldots, b_{256})$ are members of $\{0,1\}^{256}$. Given the small size of the training set, we actually set things up so that the only such regions considered as candidates were those which contained at least one member of the training set.

Every so often during the run, as the sample of weak models was accumulating, we evaluated the performance of the stochastic model based on the sample as it existed at that point in time. For example, we first evaluated performance when our random sample consisted of 22 weak models. At that point in time, the stochastic model performed at a 63.04% level of accuracy on the training set and a 62.09% level of accuracy on the test set. The next evaluation took place when the sample size was 64, and here the stochastic model performed at a 76.43% level of accuracy on the training set and a 74.75% level of accuracy on the test set.

During the course of the run, many such evaluations took place. In Figure 8, we tabulate a short selection of them, and in Figure 9, we graph both training set performance and test set performance as functions of sample size.

Size of Weak Model Sample	Training Set Performance	Test Set Performance
22	63.04%	62.09%
64	76.43%	74.75%
235	89.25%	87.80%
618	92.60%	91.06%
1953	95.94%	94.03%
4536	97.32%	95.42%
7634	98.02%	95.78%

FIG. 8. *Stochastic model performances as function of size of weak model sample.*

There are several things to note here. As predicted by the theoretical discussion in Section 2, as the number of weak models from which the stochastic model is built increases, the accuracy of the stochastic model increases. (If we continued adding weak models, the performance would continue to increase.)

Note that there is some degradation of performance in going from training to test sets; that is, there is some degradation in projectability. This is due simply to the fact that the component weak models from which the stochastic model is built are themselves not perfectly projectable, a situation caused, in

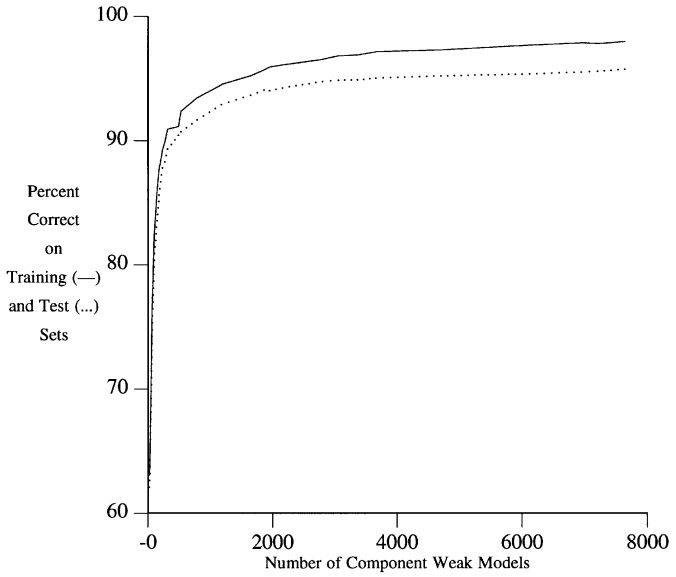

FIG. 9. *Stochastic model performances as function of size of weak model sample.*

part, by the fact that our training set is relatively small, and hence not nearly as representative as it might be. In any case, the performances on training and test sets increase together.

Due to the size of our training set, the digit recognizer produced in this particular experiment had little chance of being state of the art. That was not our intention here. We wanted a simple, yet nontrivial, context in which to *compare* different pattern recognition techniques, where all techniques had access to identical training and test sets. If one were interested in creating as accurate a digit recognizer as possible, the training set should probably contain at least 100,000 examples, rather than the 5000 we used here. See, for example, [1]. To add some perspective to this point, consider the result after we built a stochastic digit classifier as above, but this time using training (and test) sets of size only 250. In this case, the difference between training set performance and test set performance was 14.8%. Doing the same thing for training and test sets of size 500, the difference was 7.8%. And using the 4997 training and 4975 test examples as above, the difference was 2.24%. [In addition, one could expect far better performance if gray-scale features were used (rather than the thresholded binary features used here). For example, using the identical digit set to that used above, but this time training on 256 gray-scale features, after 7662 weak models, the stochastic classifier achieved a rate of 98.40% correct on the training set and 96.58% correct on the test set, with a difference between the two of 1.82%.]

Again, we were not interested here in trying to produce the best possible handwritten digit recognizer based on stochastic discrimination. We simply want to compare our results to those achieved by other techniques under identical conditions.

Toward this end, we evaluated both a nearest-neighbor algorithm and a k-nearest-neighbor (with $k = 3$) algorithm on the same test set of 4975 records using the Euclidean metric along with the 4997 training records as templates (see [4] and [17]). On the test set, the nearest-neighbor algorithm performed at a 95.22% level of accuracy, and the k-nearest-neighbor algorithm performed at a 95.14% level of accuracy. We also trained a (feed-forward/back-propagation) neural network on these same training records, and its level of accuracy when evaluated on the test set of 4975 records was 94.67%.

The performance of the stochastic model will continue to increase as we add weak models, but with as few as 7634 weak models, it was performing at a level of 95.78%, more than 0.5% better than the nearest-neighbor methods. On the other hand, the 7634-component stochastic model evaluated records many times faster than either of the nearest-neighbor methods. The basis for this lies in the fact that weak models are extremely simple to evaluate. For this particular application, evaluating any weak model on a record required at most 18 Boolean computations.

For a more complete discussion of similar applications of our stochastic modeling algorithm to handwritten digit recognition, we refer the reader to [17].

There are many other experiments which have been carried out in order to further evaluate, anecdotally, the degree to which theory and practice differ when our mathematical assumptions are not strictly met. Let us briefly mention two of them. In [18], we considered a synthetic problem involving two *overlapping* classes in Euclidean 2-space. The classes were produced by simply taking Gaussian joint distributions centered at $(80, 80)$ (for class 0) and $(120, 120)$ (for class 1). In each case, x and y varied independently, with a standard deviation of 20. For this problem, the Bayes error rate (estimated by numerical integration) was 0.07866, implying a maximum correct rate of 92.13%. Pseudo-random training and test sets, each of size 2000, were constructed, and we proceeded to build a stochastic model by sampling the same space of weak models, based on large rectangular regions, used above. Despite the overlap of the classes, as the number of contributed weak models increased, performance of the stochastic model on both training and test sets increased. The Bayes (maximum) correct rate was reached after fewer than 150 weak models were chosen.

In another experiment based on data provided by Project Statlog ([19]), the object was to classify boundary types in DNA sequences. Feature vectors were derived from contiguous blocks within sequences, and each consisted of 180 binary fields. This was a problem formally defined in terms of specific training and test sets (of sizes 2000 and 1186, respectively), and over time, many of the different pattern recognition techniques were tried out on it in order to evaluate their relative levels of performance. The resulting accuracies of more than 20 different methods are reported in [19]. These extremely interesting results show, for example, test set performances of KNN at 85.4%, BackProp at 91.2% and Cart at 91.5%. The best result (Polar) reported 95.9%. We built a stochastic model for this problem involving 6451-many weak models sampled from exactly the same sort of space, based on large rectangular parallelopipeds, used above, and its test set performance was 96.2%.

It is important to note that for this (and every other experimental result reported here), the same stochastic model building program was used. The only user-supplied information which varied from run to run was the data in the training and test sets.

See [18] for algorithmic detail on a practical implementation of stochastic discrimination, and for the application of that implementation to these and other problems.

4. Mathematical considerations. In this section we will develop the mathematics underlying the ideas introduced in Section 2. Our first goal is to formalize the notion of a collection of subsets of the feature space (usually, a collection of weak models) being "spread." This was a key concept involved with our initial discussion of indiscernibility/discernibility as applied to the various training and test sets, and it was never really pinned down at that time.

As before, we assume that our underlying feature space, F, is finite. We will use the counting measure μ on F and the counting measure ν on the power set of F, $\mathbf{F}$, to calculate probabilities whenever necessary.

NOTATION. We denote by r the function from $\mathbf{F} \times \mathbf{F}$ into the reals given by

$$\text{for each pair of subsets } (S,T) \text{ of } F, T \neq \emptyset, \; r(S,T) = \frac{\mu(S \cap T)}{\mu(T)}.$$

Viewing F as a sample space, $r(S,T)$ is just $P(S|T)$, that is, the probability of S given T. This function r will be an extremely useful notational device for us.

NOTATION. Given a real number x, a member A of $\mathbf{F}$ and a subset $\mathbf{M}$ of $\mathbf{F}$, let $\mathbf{M}_{x,A}$ denote the set of those M in $\mathbf{M}$ such that $r(M, A) = x$.

Intuitively, for any two subsets S and T of F, $r(S,T)$ equals the fractional amount of T which is captured by S. Thus, we might make the following trivial observation: given any real number x, member A of $\mathbf{F}$, subset $\mathbf{M}$ of $\mathbf{F}$ and members M and N of $\mathbf{M}_{x,A}$, the probability that M captures a point in A is equal to the probability that N captures a point in A, that is,

$$P_F(\, p \in M | p \in A) = P_F(\, p \in N | p \in A).$$

(Since we will be dealing with several different probability spaces in what follows, there might be times when confusion could arise as to just which space we are taking probabilities with respect to. At times of such potential ambiguity, we will use P_T to denote probabilities taken with respect to the space T.)

The notion of "spread" derives from what might be viewed as the "dual" of this trivial observation.

DEFINITION. Let $\mathbf{M}$ be a collection of subsets of a given feature space F, and let A be a subset of F. Then $\mathbf{M}$ is said to be *A-uniform* if, for any real number x such that $\mathbf{M}_{x,A}$ is nonempty, given any two points p and q in A, the probability *relative to the space* $\mathbf{F}$ that p is captured by a member of $\mathbf{M}_{x,A}$ is equal to the probability that q is captured by a member of $\mathbf{M}_{x,A}$, that is,

$$P_{\mathbf{F}}(\, p \in M | M \in \mathbf{M}_{x,A}) = P_{\mathbf{F}}(q \in M | M \in \mathbf{M}_{x,A}).$$

An A-uniform collection of subsets of F is, in some intuitive sense, "spread over A."

While it is easy to construct examples of collections of subsets of F which are not A-uniform, it is also the case that examples of such uniformity occur quite naturally. In fact, given any subset A of F, a simple counting argument shows that, for any real number x such that $\mathbf{F}_{x,A}$ is nonempty, the probability relative to the space $\mathbf{F}$ that a member of A is captured by a member of $\mathbf{F}_{x,A}$ is x. Thus, the full power set of F is always A-uniform for every subset A of F.

It turns out that $P_{\mathbf{F}}(q \in M | M \in \mathbf{M}_{x,A}) = x$ holds for *any* collection of subsets $\mathbf{M}$ which is A-uniform.

LEMMA 1. *Let F be a given feature space, and let A be a subset of F. Suppose $\mathbf{M}$ is an A-uniform collection of subsets of F. Then, for any real number x such that $\mathbf{M}_{x,A}$ is nonempty,*

$$P_{\mathbf{F}}(q \in M | M \in \mathbf{M}_{x,A}) = x$$

for every q in A.

PROOF. Since $\mathbf{M}$ is A-uniform, we know that, for some real y,

$$P_{\mathbf{F}}(q \in M | M \in \mathbf{M}_{x,A}) = y$$

for every q in A. Let a denote the (common) number of elements in $M \cap A$ for (any) M in $\mathbf{M}_{x,A}$, let b denote the number of elements in A, let c denote the (common) number of sets M in $\mathbf{M}_{x,A}$ such that (any given) q (in A) is a member of M, and let d denote the number of elements of $\mathbf{M}_{x,A}$. We can count the number of ordered pairs (q, M) such that $q \in M \cap A$ and $M \in \mathbf{M}_{x,A}$ in either of two ways, getting bc one way and ad the other. Thus, $a/b = c/d$, and since we chose a, b, c and d such that $a/b = x$ and $c/d = y$, we have shown that $x = y$. □

In light of this result, we are now in a position to formally define the notion of "spread" needed for our work. In considering the following notation and definition, it would be helpful for the reader to picture an m-class pattern recognition problem in the feature space F, where, for each i, C_i is the training set associated with class i.

NOTATION. Given a positive integer m, a sequence $\mathbf{C} = \langle C_1, C_2, \ldots, C_m \rangle$ of subsets of F and a sequence $\mathbf{x} = \langle x_1, x_2, \ldots, x_m \rangle$ of reals, $\mathbf{M}_{\mathbf{x},\mathbf{C}}$ denotes the set of those M in $\mathbf{M}$ such that, for each j, $1 \le j \le m$, $r(M, C_j) = x_j$.

DEFINITION. For a given sequence of subsets $\mathbf{C} = \langle C_1, C_2, \ldots, C_m \rangle$ of F, a subset $\mathbf{M}$ of $\mathbf{F}$ is said to be *$\mathbf{C}$-uniform* if, for every j, $1 \le j \le m$, every member q of C_j and every sequence $\mathbf{x} = \langle x_1, x_2, \ldots, x_m \rangle$ of real numbers such that $\mathbf{M}_{\mathbf{x},\mathbf{C}}$ is nonempty,

$$P_{\mathbf{F}}(q \in M | M \in \mathbf{M}_{\mathbf{x},\mathbf{C}}) = x_j.$$

Given an m-class pattern recognition problem in F with training sets $\mathbf{TR} = (TR_1, TR_2, \ldots, TR_m)$, any *$\mathbf{TR}$-uniform* collection of subsets of F is "sufficiently spread over the training sets" for our development to succeed.

Let us make a comment here concerning the formal definition of uniformity given above. Throughout this paper we will give a number of definitions involving strict equalities such as, for example, the equality

$$P_{\mathbf{F}}(q \in M | M \in \mathbf{M}_{\mathbf{x},\mathbf{C}}) = x_j$$

given above. In actual practice, however, such equalities might only be satisfied to within some small real value ε. Depending on just how far off we are in any given instance, there may be some commensurate drop in various other factors predicted by our theory. However, small errors in satisfying equalities in definitions result in small errors in predicted outcome. In this section our goal is to present the underlying theory of stochastic modeling as simply as possible, without the additional complication of estimating derivative error. For a rigorous treatment of these issues, we refer the reader to [2] and [3].

We now attack the issue of indiscernibility between corresponding training and test sets. As already discussed in Section 2, this concept is dependent not only on the training and test sets themselves, but also on a class of subsets of the feature space (such as a class of thick regions). As discussed previously, in order for two given subsets A and B of F to be indiscernible with respect to a collection $\boldsymbol{M}$ of subsets of F, we would certainly require that $r(M, A)$ be (approximately) equal to $r(M, B)$ for every member M of $\boldsymbol{M}$. However, in light of our recently completed discussion concerning uniformity, it seems reasonable that if A were truly indiscernible from B with respect to sets in the collection $\boldsymbol{M}$, then if $\boldsymbol{M}$ were "spread" over A, $\boldsymbol{M}$ would also have to be "spread" over B. On the other hand, the notion of indiscernibility exists independently of any uniformity which may or may not hold for the sets involved, so our definition must consider concepts related to the degree of spread of sets in $\boldsymbol{M}$ over the regions A and B. Formalizing this to our current context of training and test subsets of an m-class pattern recognition problem, we introduce the following definitions which, as before, will be easier to consider if the reader pictures an m-class pattern recognition problem in the feature space F, where, for each i, C_i is the training set associated with class i and D_i is the test set associated with class i.

DEFINITION. Given any subset $\boldsymbol{M}$ of $\boldsymbol{F}$, any positive integer m, any sequence $\boldsymbol{C} = \langle C_1, C_2, \ldots, C_m \rangle$ of subsets of F and any sequence $\boldsymbol{x} = \langle x_1, x_2, \ldots, x_m \rangle$ of reals such that $\boldsymbol{M}_{\boldsymbol{x},\boldsymbol{C}}$ is nonempty, for any j, $1 \leq j \leq m$, $f^j_{\boldsymbol{M},\boldsymbol{x},\boldsymbol{C}}$ is the random variable defined on C_j whose value at any q is given by

$$f^j_{\boldsymbol{M},\boldsymbol{x},\boldsymbol{C}}(q) = P_{\boldsymbol{F}}(q \in M | M \in \boldsymbol{M}_{\boldsymbol{x},\boldsymbol{C}}).$$

In some sense, the random variables $f^j_{\boldsymbol{M},\boldsymbol{x},\boldsymbol{C}}$ provide "profiles of coverage" of the sets in $\boldsymbol{C}$ by members of $\boldsymbol{M}$. We are thus led naturally to the following definition.

DEFINITION. Given any subset $\boldsymbol{M}$ of $\boldsymbol{F}$, any positive integer m and any two sequences $\boldsymbol{C} = \langle C_1, C_2, \ldots, C_m \rangle$ and $\boldsymbol{D} = \langle D_1, D_2, \ldots, D_m \rangle$ of subsets of F, we say that $\boldsymbol{C}$ is $\boldsymbol{M}$-*indiscernible* from $\boldsymbol{D}$ if:

(a) for any j, $1 \leq j \leq m$, and any M in $\boldsymbol{M}$, $r(M, C_j) = r(M, D_j)$;

(b) for any sequence $\boldsymbol{x} = \langle x_1, x_2, \ldots, x_m \rangle$ of reals and for any j, $1 \leq j \leq m$, the random variables $f^j_{\boldsymbol{M},\boldsymbol{x},\boldsymbol{C}}$ and $f^j_{\boldsymbol{M},\boldsymbol{x},\boldsymbol{D}}$ have the same probability density functions.

It is immediate that the notion of $\boldsymbol{M}$-indiscernibility is an equivalence relation. Furthermore, $\boldsymbol{M}$-indiscernibility preserves uniformity, as shown by the following result.

LEMMA 2. *Given any two sequences* $\boldsymbol{C} = \langle C_1, C_2, \ldots, C_m \rangle$ *and* $\boldsymbol{D} = \langle D_1, D_2, \ldots, D_m \rangle$ *of subsets of* F, *if* $\boldsymbol{C}$ *is* $\boldsymbol{M}$*-indiscernible from* $\boldsymbol{D}$ *and* $\boldsymbol{M}$ *is* $\boldsymbol{C}$*-uniform, then* $\boldsymbol{M}$ *is* $\boldsymbol{D}$*-uniform.*

PROOF. We simply note that for any sequence of subsets $\boldsymbol{B} = \langle B_1, B_2, \ldots, B_m \rangle$ of F, $\boldsymbol{M}$ is $\boldsymbol{B}$-uniform iff for every sequence $\boldsymbol{x} = \langle x_1, x_2, \ldots, x_m \rangle$ of real numbers such that $\boldsymbol{M}_{\boldsymbol{x},\boldsymbol{B}}$ is nonempty, for every j, $1 \leq j \leq m$, the pdf of the random variable $f^j_{\boldsymbol{M},\boldsymbol{x},\boldsymbol{B}}$ is the function which is 1 at x_j and 0 elsewhere. The lemma is now immediate. □

Let us now assume for the duration of this paper that we have been given an m-class pattern recognition problem in a feature space F; that is, assume that we have been given, for some positive integer m and finite feature space F, an m-sequence $\boldsymbol{TR} = \langle TR_1, TR_2, \ldots, TR_m \rangle$ of nonempty (training) subsets of F and an m-sequence $\boldsymbol{TE} = \langle TE_1, TE_2, \ldots, TE_m \rangle$ of nonempty (test) subsets of F.

Based on our discussion in Section 2, we know that in order for this problem to be perfectly solvable, there must exist some $\boldsymbol{TR}$-uniform collection, $\boldsymbol{M}$, of subsets of F such that $\boldsymbol{TR}$ is $\boldsymbol{M}$-indiscernible from $\boldsymbol{TE}$, and such that the different training sets in $\boldsymbol{TR}$ are "discernible" with respect to sets in $\boldsymbol{M}$. For the duration of this paper, let us assume that $\boldsymbol{M}$ is a fixed $\boldsymbol{TR}$-uniform collection of subsets of F such that $\boldsymbol{TR}$ is $\boldsymbol{M}$-indiscernible from $\boldsymbol{TE}$. In order to formalize the concept of "discernible," we give the following definition.

DEFINITION. For $1 \leq i \leq m$ and $1 \leq j \leq m$, the (i, j)-*enrichment degree* of a subset $\boldsymbol{N}$ of $\boldsymbol{M}$ (written $e^{\boldsymbol{N}}_{i,j}$) is defined to be

$$\inf\left\{\left| r(M, TR_i) - r(M, TR_j) \right| \middle| M \in \boldsymbol{N}\right\}.$$

The subset $\boldsymbol{N}$ is said to be (i, j)-*enriched* if $e^{\boldsymbol{N}}_{i,j} > 0$.

If $\boldsymbol{N}$ is (i, j)-enriched for a particular i and j, then TR_i and TR_j are "discernible" from one another with respect to $\boldsymbol{N}$.

As we mentioned earlier, for each j, $1 \leq j \leq m$, $r(M, TR_j)$ measures the degree to which the set M captures points in TR_j. Thus, if M were equal to F, this value would be 1, and if M were equal to $\varnothing$, this value would be 0. For a fixed value of j, if we were to view r as a random variable defined on the space $\boldsymbol{F} \times \{TR_j\}$, then it is clear that its expectation would be $\frac{1}{2}$. For if one were to go through a process deciding, with equal probability, whether each point in a given set was to be removed or not, one would be expected, when done, to have removed half the points in the set.

Although it is by no means the only way to do it, one might view the process of creating an (i, j)-enriched, $\boldsymbol{TR}$-uniform collection of subsets of F as one which randomly generates subsets of F and selects those M such that $r(M, TR_i)$ and $r(M, TR_j)$ sit on opposite sides of the expected value $\frac{1}{2}$. This process need have no effect on the values of $r(M, TR_k)$ for k different from i and j, and if m is greater than 2, it is, in fact, desirable to select M which not only push $r(M, TR_i)$ and $r(M, TR_j)$ away from $\frac{1}{2}$, but which keep all other $r(M, TR_k)$ close to $\frac{1}{2}$. In some sense, such collections are "neutral" with respect to classes other than i and j. With this in mind, we give the following definition.

DEFINITION. For $1 \leq i \leq m$ and $1 \leq j \leq m$ the (i, j)-*neutrality degree* of a subspace $\boldsymbol{N}$ of $\boldsymbol{M}$ (written $n^{\boldsymbol{N}}_{i,j}$) is defined to be

$$\sup\left\{ \frac{r(M, TR_k) - r(M, TR_j)}{r(M, TR_i) - r(M, TR_j)} \middle| M \in \boldsymbol{N},\ k \neq i, j \right\}.$$

The subset $\boldsymbol{N}$ is said to be (i, j)-*neutral* if $n^{\boldsymbol{N}}_{i,j} < 1$.

The virtues of enrichment and neutrality will be made clear shortly. For the moment, however, let us simply note that if a given $\boldsymbol{TR}$-uniform space of weak models has a subspace which is either enriched or neutral, then it has a $\boldsymbol{TR}$-uniform subspace which is, similarly, enriched or neutral. For it is clear that any subspace of a $\boldsymbol{TR}$-uniform space $\boldsymbol{W}$ which can be written as a union of subspaces of the form $\boldsymbol{W}_{\boldsymbol{x}, \boldsymbol{TR}}$ must also be $\boldsymbol{TR}$-uniform.

We now formalize the random variable $X_{(i,j)}$ introduced in Section 2.

DEFINITION. For any pair of integers (i, j), $1 \leq i \leq m$ and $1 \leq j \leq m$, let $X_{(i,j)}$ be the function defined on $F \times \boldsymbol{M}$ as follows: for any pair (q, S) in $F \times \boldsymbol{M}$,

$$X_{(i,j)}(q, S) = \begin{cases} 2\left(\dfrac{\chi_S(q) - r(S, TR_j)}{r(S, TR_i) - r(S, TR_j)} \right) - 1, & \text{if } r(S, TR_i) \neq r(S, TR_j), \\ 0, & \text{if } r(S, TR_i) = r(S, TR_j), \end{cases}$$

where χ_S denotes the characteristic function of the set S.

Since both F and $\boldsymbol{M}$ are finite sets, we can view them, under counting measures, as finite measure spaces; as such, $X_{(i,j)}$ can be viewed as a random variable defined on the sample space $F \times \boldsymbol{M}$.

Often, when dealing with functions of several variables (such as r or $X_{(i,j)}$), we will wish to hold several of the variables fixed at constant values and consider the resulting expression to be a function of the remaining (nonfixed) variables. The following well-known lambda notation is helpful in representing such functions: if f is a function of $n + k$ variables $x_1, x_2, \ldots, x_n$

and $y_1, y_2, \ldots, y_k$, and $a_1, \ldots, a_k$ are k constants, then

$$\lambda x_1 \cdots \lambda x_n [f(x_1, x_2, \ldots, x_n, a_1, a_2, \ldots, a_k)]$$

denotes the function of n variables which results from taking f and holding its k variables $y_1, y_2, \ldots, y_k$ constant at $a_1, \ldots, a_k$, respectively.

We now consider the question of the expectations of the $X_{(i,j)}$.

LEMMA 3. *If A is a subset of F, $\mathbf{N}$ is an A-uniform subspace of $\mathbf{M}$ and q is a member of A, then the expectations of the random variables $\lambda M[\chi_M(q)]$ and $\lambda M[r(M, A)]$ (both restricted to the sample space $\mathbf{N}$) are identical.*

PROOF. By Lemma 1, for any real x, $P_{\mathbf{F}}(q \in M | r(M, A) = x) = x$. Thus, the expectation of $\lambda M[\chi_M(q)]$, restricted to $\mathbf{N}_{x,A}$, is x. Clearly, the expectation of $\lambda M[r(M, A)]$, restricted to $\mathbf{N}_{x,A}$, is also x. Since $\mathbf{N}$ can be written as a finite disjoint union of sets of the form $\mathbf{N}_{x,A}$, the lemma is now immediate. □

Thus, if $\mathbf{N}$ is an A-uniform subspace of $\mathbf{M}$, given any two members p and q of A, the random variables $\lambda M[\chi_M(p)]$ and $\lambda M[\chi_M(q)]$ have the same probability density functions.

LEMMA 4. *Let i and j be distinct integers, $1 \leq i \leq m$ and $1 \leq j \leq m$, and let $\mathbf{N}$ be a $\mathbf{TR}$-uniform, (i, j)-enriched subspace of $\mathbf{M}$. Then, for any k, $1 \leq k \leq m$, as q ranges over TR_k, the random variables which result from restricting $X_{(i,j)}$ to sample spaces of the form $\{q\} \times \mathbf{N}$, are identically distributed. Let E_k denote the common expectation of these random variables. Then $E_i = 1$, $E_j = -1$ and, if $k \neq i, j$, $E_k \leq 2n^{\mathbf{N}}_{(i,j)} - 1$.*

PROOF. If the space $\mathbf{N}$ is of the form $\mathbf{T}_{\boldsymbol{x},\mathbf{TR}}$ for some sequence $\boldsymbol{x} = \langle x_1, x_2, \ldots, x_m \rangle$ of real numbers, and some subspace $\mathbf{T}$ of $\mathbf{M}$, this lemma follows easily from Lemma 3. However, the subspace $\mathbf{N}$ can clearly be written as a finite disjoint union of nonempty sets of the form $\mathbf{N}_{\boldsymbol{x},\mathbf{TR}}$, and so the full lemma is now immediate. □

Until further notice, let us assume that we are dealing with a two-class problem ($m = 2$), and that there exists a subset $\mathbf{N}$ of $\mathbf{F}$ which is $\mathbf{TR}$-uniform and $(1, 2)$-enriched, such that $\mathbf{TR}$ and $\mathbf{TE}$ are $\mathbf{N}$-indiscernible. As in Section 2, $\mathbf{N}$ is called a space of weak models. Given our current restriction to two-class problems, our description of a stochastic discrimination model based on $\mathbf{N}$ can be made even simpler than that presented in Section 2. Indeed, our model is now basically that which classifies a point q in F to be of type 1 (type 2) if the average value of $\lambda M[X_{(1,2)}(q, M)]$ on a sufficiently large random sample from $\mathbf{N}$ is greater than or equal to 0 (is less than 0).

To make this precise, let t be a given positive integer, and let us denote by $X^k_{(1,2)}$ the random variable corresponding to $X_{(1,2)}$ associated with the kth of t trials, that is, the random variable defined on the sample space $F \times \mathbf{N}^t$ whose value at any point $(q, (S_1, S_2, \ldots, S_t))$ is $X_{(1,2)}(q, S_k)$. Let $Y^t_{(1,2)}$ denote

the random variable

$$\left(\sum_{k=1}^{t} X_{(1,2)}^{k} \right) \Big/ t.$$

By the central limit theorem, as t increases, the probability density function of $Y_{(1,2)}^{t}$ approaches a normal probability density function having expectation that of $X_{(1,2)}$ and having variance $1/t$ that of $X_{(1,2)}$.

In particular, if, for a given t and member $\mathbf{s}^t = (S_1, S_2, \ldots, S_t)$ of $\boldsymbol{N}^t$, we define the discrimination model $M_{\mathbf{s}^t}$ to be that which classifies a point q to be of class 1 if the value $Y_{(1,2)}^{t}(q, (S_1, S_2, \ldots, S_t))$ is greater than or equal to 0, and to be of class 2 if the value $Y_{(1,2)}^{t}(q, (S_1, S_2, \ldots, S_t))$ is less than 0, then the probability that $M_{\mathbf{s}^t}$ makes an error in classifying any point in the training set approaches 0 as t approaches ∞. Furthermore, given the $\boldsymbol{N}$-indiscernibility of $\boldsymbol{TR}$ and $\boldsymbol{TE}$, the probability that $M_{\mathbf{s}^t}$ makes an error in classifying any point in the test set also approaches 0 as t approaches ∞.

However, let us be careful here. The notion of "probability of error" in the context of pattern recognition almost always relates to the probability *relative to the space of points being classified* that a given point is improperly classified. In the discussion above concerning stochastic models, however, this is not what we are talking about. Any conclusions we have reached thus far concerning the accuracy of stochastic models is *relative to the space* $\boldsymbol{N}^t$ *of t-tuples of subsets of* $\boldsymbol{N}$. In other words, what we have shown is that for any given point q in the training or test set, if $m(t)$ denotes the probability that a member $\mathbf{s}^t$ of the space $\boldsymbol{N}^t$ leads to a derivative model $M_{\mathbf{s}^t}$ which misclassifies q, then as t approaches ∞, $m(t)$ approaches 0. Thus, while from the general perspective of producing solutions to pattern recognition problems, we have described a general modeling technique which succeeds in discriminating between classes, we would like to analyze our approach in such a way as to produce more conventional estimates of model accuracy.

We begin with a couple of definitions.

DEFINITION. Given a collection $\boldsymbol{N}$ of subsets of F, a *level-t stochastic model built from* $\boldsymbol{N}$ is a model of the form $M_{\mathbf{s}^t}$ for some member $\mathbf{s}^t$ of $\boldsymbol{N}^t$.

For binary discrimination problems, it is standard to equate any discrimination model with the set of points in the feature space which the model classifies as being of type 1. Given this, the following definition is reasonable.

DEFINITION. Given a binary discrimination model M and a sequence $\boldsymbol{T} = \langle T_1, T_2 \rangle$ of subsets of F, the *accuracy* of M relative to (separating the classes) T_1 and T_2 [denoted $a(M, \boldsymbol{T})$] is given by

$$a(M, \boldsymbol{T}) = r(M, T_1) - r(M, T_2).$$

Here $a(M, \boldsymbol{T})$ ranges between 1 and -1. If $a(M, \boldsymbol{T}) = 1$, the classification model M is perfect.

For the case at hand, every point $\mathbf{s}^t$ of $\boldsymbol{N}^t$ leads to a (potentially different) model $M_{\mathbf{s}^t}$. Thus, we are interested not so much in $a(M_{\mathbf{s}^t}, \boldsymbol{TR})$ for any particular $\mathbf{s}^t$, but rather in the expected value of $a(M_{\mathbf{s}^t}, \boldsymbol{TR})$ as a function of t. In particular, we would like some idea as to how large t must be so that the expected value of $a(M_{\mathbf{s}^t}, \boldsymbol{TR})$ is greater than, say, $1 - 1/u$.

Assume $\mathbf{s}^t = (S_1, S_2, \ldots, S_t)$ is a given member of $\boldsymbol{N}^t$. Since $M_{\mathbf{s}^t}$ is equal to

$$\left\{q | Y^t_{(1,2)}(q, \mathbf{s}^t) \geq 0\right\},$$

if, for each subset C of F, we let $g^C_{\mathbf{s}^t}$ denote the probability density function of the random variable $\lambda q[Y^t_{(1,2)}(q, \mathbf{s}^t)]$ defined on C, then it is clear that, for each k, $1 \leq k \leq 2$, $r(M_{\mathbf{s}^t}, TR_k)$ is equal to

$$\int_{[0,\infty)} g^{TR_k}_{\mathbf{s}^t}.$$

Thus,

$$a(M_{\mathbf{s}^t}, \boldsymbol{TR}) = \int_{[0,\infty)} \left(g^{TR_1}_{\mathbf{s}^t} - g^{TR_2}_{\mathbf{s}^t}\right).$$

As a result, we are led to look at the density functions $g^{TR_1}_{\mathbf{s}^t}$ and $g^{TR_2}_{\mathbf{s}^t}$. Given, however, our interest in expected accuracies of stochastic models and given that these pdf's $g^{TR_1}_{\mathbf{s}^t}$ and $g^{TR_2}_{\mathbf{s}^t}$ are themselves functions of $\mathbf{s}^t$, our interest is in examining the "expected pdf's," that is, the pdf's $g^{TR_1}_t$ and $g^{TR_2}_t$ where, for each subset C of F and real number r, $g^C_t(r)$ is equal to the expectation of the random variable $\lambda \mathbf{s}^t[g^C_{\mathbf{s}^t}(r)]$ defined on $\boldsymbol{N}^t$.

Assume t has been fixed. Any given member $\mathbf{s}^t = (S_1, S_2, \ldots, S_t)$ of $\boldsymbol{N}^t$ induces, in a natural way, partitions of TR_1 and TR_2 which we might describe as follows: given a function f from the set $\{1, 2, \ldots, t\}$ into $\{0, 1\}$ (i.e., f is a member of the set ${}^t2$), define, for each j, $1 \leq j \leq t$, S^f_j to be S_j if $f(j) = 1$, and $F - S_j$ if $f(j) = 0$. Then

$$\left\{ \bigcap_{j=1}^{t} S^f_j \middle| f \in {}^t2 \right\}$$

is a partition of F, which, in turn, induces partitions of TR_1 and TR_2.

In order to evaluate the pdf $g^{TR_1}_{\mathbf{s}^t}$, one must evaluate the sizes of the sets in the partition of TR_1; given our interest here in studying the "expected" pdf, we start by determining the expected sizes of the sets in this partition.

Let us fix a sequence of pairs of reals in $[0, 1]$:

$$\mathbf{z} = \langle (a_1, b_1), (a_2, b_2), \ldots, (a_t, b_t) \rangle,$$

such that the sample space

$$\mathbf{N} = \prod_{1 \leq j \leq t} \boldsymbol{N}_{(a_j, b_j), \boldsymbol{TR}}$$

is nonempty. Let us define, for each j, $1 \le j \le t$, a_j^f to be a_j if $f(j) = 1$, and $1 - a_j$ if $f(j) = 0$. Then it is easy to see that, for any member f of ${}^t2$, the random variable $x_{\mathbf{z}}^f$ defined on $\mathbf{N}$, whose value at any point $(S_1, S_2, \ldots, S_t)$ is $\mu(TR_1 \cap \bigcap_{i=1}^t S_i^f)$, has expectation

$$E(x_{\mathbf{z}}^f) = \mu(TR_1) \prod_{1 \le j \le t} a_j^f.$$

For a given member $\mathbf{s}^t = (S_1, S_2, \ldots, S_t)$ of $\mathbf{N}$, consider the (vector-valued) random variable

$$U_{\mathbf{s}^t} = \lambda p\left[\left(\chi_{S_1}(p), \chi_{S_2}(p), \ldots, \chi_{S_t}(p)\right)\right]$$

defined on TR_1. For any f in ${}^t2$, we clearly have that

$$P(U_{\mathbf{s}^t} = f) = \frac{\mu\left(TR_1 \cap \bigcap_{i=1}^t S_i^f\right)}{\mu(TR_1)}.$$

Thus,

$$E_{\mathbf{N}}(P(U_{\mathbf{s}^t} = f)) = \frac{E(x_{\mathbf{z}}^f)}{\mu(TR_1)} = \prod_{1 \le j \le t} a_j^f.$$

Let us now pick a member q of TR_1, and consider the (vector-valued) random variable

$$V_q = \lambda(M_1, M_2, \ldots, M_t)\left[\left(\chi_{M_1}(q), \chi_{M_2}(q), \ldots, \chi_{M_1}(q)\right)\right]$$

defined on $\mathbf{N}$. Given the $\boldsymbol{TR}$-uniformity of $\boldsymbol{N}$, the probability that the value of the jth variable in this vector takes on the value 1 is $r(M_j, TR_1) = a_i$. Thus, given the independence of the components of V_q, for any f in ${}^t2$, we clearly have that

$$P(V_q = f) = \prod_{1 \le j \le t} a_j^f.$$

Let us now consider, instead of $U_{\mathbf{s}^t}$ and V_q, the random variables $\lambda p[Y_{(1,2)}^t(p, \mathbf{s}^t)]$ (defined on TR_1) and $\lambda \mathbf{m}^t[Y_{(1,2)}(q, \mathbf{m}^t)]$ (defined on $\mathbf{N}$). Given the domain restrictions, it is clear that for any real number r there exists a subset $\boldsymbol{f}_r$ of ${}^t2$ such that

$$P\left(\lambda p\left[Y_{(1,2)}^t(p, \mathbf{s}^t)\right] = r\right) = \sum_{f \in \boldsymbol{f}_r} P(U_{\mathbf{s}^t} = f)$$

and

$$P\left(\lambda \mathbf{m}^t\left[Y_{(1,2)}(q, \mathbf{m}^t)\right] = r\right) = \sum_{f \in \boldsymbol{f}_r} P(V_q = f).$$

Thus, for any real number r,

$$E_{\mathbf{N}}\left(P\left(\lambda p\left[Y_{(1,2)}^t(p, \mathbf{s}^t)\right] = r\right)\right) = P\left(\lambda \mathbf{m}^t\left[Y_{(1,2)}(q, \mathbf{m}^t)\right] = r\right) = \sum_{f \in \boldsymbol{f}_r} \prod_{1 \le j \le t} a_j^f.$$

Since $\boldsymbol{N}^t$ can be partitioned into a disjoint union of such sets $\mathbf{N} = \prod_{1 \le j \le t} \boldsymbol{N}_{(a_j, b_j), \boldsymbol{TR}}$, and given that the argument above can just as well be carried out if we restrict our attention to TR_2 rather than TR_1, we have thus established the following result.

LEMMA 5 (The duality theorem). *Given a* $\boldsymbol{TR}$*-uniform subspace* $\boldsymbol{N}$ *of* $\boldsymbol{M}$ *and a positive integer* t*, for any* k*,* $1 \le k \le 2$*, and any* q *in* TR_k*,* $g_t^{TR_k}$ *is equal to the pdf of the random variable* $\lambda \mathbf{m}^t[Y_{(1,2)}(q, \mathbf{m}^t)]$ *defined on* $\boldsymbol{N}^t$*.*

Since the expected accuracy at separating the sets in $\boldsymbol{T} = (T_1, T_2)$ of a level-t stochastic model built from a member of $\boldsymbol{N}^t$ [henceforth denoted $e(t, \boldsymbol{T})$] satisfies

$$e(t, \boldsymbol{T}) = \int_{[0,\infty)} \left(g_t^{T_1} - g_t^{T_2}\right),$$

Lemma 5 allows us to use the pdf of the random variable $\lambda \mathbf{m}^t[Y_{(1,2)}(q, \mathbf{m}^t)]$ defined on $\boldsymbol{N}^t$ in our evaluation. And since $\lambda \mathbf{m}^t[Y_{(1,2)}(q, \mathbf{m}^t)]$ is a sum of independent identically distributed random variables, we can use Chebyshev's inequality to see that, for each k, $1 \le k \le 2$, given any q_k in TR_k and any h,

$$P_{\boldsymbol{N}^t}\left(\left|\frac{\sum_{i=1}^{t} \lambda M\left[X_{(1,2)}^i(q_k, M)\right]}{t} - E\left(\lambda M\left[X_{(1,2)}(q_k, M)\right]\right)\right| < \frac{1}{h}\right) > 1 - \frac{\sigma_k^2 h^2}{t},$$

where σ_k^2 is the variance of $\lambda M[X_{(1,2)}(q_k, M)]$. By Lemma 4, the distance from either $E(\lambda M[X_{(1,2)}(q_1, M)])$ or $E(\lambda M[X_{(1,2)}(q_2, M)])$ to 0 is 1. Thus, by taking h equal to 1, we immediately have that

$$\int_{[0,\infty)} g_t^{TR_1} > 1 - \frac{\sigma_1^2}{t}$$

and

$$\int_{[0,\infty)} g_t^{TR_2} < \frac{\sigma_2^2}{t}.$$

Thus,

$$e(t, \boldsymbol{TR}) > 1 - \frac{\sigma_1^2 + \sigma_2^2}{t}.$$

Since $\boldsymbol{N}$ is $(1,2)$-enriched, its enrichment degree, $d = e_{(1,2)}^{\boldsymbol{N}}$, is greater than 0. Clearly, both σ_1 and σ_2 are less than $4/d$. However, if one simply carries out the calculation [and uses the fact that when x is equal to $\frac{1}{2}$, the function $f(x) = x - x^2$ achieves its maximum value of $\frac{1}{4}$], it is not difficult to see that each of σ_1 and σ_2 is, in fact, less than $1/d$. As a result,

$$e(t, \boldsymbol{TR}) > 1 - \frac{2}{d^2 t},$$

and so, if we take t to be greater than $2u/d^2$, we would have

$$e(t, \boldsymbol{TR}) > 1 - \frac{1}{u}.$$

We summarize this discussion with the following theorem.

THEOREM 1. *For a given real number u, let $t^{\mathbf{N}}(u)$ denote the least t such that the expected accuracy, $e(t, \mathbf{TR})$, of level-t stochastic models built from a $\mathbf{TR}$-uniform, $(1,2)$-enriched space, $\mathbf{N}$, is greater than $1 - 1/u$. Then $t^{\mathbf{N}}(u)$ is bounded above by a value which is directly proportional to u and inversely proportional to the square of the enrichment degree of $\mathbf{N}$.*

We wish to note that we have made no effort here to establish tight bounds. This analysis was simply intended to get some rough theoretical sense for the computational feasibility of our method.

All of the discussion above concerning the expected accuracy of binary stochastic models involved accuracy only as measured on the training set $\mathbf{TR}$. What, then, can we conclude about the expected accuracy of such models when measured on the test set $\mathbf{TE}$? It turns that the expected accuracies are the same.

THEOREM 2. *Given any sequence $\mathbf{C} = \langle C_1, C_2 \rangle$ of subsets of F, if $\mathbf{TR}$ is $\mathbf{N}$-indiscernible from $\mathbf{C}$, then $e(t, \mathbf{TR}) = e(t, \mathbf{C})$.*

PROOF. Assume we are given a sequence $\mathbf{C} = \langle C_1, C_2 \rangle$ of subsets of F, such that $\mathbf{TR}$ is $\mathbf{N}$-indiscernible from $\mathbf{C}$. Consider the effect of carrying out our entire development so far, but everywhere replacing TR_k with C_k, $1 \leq k \leq 2$. Since $\mathbf{TR}$ is $\mathbf{N}$-indiscernible from $\mathbf{C}$, $r(M, TR_k) = r(M, C_k)$ for $1 \leq k \leq 2$, and so the definition of the random variable $X_{(i,j)}$ is unaffected (as are all factors associated with the enrichment or neutrality of $\mathbf{N}$). By Lemma 2, $\mathbf{N}$ is $\mathbf{C}$-uniform. Thus, the duality theorem applies, and so if we choose some member p of C_1, we have that $g_t^{C_1}$ is equal to the pdf of the random variable $\lambda \mathbf{m}^t[Y_{(1,2)}(p, \mathbf{m}^t)]$ defined on $\mathbf{N}^t$. However, by Lemma 3, since $r(M, TR_1) = r(M, C_1)$, for each i, $1 \leq i \leq t$, $\lambda M[X^i_{(1,2)}(q, M)]$ and $\lambda M[X^i_{(1,2)}(p, M)]$ are identically distributed. Thus, $\lambda \mathbf{m}^t[Y_{(1,2)}(q, \mathbf{m}^t)]$ and $\lambda \mathbf{m}^t[Y_{(1,2)}(p, \mathbf{m}^t)]$ (both defined on $\mathbf{N}^t$) are identically distributed. We have thus proved that $g_t^{TR_1} = g_t^{C_1}$. Similarly, $g_t^{TR_2} = g_t^{C_2}$. Since, for $1 \leq k \leq 2$, the expected value of $r(M_{\mathbf{s}^t}, TR_k)$ $(r(M_{\mathbf{s}^t}, C_k))$ is equal to $\int_{[0,\infty)} g_t^{TR_k}$ $[\int_{[0,\infty)} g_t^{C_k}]$, our proof is complete. □

In terms of our given pattern recognition problem, we see that since $\mathbf{TR}$ is $\mathbf{N}$-indiscernible from $\mathbf{TE}$, the expected performance of models produced by stochastic discrimination does not degrade when moving from the training set to the test set.

Furthermore, one can check that, for any k, $1 \leq k \leq 2$, the probability that $g_{\mathbf{s}^t}^{TR_k}$ is close to $g_t^{TR_k}$ approaches 1 as t goes to ∞, and so our confidence that the actual accuracy of any particular level-t stochastic model is close to the expected accuracy of level-t stochastic models, approaches 1 as t approaches ∞.

Given the discussion above directed at showing the rapid rate of convergence of stochastic modeling and given that the general question of accuracy for stochastic models is addressed in terms of various probability distribu-

tions which, by the central limit theorem, may be assumed to be essentially normal, it is fair to ask just how rapidly such distributions converge to normality. For this, we refer the reader to [16], where it is shown that this convergence is polynomial in the complexity of the problem being considered.

Thus far, we have been working with the random variables $X_{(i,j)}(q, M)$ because shortly, when we get to general multiclass pattern recognition ($m > 2$), we will require the normalized expectations provided by these variables. However, if one never planned to discriminate among more than two classes, we could have carried out our analysis using $\chi_M(q)$ instead of $X_{(i,j)}(q, M)$.

The key difference appears in Lemma 4, where instead of having $E_1 = 1$ and $E_2 = -1$, we would have $E_1 = E(\lambda M[r(M, TR_1)])$ and $E_2 = E(\lambda M[r(M, TR_2)])$. However, if we require that $r(M, TR_1) > r(M, TR_2)$ for every M in $\boldsymbol{N}$ (a somewhat stronger case of enrichment of $\boldsymbol{N}$ than before), we must then have that $E(\lambda M[r(M, TR_1)])$ is greater than $E(\lambda M[r(M, TR_2)])$. Thus, if ν denotes the mean of these two expectations, and if (similarly to the definition of the random variables $Y_{(i,j)}$) we let U^t denote the random variable

$$\left(\sum_{k=1}^{t} \chi_M^k(q) \right) \Big/ t,$$

we can then define the (type-U) stochastic discrimination model $T_{\mathbf{s}^t}$ to be that which classifies a point q to be of class 1 if the value $U^t(q, (S_1, S_2, \ldots, S_t))$ is greater than or equal to ν, and to be of class 2 if the value $U^t(q, (S_1, S_2, \ldots, S_t))$ is less than ν.

Otherwise, everything carries through as before. In fact, given the simpler form of our base random variable, we have a somewhat more general result concerning projectability.

THEOREM 3. *If $\boldsymbol{C}$ and $\boldsymbol{D}$ are $\boldsymbol{N}$-indiscernible from one another, then the expected accuracies of (type-U) binary stochastic models built from $\boldsymbol{N}$, at separating the sets in $\boldsymbol{C}$ and at separating the sets in $\boldsymbol{D}$, are equal.*

PROOF. See the proof of Theorem 2.

We now consider the notion of discrimination model when there are more than just two classes of points. For motivational purposes, let us once again consider the two-class case, but this time with a somewhat different slant. For suppose that in addition to just choosing a single random sample $\mathbf{s}^t$ of size t from a $(1, 2)$-enriched, uniform subspace of $\boldsymbol{M}$, we also choose a random sample $\mathbf{u}^t$ of size t from a (possibly different) $(2, 1)$-enriched, uniform subspace of $\boldsymbol{M}$. Then we have two functions to consider, namely, $k_1(q) = Y^t_{(1,2)}(q, \mathbf{s}^t)$ and $k_2(q) = Y^t_{(2,1)}(q, \mathbf{u}^t)$. And, based on the discussion above, we would expect the value of $k_1(q)$ to be greater than the value of $k_2(q)$ for points q of class 1, and we would expect the value of $k_1(q)$ to be smaller than the value of $k_2(q)$ for points q of class 2. Thus, the function k_1, which is

derived from a space of sets enriched with respect to points of class 1 (at the expense of points of class 2), is a measure of the degree to which a given point is 1-like (the higher the value of k_1 the more 1-like a point is), and the function k_2, which is derived from a space of sets enriched with respect to points of class 2 (at the expense of points of class 1), is a measure of the degree to which a given point is 2-like. So, to classify a given point q, we would simply evaluate $k_i(q)$ for $1 \leq i \leq 2$, and if $k_n(q)$ turned out to be the greater of the two values, then n would be the classification we gave q.

In just this way, we will produce models for discriminating among many classes. Indeed, for any positive integer m, given an m-class discrimination problem, we will produce functions k_i for each i between 1 and m, where each k_i will be based on a random sample from a space enriched with respect to points of class i. As before, $k_i(q)$ will, in some sense, measure the degree to which a given point q is i-like; given a point q, our discrimination model will simply classify q as being of class i if $k_i(q)$ is the largest value among $\{k_1(q), k_2(2), \ldots, k_m(q)\}$.

There are a number of ways in which we can produce these functions k_i. For example, we could deal simply with m subspaces of $\boldsymbol{M}$ where, for each i, $1 \leq i \leq m$, the expectation of $\lambda M[r(M, TR_i)]$ restricted to the ith subspace was greater than any of the expectations of the $\lambda M[r(M, TR_j)]$ for j not equal to i. However, given potential problems involved with achieving neutrality in practical applications, a more refined and accurate approach involves carrying out all enrichment in terms of binary pairs. We proceed as follows:

for any given pair of distinct integers i and j, assume that $\boldsymbol{N}_{(i,j)}$ is a uniform, (i, j)-enriched, (i, j)-neutral subspace of $\boldsymbol{M}$ (with $\boldsymbol{TR}$ $\boldsymbol{N}_{(i,j)}$-indiscernible from $\boldsymbol{TE}$), and assume that we have chosen a random sample, $\mathbf{s}^t_{(i,j)}$ of size t from $\boldsymbol{N}_{(i,j)}$. Then we have $m(m\text{-}1)$ functions, namely, the functions $\lambda q[Y^t_{(i,j)}(q, \mathbf{s}^t_{(i,j)})]$ discussed above. We now define m new functions k^t_i by simply setting, for each i between 1 and m, $k^t_i(q)$ equal to

$$\left(\sum_{\substack{j=1 \\ j \neq i}}^{m} Y^t_{(i,j)}\left(q, \mathbf{s}^t_{(i,j)}\right) \right) \Big/ (m - 1).$$

The following points are now immediate from our earlier discussion: for each i between 1 and m:

1. The expected value of the function k^t_i, when it is restricted to points of class i, is 1 (see Lemma 4).
2. If n_i denotes the largest of the neutrality degrees $n^{\boldsymbol{N}}_{(i,j)}$ for $1 \leq j \leq m$, $j \neq i$, then the expected value of the function k^t_i, when it is restricted to points of class other that i, is less than or equal to $2n_i - 1$ (< 1) (see Lemma 4).

3. The probability that k_i^t deviates greatly from these expected values approaches 0 as t approaches ∞; the rate of this convergence is rapid in t.
4. These facts hold whether we view k_i^t as being a function of points in the union of the TR_i, or a function of points in the union of the TE_i.

As a result, we may define our m-class discrimination model, E^t, as follows:

> given any point q, evaluate $k_i^t(q)$ for each i between 1 and m; find the least i such that $k_i^t(q)$ is greater than or equal to $k_j^t(q)$ for each j between 1 and m; now classify q as being of class i.

In light of our discussion above, the expected accuracy of this discrimination model E^t can be made as high as desired by choosing t to be sufficiently large.

The reader might have noticed that from a theoretical point of view, multiclass discrimination could just as well have been carried out by simply defining, for each i between 1 and m, k_i^t to be $Y^t_{(i,(i+1)\bmod(m))}$. However, given the way we did define k_i^t, there is what might be called a *secondary stochastic effect* taking place here. For the essence of stochastic discrimination lies in the fact that as one forms a random variable by taking a sum of independent random variables, the variance of that sum approaches 0 as the number of component variables approaches ∞. Thus, in the case of k_i^t, a sum of $m - 1$ variables, each of which is itself a sum of t variables, large values of m contribute to the desired small variance just as do large values of t. Furthermore, in practical applications, where both uniformity and neutrality are sometimes difficult to achieve, the impact from defining k_i^t as the *sum* of $m - 1$ random variables on ameliorating deficiencies in strict uniformity or neutrality is often significant.

5. A geometric interpretation of binary stochastic modeling. Let $\boldsymbol{N}$ be a fixed $(1, 2)$-enriched, uniform subspace of $\boldsymbol{M}$. Let q be any given point in F. Then by carrying out random sampling (with replacement) from $\boldsymbol{N}$, we can build a record of numeric fields associated with q by simply taking the sequence of values of $\lambda M[X_{(1,2)}(q, M)]$ on successive points in the sample. In this way, we can map the original feature space F into higher and higher dimensional Euclidean spaces (in effect, new feature spaces), and we claim that no matter how accurate a model one desires, if we go to a high enough dimension, t, then (the expectation is that) some $t - 1$-dimensional hyperplane in that Euclidean t-space would affect a model with that degree of accuracy. In other words, given a desired degree of accuracy, $1 - \varepsilon$, there exists a t such that the expected accuracy of the discrimination model $M_{\boldsymbol{H}}$ determined by some $t - 1$-dimensional hyperplane $\boldsymbol{H}$ in Euclidean t-space is greater than $1 - \varepsilon$. ($M_{\boldsymbol{H}}$ classifies a point to be of class 1 if it is on one side of $\boldsymbol{H}$, and of class 2 if it is on the other side of $\boldsymbol{H}$.) In effect, we have a method here for adding features so that models based on linear discriminant analysis

continue to increase in accuracy monotonically (to any desired degree) as we increase the number of features. [Viewing our method in this way provides another perspective on its virtue as a general tool for pattern recognition. For as Bellman (see [5]) and others have demonstrated, adding features often leads to a decrease in model projectability. Thus, this well-known "curse of dimensionality" is completely reversed if one adds features as described above.]

In order to see this, let t be a fixed positive integer, and let us consider, given a random sample $\mathbf{s}^t = (S_1, S_2, \ldots, S_t)$ of size t from $\boldsymbol{N}$, the embedding $e_{\mathbf{s}^t}$ from F into Euclidean t-space which sends any q in F to the vector whose kth coordinate (for $1 \leq k \leq t$) is equal to $X_{(i,j)}(q, S_k)$. Let $\boldsymbol{H}(t)$ denote the $t-1$-dimensional hyperplane in Euclidean t-space which passes through the origin and which is orthogonal to the vector $\boldsymbol{\nu}_t$, all of whose coordinates are equal to 1. Working with the metric on Euclidean t-space which results when one contracts the standard Euclidean metric by the factor $1/\sqrt{t}$, it is easy to see that for any random sample $\mathbf{s}^t$ of size t from $\boldsymbol{N}$, given a point q in F, the signed distance from $e_{\mathbf{s}^t}(q)$ to $\boldsymbol{H}(t)$ is equal to $Y^t_{(1,2)}(q, \mathbf{s}^t)$. [The sign of the distance is "+" if q is on the same side of $\boldsymbol{H}(t)$ as the point $\boldsymbol{\nu}_t$; it is "−" otherwise.] Thus, intuitively, the probability density function for the random variable $\lambda \mathbf{s}^t[Y^t_{(1,2)}(q, \mathbf{s}^t)]$ really represents the probable locations where one can expect to find the point q in Euclidean t-space [vis-à-vis its distance from the hyperplane $\boldsymbol{H}(t)$] after q is passed through a "generic" embedding. In light of our discussion above concerning the accuracy of binary stochastic models, we see that the density function $g^{TR_1}_{\mathbf{s}^t}$ portrays the distribution of points in TR_1 with respect to their signed distance from the hyperplane $\boldsymbol{H}(t)$, and the function $g^{TR_2}_{\mathbf{s}^t}$ portrays the distribution of points in TR_2 with respect to their signed distance from $\boldsymbol{H}(t)$.

By the central limit theorem, as t increases, we can think of $g^{TR_1}_{\mathbf{s}^t}$ and $g^{TR_2}_{\mathbf{s}^t}$ as normal functions with variances inversely proportional to t, and since these functions are centered at 1 and -1, respectively, we see, given our previous discussion, that as t increases without bound, the likelihood of finding members of F of like class on opposite sides of the hyperplane $\boldsymbol{H}(t)$ decreases to 0. Thus, as t approaches ∞, the expected accuracy of such a hyperplane-based discrimination model goes to 1. In fact, the hyperplane-based model $M_{\boldsymbol{H}(t)}$ which classifies a point q to be of class 1 (2) if the sign of the distance of $e_{\mathbf{s}^t}(q)$ to $\boldsymbol{H}(t)$ is "+" ("−") is identical to the model $M_{\mathbf{s}^t}$ defined in Section 4.

Acknowledgments. The author wishes to thank R. Berlind, T. Biehler, D. Bowen and T. K. Ho for their useful comments based on preliminary versions of this paper. Many of their suggestions have been included here. In addition, Dr. Ho provided access to many of the data sets used in this paper. Her help in dealing with these, as well as in supplying nearest-neighbor results included in this paper, in writing the graphical tools used in producing the pictures of Section 2, and, in general, with supplying perspective relating our methods to more traditional methods, is greatly appreciated. We

also wish to thank E. G. Kleinberg for sharing many helpful insights and for technical assistance in producing some of the neural network results discussed here. Finally, we thank L. D. Brown for his patient help in suggesting many revisions of our original manuscript.

REFERENCES

[1] AMIT, Y., GEMAN, D. and WILDER, K. (1996). Recognizing shapes from simple queries about geometry. Unpublished manuscript.
[2] BERLIND, R. (1994). An alternative method of stochastic discrimination with applications to pattern recognition. Ph.D. dissertation, Dept. Mathematics, State Univ. New York, Buffalo.
[3] BERLIND, R. (1994). Almost uniformity in stochastic modeling. Unpublished manuscript.
[4] COVER, T. M. and HART, P. E. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory* **IT-13** 21–27.
[5] DUDA, R. O. and HART, P. E. (1973). *Pattern Classification and Scene Analysis*. Wiley, New York.
[6] FREUND, Y. (1995). Boosting a weak learning algorithm by majority. *Inform. and Comput.* **121** 256–285.
[7] GEMAN, S., BIENENSTOCK, E. and DOURSAT, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation* **4** 1–58.
[8] GOLDMAN, S. A., KEARNS, M. J. and SCHAPIRE, R. E. (1995). On the sample complexity of weakly learning. *Inform. and Comput.* **117** 276–287.
[9] GUYON, I. (1991). Applications of neural networks to character recognition. In *Character and Handwriting Recognition* (P. S. P. Wang, ed.). World Scientific, Singapore.
[10] HARALICK, R. M. (1976). The table look-up rule. *Comm. Statist. Theory Methods* **5** 1163–1191.
[11] HECHT-NIELSEN, R. (1990). *Neurocomputing*. Addison-Wesley, Reading, MA.
[12] HO, T. K. (1992). A theory of multiple classifier systems and its application to visual word recognition. Ph.D. thesis, Dept. Computer Science, State Univ. New York, Buffalo.
[13] HO, T. K. (1993). Recognition of handwritten digits by combining independent learning vector quantizations. In *Proceedings of the Second International Conference on Document Analysis and Recognition* (M. Kavanaugh, ed.) 818–821. IEEE Computer Society Press, New York.
[14] HO, T. K. (1995). Random decision forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition* (M. Kavanaugh and P. Storms, eds.) 278–282. IEEE Computer Society Press, New York.
[15] KANAL, L. (1974). Patterns in pattern recognition: 1968–1974. *IEEE Trans. Inform. Theory* **IT-20** 697–722.
[16] KLEINBERG, E. M. (1990). Stochastic discrimination. *Annals of Mathematics and Artificial Intelligence* **1** 207–239.
[17] KLEINBERG, E. M. and HO, T. K. (1993). Pattern recognition by stochastic modeling. In *Proceedings of the Third International Workshop on Frontiers in Handwriting Recognition* (M. Bosker, R. Casey, et al., eds.) 175–183. Partners Press, Buffalo, NY.
[18] KLEINBERG, E. M. and HO, T. K. (1996). Building projectable classifiers of arbitrary complexity. In *Proceedings of the 13th International Conference on Pattern Recognition* (M. E. Kavanaugh and B. Werner, eds.) 880–885. IEEE Computer Society Press, New York.
[19] PROJECT STATLOG (1992). *LIACC*. Univ. Porto. Internet address: ftp.ncc.up.pt:pub/statlog/datasets.
[20] SCHAPIRE, R. E. (1990). The strength of weak learnability. *Machine Learning* **5** 197–227.

DEPARTMENT OF MATHEMATICS
STATE UNIVERSITY OF NEW YORK
BUFFALO, NEW YORK 14214
E-MAIL: kleinbrg@cs.buffalo.edu