Neural network classification of quark and gluon jets

M. A. Graham and L. M. Jones
Physics Department, University of Illinois, 1110 W. Green St., Urbana, Illinois 61801

S. Herbin
32/132 rue Basse, 59800 Lille, France

(Received 22 October 1993)

We demonstrate that there are characteristics common to quark jets and to gluon jets regardless of the interaction that produced them. The classification technique we use depends on the mass of the jet as well as the center-of-mass energy of the hard subprocess that produces the jet. In addition, we present the quark-gluon separability results of an artificial neural network trained on three-jet e^+e^- events at the Z^0 mass, using a back-propagation algorithm. The inputs to the network are the longitudinal momenta of the leading hadrons in the jet. We tested the network with quark and gluon jets from both $e^+e^- \to 3$ jets and $p\bar{p} \to 2$ jets. Finally, we compare the performance of the artificial neural network with the results of making well chosen physical cuts.

PACS number(s): 13.87.−a, 14.65.−q, 14.70.Dj, 42.79.Ta

I. INTRODUCTION

Recent interest in using neural networks for quark-gluon jet discrimination has, in particular, addressed the issue of finding a method of recognizing quark and gluon jets from e^+e^- interactions [1, 2]. In this paper, we demonstrate that quark jets and gluon jets have characteristics which do not depend on the interaction that produces them: They are the same in $e^+e^- \to q\bar{q}$, $e^+e^- \to q\bar{q}g$, and $p\bar{p} \to 2$ jets+anything, provided one specifies the jet mass and the center-of-mass energy of the hard subprocess that produces the jet. This means it is possible to train a network on e^+e^- jets and to use it for jet identification in the more complex $p\bar{p}$ situation.

We also find that a linear discriminant function achieves the same performance as a nonlinear neural network in distinguishing light quark jets from gluon jets. The performance of the neural networks is marginally better than cuts based on the longitudinal momentum fraction of the fastest two particles in the jet or on the jet "bin label" number.

In the first section we describe the method we use to simulate the data. Next, we define the invariant longitudinal momentum fraction for a hadron in a jet and use this momentum fraction to define a single variable that contains information about the nine leading hadrons in the jet.

In Sec. IV, histograms of the jet variable for quark jets produced from different reactions are compared, as well as histograms from the same reaction at different center-of-mass energies.

We present a brief discussion of the back-propagation technique in general (Sec. V) and, in particular, the case where a back propagation network reduces to a linear discriminant function (Sec. VI).

In order to determine whether a neural network is able to achieve superior quark-gluon jet separability, we calculate in Sec. VII the fraction of quark and gluon jets that can be correctly classified by making simple cuts in the input data.

Next, we examine the performance of a two-layer and a three-layer back-propagation neural network (Sec. VIII). We train the networks with three different parameter sets and compare the performance of (1) the three-layer network to the two-layer network and (2) the performance of each network as it depends on the training parameters.

Finally, in Sec. IX, we summarize our results and discuss the outlook for quark-gluon jet identification.

II. CALCULATIONS

We used the high-energy physics Monte Carlo event generator HERWIG [3] to simulate the reactions $e^+e^- \to 3$ jets at 60 and 91.2 GeV center-of-mass energies, $e^+e^- \to 2$ jets at 36, 60, and 91.2 GeV, and $p\bar{p} \to 2$ jets at Fermilab Tevatron energies with varying subenergies for the hard subprocess center of mass. Only light quark (u, d, s) events were considered.

In order to obtain a three-jet sample with well-separated jets, we considered an e^+e^- event to have three jets only if the parton thrust was less than 0.9. More than 100 000 such events were produced for center-of-mass energy 91.2 GeV.

The final state hadrons in each event were separated into jets using the algorithm described in Appendix A. This algorithm is not one of the standard algorithms used in the experimental determination of jets. In fact, unlike the currently standard algorithms, it is not infrared safe. However, its computer implementation is simple and fast. We believe that the particle content of the jets identified with our algorithm should be nearly the same as the particle content of jets identified with one of the standard algorithms, at least for most jets. Thus we believe that the results obtained in this paper will carry over to the results with other algorithms. Studies with other algorithms are underway.

We seek to classify jets as being either quark jets or

0556-2821/95/51(9)/4789(19)/$06.00

gluon jets. We understand by a "quark jet" a jet that originated from a hard interaction as a high transverse momentum quark, while a "gluon jet" is one that originated from a gluon. Thus we consider that each jet arises from a parent parton that is either a quark or a gluon. We seek to identify the identity of the parent parton with as much statistical reliability as possible based on the configuration of the final state hadrons in the jet. There is a certain ambiguity inherent in the concept of a jet emerging from a single parent parton. For our computer-generated events, we settled this ambiguity by defining a jet as originating from the parton whose four-momentum most closely matched the jet four-momentum. We rejected jets whose four-momentum did not correspond sufficiently well with *any* parton. The cut that we used for this purpose was

$$(E_{\text{jet}} - E_{\text{part}})^2 + (p^1_{\text{jet}} - p^1_{\text{part}})^2 + (p^2_{\text{jet}} - p^2_{\text{part}})^2 + (p^3_{\text{jet}} - p^3_{\text{part}})^2 \leq R^2,$$

where we have chosen $R = 10$ GeV.

A method is needed which can deal with the facts that (i) the number of particles varies from jet to jet, and (ii) "wee" particles are hard to measure and difficult to assign to a particular jet. Because jets *are* "jetty," longitudinal momentum probably carries most of the information. A jet classification scheme discussed in previous papers [6, 7] allows one to plot the density in (multidimensional) longitudinal phase space as a function of one variable, thus simplifying visualization of the distributions. We use this technique in the sections below to compare quark and gluon jets from the same reactions as well as comparing quark jets from different reactions and gluon jets from different reactions. For the convenience of the reader, we briefly summarize the approach here; more details are given in the original reference.

III. JET VARIABLE

Our method of classification uses a binning procedure that uniquely assigns an integer to each jet based on the longitudinal momenta of the hadrons that make up the jet. Define the invariant longitudinal momentum fraction of each hadron to be

$$z_i = \frac{(E_i + p_i^{\text{long}})}{(E_{\text{jet}} + p_{\text{jet}})}.$$

Any hadron with a longitudinal momentum fraction less than 0.1 is ignored. Because we ignore $z_i \leq 0.1$, there are at most nine-hadrons to consider. We thus are dealing with a nine-dimensional longitudinal phase space. We wish to compare the jet density in this space for both quark and gluon jets; thus we need to find a good "grid" to impose on the space.

Our grid is formed by making hypercubes of side $\mid \Delta z \mid = 0.1$ in the longitudinal phase space. The ith hadron is associated with an integer j_i where j_i is the integer part of $(10z_i)$ with $0 \leq j_i < 10$. The resulting nine j_i's can be ordered and used to define a "feature vector" for each jet, $\mathbf{x} = (j_1, j_2, ..., j_9)$, where the "feature space" is nine dimensional and discrete. Each dimension of the feature space represents the discretized longitudinal momentum of a given hadron in the jet; therefore, this is basically a multiparticle longitudinal phase space.

There are regions of feature space that cannot be populated because of momentum conservation and the ordering constraint, and in fact, we find that there are only 97 hypercubes that may be occupied. It would be convenient to project this nine-dimensional space onto a lower dimensional space. We can represent the space on a one-dimensional line by using the 97 hypercubes as bins in a histogram; contributions to each bin show the jet population of the corresponding region in feature space. The bins are roughly ordered according to fast leading particles, with bin 1 having a single hard particle and bin 97 having no hard particles. Our bin patterns are given in Table I. Histograms of the $e^+e^- \rightarrow 3$ jets at 91.2 GeV and $p\bar{p} \rightarrow 2$ jets events are given in Fig. 1. We can see immediately that the distributions for quark jets differ greatly from those for gluon jets.

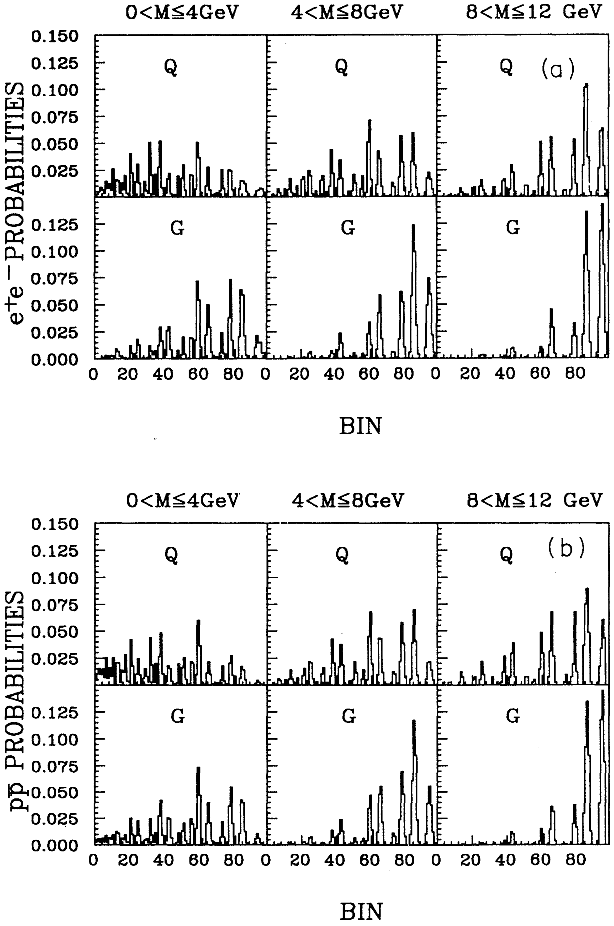

FIG. 1. Plots of correlation bin frequencies for quark and gluon jets. The events are divided into these ranges of jet mass (0–4, 4–8, and 8–12 GeV). (a) Quark and gluon jets produced in the reaction $e^+e^- \rightarrow 3$ jets. (b) Quark and gluon jets produced from $p\bar{p} \rightarrow 2$ jets.

TABLE I. Definition of the correlation bins.

Bin		Bin		Bin		Bin	
1	9	26	5	51	3,3	76	2,2,1,1,1
2	8,1	27	4,4,1	52	3,2,2,2	77	2,2,1,1
3	8	28	4,4	53	3,2,2,1,1	78	2,2,1
4	7,2	29	4,3,2	54	3,2,2,1	79	2,2
5	7,1,1	30	4,3,1,1	55	3,2,2	80	2,1,1,1,1,1,1,1
6	7,1	31	4,3,1	56	3,2,1,1,1,1	81	2,1,1,1,1,1,1
7	7	32	4,3	57	3,2,1,1,1	82	2,1,1,1,1,1
8	6,3	33	4,2,2,1	58	3,2,1,1	83	2,1,1,1,1
9	6,2,1	34	4,2,2	59	3,2,1	84	2,1,1,1
10	6,2	35	4,2,1,1,1	60	3,2	85	2,1,1
11	6,1,1,1	36	4,2,1,1	61	3,1,1,1,1,1,1	86	2,1
12	6,1,1	37	4,2,1	62	3,1,1,1,1,1	87	2
13	6,1	38	4,2	63	3,1,1,1,1	88	1,1,1,1,1,1,1,1,1
14	6	39	4,1,1,1,1,1	64	3,1,1,1	89	1,1,1,1,1,1,1,1
15	5,4	40	4,1,1,1,1	65	3,1,1	90	1,1,1,1,1,1,1
16	5,3,1	41	4,1,1,1	66	3,1	91	1,1,1,1,1,1
17	5,3	42	4,1,1	67	3	92	1,1,1,1,1
18	5,2,2	43	4,1	68	2,2,2,2,1	93	1,1,1,1
19	5,2,1,1	44	4	69	2,2,2,2	94	1,1,1
20	5,2,1	45	3,3,3	70	2,2,2,1,1,1	95	1,1
21	5,2	46	3,3,2,1	71	2,2,2,1,1	96	1
22	5,1,1,1,1	47	3,3,2	72	2,2,2,1	97	no hard particles
23	5,1,1,1	48	3,3,1,1,1	73	2,2,2		
24	5,1,1	49	3,3,1,1	74	2,2,1,1,1,1,1		
25	5,1	50	3,3,1	75	2,2,1,1,1,1		

IV. STATISTICAL RESULTS

In order to compare the samples of quark and gluon jets from e^+e^- and $p\bar{p}$, we used the Kolmogorov-Smirnov (KS) two-sample test. This test is described in Appendix B and in a previous paper [7].

Our method involves representing the jet in terms of the bin number introduced in Sec. III. These bin numbers are histogrammed and the KS statistic used to compare the histograms. Consider for example the quark jets from $e^+e^- \to 2$ jets at varying center-of-mass energies 36, 60, and 91.2 GeV, shown in Fig. 2. There is substantial dependence on the mass of the jets. In addition there is a (slower) dependence on the energy of the hard process producing the jet. The KS statistics computed for all combinations of these events are given in Tables II, III, and IV. In all three of the tables, there is at least one occurrence of a KS statistic that is larger than 2. Pairs of histograms for which the KS statistic is larger than 2 cannot be derived from the same distribution. Therefore, there is a dependence of the jet bin number on center-of-mass energy $\hat{s}$ as well.

Next we compare jets from $e^+e^- \to 2$ jets with the quark jets from $p\bar{p} \to 2$ jets; see Table V. When the histograms include jets of subprocess center-of-mass energies, between 1000 and 1600 GeV^2, the comparison gives a KS statistic below 2, and therefore indicates that these histograms may be drawn from the same distribution. The same can be said for the $\hat{s}$ ranges of 1500 GeV^2 $\leq \hat{s} \leq$ 4500 GeV^2 and $\hat{s} \geq$ 2500 GeV^2.

In order to make a comparison using gluon jets from e^+e^- we need to define a center of mass for the hard subprocess. We use a definition based on classical physics: $\hat{s} = (p_2^\mu + p_g^\mu)^2$ where p_2^μ is the four-momentum of the quark (or antiquark) jet that is observed to have a smaller angle with the gluon jet, and p_g^μ is the four-momentum

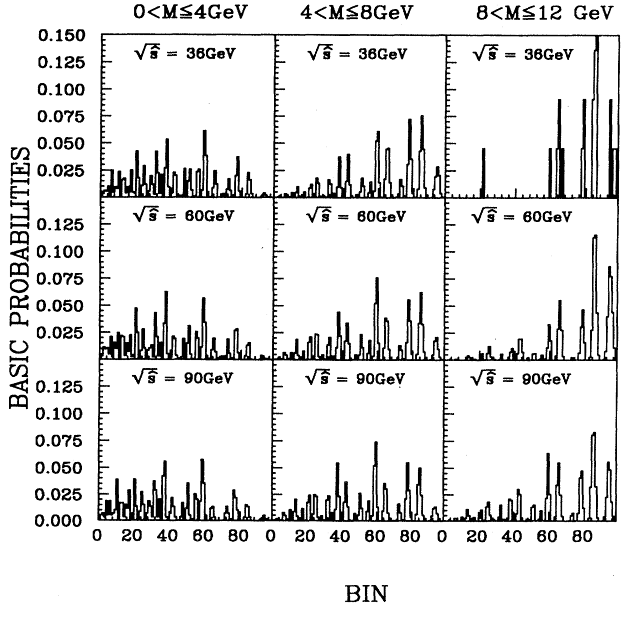

FIG. 2. Plots of correlation bin frequencies for quark jets from $e^+e^- \to 2$ jets. The plots are divided according to jet masses (0–4, 4–8, and 8–12 GeV) and center-of-mass energy (36, 60, and 90 GeV).

TABLE II. Values of the KS statistic for comparison of quark jets created in $e^+e^- \to 2$ jets at center-of-mass energy 36 GeV with those created at 60 GeV.

Jet mass	D_{MN}	M^a	N^b
0–4 GeV	1.37	2281	1864
4–8 GeV	5.45	2559	8733
8–12 GeV	0.63	22	2309

[a] M is the number of jets in the 36 GeV histogram.
[b] N is the number of jets in the 60 GeV histogram.

TABLE IV. Values of the KS statistic for comparison of quark jets created in $e^+e^- \to 2$ jets at center-of-mass energy 36 GeV with those created at 90 GeV.

Jet mass	D_{MN}	M^a	N^b
0–4 GeV	1.06	2281	593
4–8 GeV	7.64	2559	5986
8–12 GeV	1.30	22	4518

[a] M is the number of jets in the 36 GeV histogram.
[b] N is the number of jets in the 90 GeV histogram.

of the gluon jet.

Now consider the comparison of jets from $e^+e^- \to 3$ jets and $p\bar{p} \to 2$ jets. The KS statistic for comparing histograms of definite mass and $\hat{s}$ is less than 2, indicating a reasonable agreement in the distributions. This is true for the e^+e^- events created at 60 and 90 GeV, Tables VI and VII.

To quantify the large differences between quarks and gluons, we compare quark and gluon jets from the same reaction. Table VIII gives the result of comparing quark jets and gluon jets from $e^+e^- \to 3$ jets at 91.2 GeV. In comparing histograms that have sufficient statistics ($\sim$ 1000 entries), we find that the KS statistic is much larger than 2. This is true for comparing e^+e^- quark and gluon jets produced at 60 GeV, Table IX, and $p\bar{p}$ quark and gluon jets, Table X, as well.

The large differences between quark and gluon distributions lead us to believe that individual jets can be labeled as either quark of gluon, to some level of accuracy. The rest of this paper is a study of the level of discrimination which can be produced by standard methods.

V. BACK-PROPAGATION APPROACH

For this study, we used JETNET [8] as our back-propagation neural network. Our network is a feed-forward multilayer perceptron network using the gradient descent minimization technique and the back-propagation updating algorithm [8].

Back propagation is a functional fitting of the outputs (jet type) to the inputs (physics variables). It is applied to "neurons" arranged in a layered structure. There is an input layer of neurons, an output layer, and one or more hidden layers. There are no feedback connections and no connections that bypass one layer to go directly to another layer. The inputs to and outputs from each hidden neuron are determined by connection weights [9, 8].

TABLE III. Values of the KS statistic for comparison of quark jets created in $e^+e^- \to 2$ jets at center-of-mass energy 60 GeV with those created at 90 GeV.

Jet mass	D_{MN}	M^a	N^b
0–4 GeV	0.31	1864	593
4–8 GeV	3.68	8733	5986
8–12 GeV	7.53	2309	4518

[a] M is the number of jets in the 60 GeV histogram.
[b] N is the number of jets in the 90 GeV histogram.

Back propagation is an example of supervised learning for the network. An input vector is applied to the input layer, and the corresponding output vector from the output layer is computed. The calculated outputs are compared with the desired outputs to give an error function. The generalized delta rule (GDR), explained below, is used as the learning algorithm; that is, the GDR is used to determine the appropriate amount to change the connection weights. The weights are modified and the procedure is repeated until the errors are acceptable for all inputs in the training set. The training set contains a large number of events, so that the network "sees" examples of the possible data it will encounter after training. Once the network is trained (has produced the optimal set of connection weights) it may be used on data for feature recognition problems, such as the quark-gluon jet problem.

As an example, consider a three-layer network (Fig. 3). An input vector (using the notation of Rumelhart *et al.* [9]) for the pth training set, $\mathbf{x_p} = (\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N})$, is applied to the input layer. In our case, this could be, for instance, the z_i for the fastest particles in the jet. The net input to the jth neuron in the hidden layer is

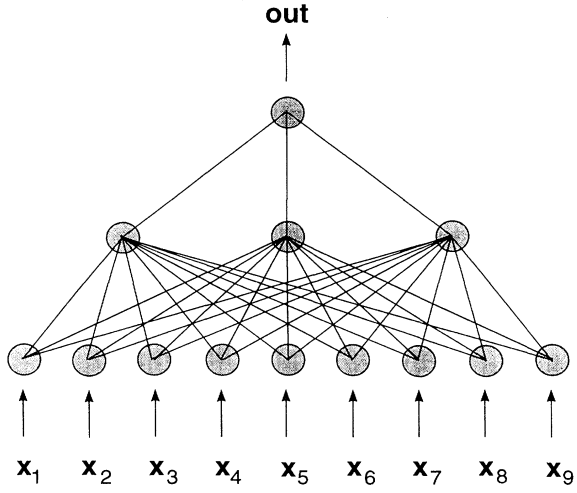

FIG. 3. A three-layer feed-forward artificial neural network.

TABLE V. Values of the KS statistic for comparison between quark jets from $e^+e^- \rightarrow 2$ jets and $p\bar{p} \rightarrow 2$ jets. The 36 and 60 GeV refer to the center-of-mass energy of the e^+e^- events. $\hat{s}$ is the center-of-mass energy squared of the hard subprocess in $p\bar{p}$.

Jet mass	$\hat{s}$	D_{MN}	M[a]	N[b]
			36 GeV e^+e^- events	
0–4 GeV	All[c]	2.20	4498	6048
4–8 GeV	All	1.66	5117	6560
8–12 GeV	All	1.77	57	781
0–4 GeV	1000–1600	1.12	4498	753
4–8 GeV	1000–1600	1.03	5117	1525
8–12 GeV	1000–1600	1.15	57	138
			60 GeV e^+e^- events	
0–4 GeV	All	2.93	1864	6048
4–8 GeV	All	7.29	8733	6560
8–12 GeV	All	1.84	2309	781
0–4 GeV	1500–4500	0.78	1864	280
4–8 GeV	1500–4500	1.68	8733	1170
8–12 GeV	1500–4500	1.64	2309	407
			90 GeV e^+e^- events	
0–4 GeV	All	1.02	593	1880
4–8 GeV	All	6.99	5986	3777
8–12 GeV	All	3.22	4518	713
0–4 GeV	over 2500	0.72	593	85
4–8 GeV	over 2500	1.61	5986	517
8–12 GeV	over 2500	1.41	4518	381

[a] M is the number of jets in the e^+e^- histogram.

[b] N is the number of jets in the $p\bar{p}$ histogram.

[c] All indicates that the entire range of $\hat{s}$ is included. For $p\bar{p}$ this is $400 \text{ GeV}^2 < \hat{s} < 30\,410 \text{ GeV}^2$.

TABLE VI. Values of the KS statistic for comparison between jets from $e^+e^- \rightarrow 3$ jets at 60 GeV and jets from $p\bar{p} \rightarrow 2$ jets. The second column refers to the range of the subprocess center-of-mass energy considered.

Jet mass	$\hat{s}$	D_{MN}	M[a]	N[b]
			Quark jets	
0–4 GeV	All[c]	3.32	5226	6048
4–8 GeV	All	4.89	11571	6560
8–12 GeV	All	2.03	1197	781
0–4 GeV	1000–4900	1.10	797	995
4–8 GeV	1000–4900	0.60	1335	2576
8–12 GeV	1000–4900	1.04	52	543
			Gluon jets	
0–4 GeV	All	3.63	6286	19792
4–8 GeV	All	3.45	8717	31052
8–12 GeV	All	1.10	922	3729
0–4 GeV	1000–4900	1.57	137	1752
4–8 GeV	1000–4900	1.37	1550	9454
8–12 GeV	1000–4900	0.69	309	2733

[a] M is the number of e^+e^- jets in the histogram.

[b] N is the number of $p\bar{p}$ jets in the histogram.

[c] All indicates that the entire range of $\hat{s}$ is included. For $e^+e^- \rightarrow 3$ jets at 60 GeV this is $200 \text{ GeV}^2 < \hat{s} < 2700 \text{ GeV}^2$.

TABLE VII. Values of the KS statistic for comparison between jets from $e^+e^- \to 3$ jets at 90 GeV and jets from $p\bar{p} \to 2$ jets. The second column refers to the range of the subprocess center-of-mass energy considered.

Jet mass	$\hat{s}$	D_{MN}	M[a]	N[b]
			Quark jets	
0–4 GeV	All[c]	0.66	7306	1880
4–8 GeV	All	5.57	40326	3777
8–12 GeV	All	1.65	17218	713
0–4 GeV	1600–2500	1.00	2396	163
4–8 GeV	1600–2500	1.42	12828	579
8–12 GeV	1600–2500	1.37	5139	133
0–4 GeV	2500–4000	0.65	684	67
4–8 GeV	2500–4000	0.79	3325	397
8–12 GeV	2500–4000	1.22	852	222
			Gluon jets	
0–4 GeV	All	2.84	6742	4355
4–8 GeV	All	1.28	32454	15030
8–12 GeV	All	2.98	14033	3299
0–4 GeV	1600–2500	0.77	555	169
4–8 GeV	1600–2500	1.43	9975	1972
8–12 GeV	1600–2500	1.65	6071	724
0–4 GeV	2500–4000	1.32	19	32
4–8 GeV	2500–4000	0.63	1193	964
8–12 GeV	2500–4000	0.81	2277	972

[a] M is the number of e^+e^- jets in the histogram.

[b] N is the number of $p\bar{p}$ jets in the histogram.

[c] All indicates that the entire range of $\hat{s}$ is included. For $e^+e^- \to$ at 90 GeV this is 600 GeV^2 $< \hat{s} <$ 6400 GeV^2.

TABLE VIII. Values of the KS statistic for comparison between quark jets and gluon jets from $e^+e^- \to 3$ jets produced with 90 GeV center-of-mass energy. The second column refers to the range of the subprocess center-of-mass energy considered in GeV^2.

Jet mass	$\hat{s}$	D_{MN}	M[a]	N[b]
0–4 GeV	All[c]	21.22	7306	6742
4–8 GeV	All	52.40	40326	32454
8–12 GeV	All	27.36	17218	14033
0–4 GeV	800–1600	17.61	4223	6168
4–8 GeV	800–1600	41.92	24166	21286
8–12 GeV	800–1600	19.93	11227	5682
0–4 GeV	1600–2500	7.89	2396	555
4–8 GeV	1600–2500	29.53	12828	9975
8–12 GeV	1600–2500	15.87	5139	6071
0–4 GeV	2500–4000	1.98	684	19
4–8 GeV	2500–4000	10.46	3325	1193
8–12 GeV	2500–4000	5.25	852	2277

[a] M is the number of quark jets in the histogram.

[b] N is the number of gluon jets in the histogram.

[c] All indicates that the entire range of $\hat{s}$ is included. For $e^+e^- \to 3$ jets at 90 GeV this is 800 GeV^2 $< \hat{s} <$ 6400 GeV^2.

TABLE IX. Values of the KS statistic for comparison between quark jets and gluon jets from $e^+e^- \to 3$ jets produced with 60 GeV center-of-mass energy. The second column refers to the range of the subprocess center-of-mass energy considered in GeV^2.

Jet mass	$\hat{s}$	D_{MN}	M[a]	N[b]
0–4 GeV	All[c]	16.23	5226	6286
4–8 GeV	All	23.78	11571	8717
8–12 GeV	All	4.83	1197	922
0–4 GeV	300–1100	15.77	4713	6226
4–8 GeV	300–1100	22.78	10747	7774
8–12 GeV	300–1100	4.13	1167	680
0–4 GeV	1100–4900	3.73	513	60
4–8 GeV	1100–4900	6.30	824	943
8–12 GeV	1100–4900	1.25	30	242

[a] M is the number of quark jets in the histogram.

[b] N is the number of gluon jets in the histogram.

[c] All indicates that the entire range of $\hat{s}$ is included. For $e^+e^- \to 3$ jets at 60 GeV this is 200 GeV$^2 < \hat{s} < 2700$ GeV^2.

$$n^h_{pj} = \sum_{i=1}^{N} \omega^h_{ji} x_{pi} + \theta^h_j,$$

where ω^h_{ij} is the weight on the connection between the ith input unit and θ^h_j is a bias term, which is included to help the weights converge to an acceptable solution. The superscript h refers to the hidden layer. The output from the jth neuron in the hidden layer can be written as some function of its input:

$$i_{pj} = f^h_j(n^h_{pj}).$$

Then, for the kth output neuron,

$$n^o_{pk} = \sum_{i=1}^{L} \omega^o_{kj} i_{pj} + \theta^o_k,$$

$$o_{pk} = f^o_k(n^o_{pk}),$$

where there are L hidden layer neurons.

The function f^o_j is chosen according to the form of the desired output. In the case of quark-gluon jet identification, a binary response is convenient, 1 for a quark or antiquark jet and 0 for a gluon jet. Since f^o_j must be differentiable, a suitable choice is a sigmoid function, $f^o_j(n^o_{jk}) = 1/[1 + \exp(-2n^o_{jk})]$, because it limits the out-

TABLE X. Values of the KS statistic for comparison between quark jets and gluon jets from $p\bar{p} \to 2$ jets. The second column refers to the range of the subprocess center-of-mass energy considered in GeV^2.

Jet mass	$\hat{s}$	D_{MN}	M[a]	N[b]
0–4 GeV	All[c]	11.07	1880	4355
4–8 GeV	All	16.28	3777	15030
8–12 GeV	All	7.58	713	3299
0–4 GeV	800–1600	10.11	1632	4151
4–8 GeV	800–1600	12.83	2627	11895
8–12 GeV	800–1600	3.97	199	1098
0–4 GeV	1600–2500	4.06	163	169
4–8 GeV	1600–2500	7.18	579	1972
8–12 GeV	1600–2500	3.09	133	724
0–4 GeV	2500–4000	0.91	67	32
4–8 GeV	2500–4000	5.34	397	964
8–12 GeV	2500–4000	3.97	222	972

[a] M is the number of quark jets in the histogram.

[b] N is the number of gluon jets in the histogram.

[c] All indicates that the entire range of $\hat{s}$ is included. For $p\bar{p}$ this is 400 GeV$^2 < \hat{s} < 30\,410$ GeV^2.

put to the range [0,1].

Define the error at the kth output neuron to be $\delta_{pk} = (y_{pk} - o_{pk})$ where the p refers to the pth training set, the y_{pk} is the desired output, and o_{pk} is the output from the kth neuron. The generalized delta rule minimizes the sum of the squares of the errors for all output neurons. If there are M output neurons, then the function to minimize is

$$E_p = \frac{1}{2}\sum_{k=1}^{M}(y_{pk} - o_{pk})^2.$$

The method of steepest descent can be used to determine in which direction the weight changes should occur. In weight space, E_p can be thought of as a surface, and $-\nabla E_p$ is the direction of steepest descent. Iterative changes of the weights are made until E_p reaches a minimum.

Differentiation of E_p gives

$$\frac{\partial E_p}{\partial \omega^o_{kj}} = -(y_{pk} - o_{pk})\frac{\partial f^o_k}{\partial n^o_{pk}}\frac{\partial n^o_{pk}}{\partial \omega^o_{kj}},$$

where

$$\frac{\partial n^o_{pk}}{\partial \omega^o_{kj}} = \left(\frac{\partial}{\partial \omega^o_{kj}}\sum_{j=1}^{L}\omega^o_{kj} i_{pj} + \theta^o_k\right) = i_{pj}.$$

This gives

$$-\frac{\partial E_p}{\partial \omega^o_{kj}} = (y_{pk} - o_{pk})\frac{\partial f^o_k}{\partial n^o_{pk}} i_{pj}.$$

The weights on the output layer are updated according to

$$\omega^o_{kj}(t+1) = \omega^o_{kj}(t) + \Delta_p \omega^o_{kj}(t),$$

where

$$\Delta_p \omega^o_{kj}(t) = \eta(y_{pk} - o_{pk})\frac{\partial f^o_k}{\partial n^o_{pk}} i_{pj},$$

where η is the learning rate parameter. The learning rate parameter is usually chosen to be a small number to ensure that the network will converge towards a solution.

Similarly, the hidden layer weights can be calculated according to [9]

$$\omega^h_{kj}(t+1) = \omega^h_{kj}(t) + \Delta_p \omega^h_{kj}(t),$$

where

$$\Delta_p \omega^h_{ji}(t) = \eta\frac{\partial f^h_j}{\partial n^h_{pj}} x_{pi}\sum_k (y_{pk} - o_{pk})\frac{\partial f^o_k}{\partial n^o_{pk}}\omega^o_{kj}.$$

The major limitation of back propagation is that there is no way to know the optimal number of hidden layers and the number of neurons in each hidden layer except by trial and error. The network may not converge if there are too few neurons. Too many neurons can cause the network to be slow and waste CPU time. Automatic methods of pruning the network have been developed by, for instance, Rumelhart *et al.* [9].

In JETNET, weight decay is included as a way to identify unnecessary nodes by allowing the connection weights to those nodes to decay to zero [8]. The updating equation with weight decay is then

$$\Delta\omega_{kj} = -\eta\frac{\partial E_p}{\partial \omega_{kj}} - \epsilon\omega_{kj},$$

where we used $\epsilon = 0.0001$.

Another problem with multiple-layer networks is that it can be difficult to interpret the internal representation that is built up by the training; the physical significance of the results is not always obvious. This can be an advantage of self-organizing networks [2], which do not use hidden neurons.

Our experimentation with self-organizing networks did not yield significantly better results than with back propagation. As we will demonstrate below, this is almost certainly because of the physical properties of jet data. These are such that our back-propagation results *are* readily interpretable.

VI. LINEAR DISCRIMINATOR

The case of a feed-forward network with no hidden layers is a linear discriminant function. A linear discriminant function is a function that partitions feature space with a hyperplane as the decision surface [10]. The type of problems that can be classified with such a surface is said to be linearly separable.

Suppose we have a feed-forward back-propagation network with no hidden layers. Then, the input to the last neuron is

$$g(\mathbf{x_p}) = \omega \cdot \mathbf{x_p} + \theta,$$

where $\mathbf{x_p} = (x_1, x_2, ..., x_N)$ is the pth feature vector, $\omega = (\omega_1, \omega_2, ..., \omega_N)$ is the weight vector, and θ is the threshold weight. Then a two-category linear classifier with classes α_1 and α_2 implements the rule [11]

$$\text{decide } \alpha_1 \text{ if } g(\mathbf{x}) > 0,$$

$$\text{decide } \alpha_2 \text{ if } g(\mathbf{x}) < 0.$$

The network reports the result $f(g(x)) = 1/\{1 + \exp[-2g(x)]\}$ which gives an alternate implementation of the rule

$$\text{decide } \alpha_1 \text{ if } f(g(\mathbf{x})) > \tfrac{1}{2},$$

$$\text{decide } \alpha_2 \text{ if } f(g(\mathbf{x})) < \tfrac{1}{2}.$$

The equation $g(\mathbf{x}) = 0$ defines a surface that separates points assigned to α_1 from points assigned to α_2.

VII. DISCRIMINATION USING SIMPLE CUTS

Before discussing our use of neural networks for quark-gluon jet discrimination, it is natural to ask what our ability to separate is by making well chosen "cuts" in the input data.

To answer this question, consider the binned data of Fig. 4. Ideally, one would choose a value of the jet variable which would optimize separation, the quark sample being to the left of the cut (a low bin number means

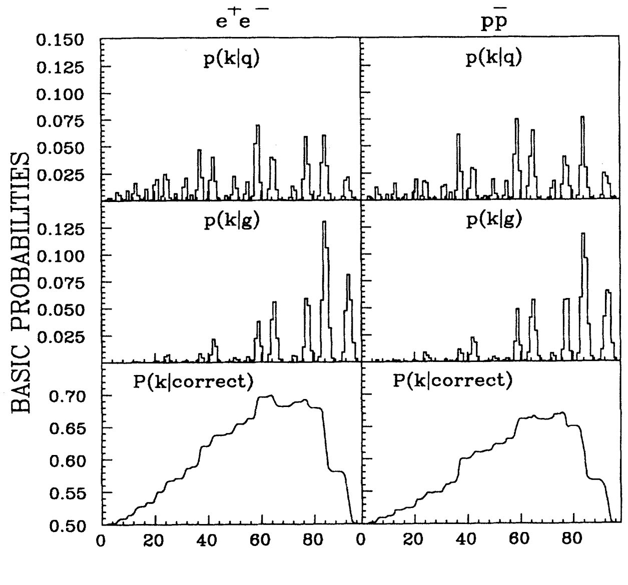

FIG. 4. The probability densities for quark jets, $p(k|q)$, and gluon jets, $p(k|g)$, and the sum of the integrated probability densities, $P(K|\text{correct}) = [P(k|q) + P(k|g)]/2$, for jets produced in the reactions $e^+e^- \to 3$ jets and $p\bar{p} \to 2$ jets. The jets are restricted to those with 4 GeV $\leq M \leq$ 8 GeV and 1600 GeV^2 $\leq \hat{s} \leq$ 2500 GeV^2. The feature space consists of the correlation bin for the jet. $P(K|\text{correct})$ is the probability of making a correct classification of jet species by cutting at bin k.

there is a leading hadron in the jet) and the gluon sample to the right of the cut (high bin number). Consider quark and gluon jet samples with jet masses in the range $4 \leq M \leq 8$ and hard-process center-of-mass energies in the range 1600 GeV^2 $\leq \hat{s} \leq$ 2500 GeV^2. The jets are produced in the reactions $e^+e^- \to 3$ jets and $p\bar{p} \to 2$ jets. To determine this cut, we defined the integrated probability for quark jets, $P(k|q) = \sum_{i=1}^{k} p(i|q)$, where $p(i|q)$ is the probability density for a quark jet to have the jet variable i. This gives the fraction of quarks jets that falls to the left of the cut. Similarly, for the gluon jets, $P(k|g) = 1 - \sum_{i=1}^{k} p(i|g)$, which is the fraction of gluon jets that falls to the right of the cut, where $p(i|g)$ is the basic probability for a gluon jet to have the jet variable i. It follows that for a cut made at bin number k the fraction of jets correctly classified is $P(k|\text{correct}) = \frac{1}{2}[P(k|g) + P(k|q)]$ [11]. See Fig. 4.

For $e^+e^- \to 3$ jets, the bin number that maximizes $P(k|\text{correct})$ is $k = 77$. At this cut, $P(k|\text{correct}) = 0.691$ where 76.9% of the quark jets occupy bins to the left of k and 61.2% of the gluon jets occupy the bins to the right of k. However, one might prefer a cut that gives similar classification results for both quark and gluon jets. If the cut is made at $k = 73$, $P(k|\text{correct}) = 0.688$ with 69.0% of the quark jets occupying bins to the left of the cut and 68.6% of the gluon jets occupying bins to the right of the cut.

Similarly, for $p\bar{p} \to 2$ jets, a cut of $k = 76$ gives $P(k|\text{correct}) = 0.670$ with 69.9% of the quark jets and 63.9% of the gluon jets correctly classified (Fig. 4). A cut of $k = 72$ gives $P(k|\text{correct}) = 0.662$ with 66.8% of quark jets occupying bins to the left of the cut and 65.6% of gluon jets to the right (Fig. 4).

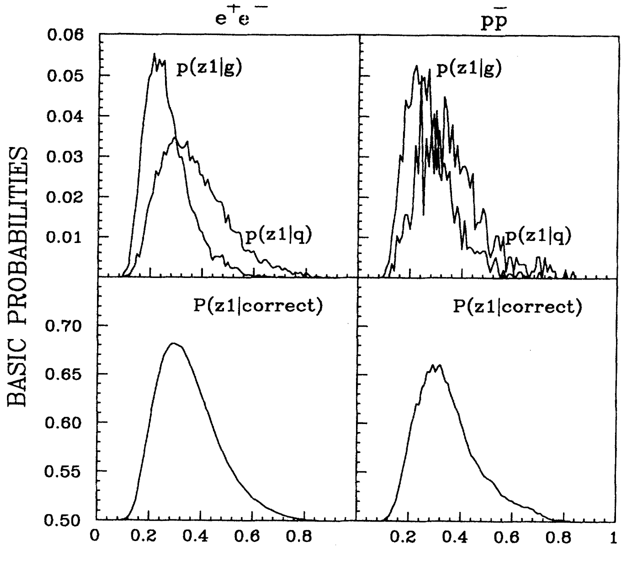

FIG. 5. The probability densities for quark jets, $p(z_1|q)$, and gluon jets, $p(z_1|g)$, and the sum of the integrated probability densities, $p(z_1|\text{correct}) = P(z_1|q) + P(z_1|g)$, for jets produced in the reactions $e^+e^- \to 3$ jets and $p\bar{p} \to 2$ jets. The jets are restricted to those with 4 GeV $\leq M \leq$ 8 GeV and 1600 GeV^2 $\leq \hat{s} \leq$ 2500 GeV^2. The feature space consists of the invariant longitudinal momentum fraction z_1 of the leading hadron in the jet.

The same test was done considering only the invariant longitudinal momentum z_1 of the leading hadron in the jet. This time, we expect quark jets to have higher invariant longitudinal momentum fractions z_1, and so we define our cut so that the quark sample is to the right of the cut and the gluon sample is to the left of the cut. When the cut is drawn at $z_1 = 0.29$, for the e^+e^- jets, $P(z_1|\text{correct}) = 0.681$ with 69.8% of the gluon jets having z_1 less than this cut and 63.1% of the quark jets having z_1 greater than this cut. For the $p\bar{p}$ jets, the same cut leads to $P(z_1|\text{correct}) = 0.660$ with 66.4% of the gluon jets having z_1 less than this cut and 65.6% of the quark jets having z_1 larger than the cut (Fig. 5).

Note that cutting in this way always leads to a higher percentage of correct classification in e^+e^- than in $p\bar{p}$. We will find similar results below in our neural network study of jets from the two cases: The jets from $p\bar{p}$ are always slightly more difficult to discriminate.

VIII. RESULTS OF THE NEURAL NETWORK

The most crucial aspect of the use of an artificial neural network is the choice of input parameters. Our goal is to train a network on data from $e^+e^- \to$ jets and then use this network to identify jets from $p\bar{p} \to$ jets. Since (for our sample) 68% of the time, the lowest-energy jet in an e^+e^- 3 jet event is the gluon jet, jet energy might be a choice as input parameter. However, we want to build up a network by training on single jets from e^+e^- that will identify jets from $p\bar{p}$ interactions, where there is not a set number of gluon jets. We have chosen to train the network on kinematic properties of the *jet* rather than use any properties of the *event.* We have restricted the input parameters to be Lorentz-invariant quantities.

Since we know from our studies with the Kolmogorov-Smirnov statistic that the distribution in jet variable (bin number) depends on the mass of the jet as well as the center-of-mass energy of the hard subprocess, we will focus on our attention on quark and gluon jets that fall within the mass range 4 GeV $\leq M \leq$ 8 GeV and within the hard subprocess center-of-mass energy range 1600 GeV^2 $\leq \hat{s} \leq$ 2500 GeV^2. For our three-jet events, we have used the definition of $\hat{s}$ that is described in Sec. IV.

A. No hidden layer

Our input parameter set consists of the invariant longitudinal momentum fractions of the nine leading hadrons in the jet, $\mathbf{x_p} = (z_1, z_2, ..., z_N)$. We begin with the simplest configuration: an input layer of nodes and a single output node, Fig. 6.

The network is trained on 10 000 jets from the e^+e^- reaction, using equal numbers of quark and gluon jets. The jets are introduced randomly to the network during the training process so that any given jet is as likely to be a gluon jet as it is to be a quark jet. The network is tested on a sample of 5000 jets from $e^+e^- \rightarrow 3$ jets as well as a sample of 2500 $p\bar{p} \rightarrow 2$ jets. The jets in the testing sample meet the same mass and hard-process center-of-mass restrictions as the training set.

As discussed in Sec. VI, this configuration is a linear discriminant function. We investigate the effect of varying some of the parameters, namely, the learning parameter η and the weight decay parameter ϵ, during training by using nine inputs to the network. The cases we consider here are (i) *simple case* $\eta = 0.01$, $\epsilon = 0$, (ii) *weight decay* $\eta = 0.01$, $\epsilon = 0.0001$, and (iii) *decreasing learning parameter* $\epsilon = 0$, $\eta = 0.1$ to start, but gradually reduced by a factor of 0.1 over each 100 training epochs. The results are reported in Tables XI, XII, and XIII.

Figure 7 shows the performance of the network with nine inputs during the training. The histogrammed output from the discriminator is shown in Fig. 8 and the output of the discriminator for a given jet versus the bin number of the jet is shown for quark and gluon jets in Fig. 9. This plot shows us that, despite the fact that the number of correctly identified jets is roughly similar, the discriminant function is not making a "simple cut" on bin number such as we did in Sec. VII. The projection onto one dimension that we get from the binning technique is not equivalent to the projection that the discriminant function is making. The linear discriminant function computes the projection of the feature vector along some weight vector ω in multidimensional space, Fig. 8. The network then uses a sigmoid function $f(\omega \cdot x + \theta)$ to push the output towards a binary answer.

After 1000 training cycles the network with nine input nodes produced the following weight vectors for the three cases listed above.

(i) *Simple case:* $\omega = (5.23, 5.38, 1.85, 2.12, 1.52, 1.42, 0.592, -0.254, -0.498)$.
(ii) *Weight decay:* $\omega = (3.30, 2.48, 0.832, 0.657, 0.297, 0.0495, 0.00768, 0.0, 0.0)$.
(iii) *Decreasing* η: $\omega = (5.11, 5.13, 1.79, 2.13, 1.13, 0.0680, 0.868, -0.463, -0.517)$.

Inspection of the weight vector verifies our expectation that the leading hadrons in the jet are the crucial factor in determining which are quark and which are gluon jets. In order to investigate the decrease in performance as the number of inputs (and therefore hadrons in the jet) is varied, we repeated the training of our linear discriminant function for the three cases above, varying the number of inputs. The results are given in Tables XI, XII, and XIII. We see that the bulk of the information is supplied by the two fastest particles in the jet.

B. Multilayer network

Now we ask the following: If we add a hidden layer to the network, is it able to do better than the linearly separable solution? We consider the three layer network configuration: nine nodes on the input layer, one node on the output layer, and a variable number of nodes on the hidden layer [12].

We consider the three cases described above, that is, (i) *the simple case* $\eta = 0.01$, $\epsilon = 0$, (ii) *weight decay* $\eta = 0.01$, $\epsilon = 0.0001$, and (iii) *decreasing learning parameter* $\epsilon = 0$, $\eta = 0.1$, to start, but gradually reduced by a factor of 0.1 over each 100 training epochs. Tables XIV, XV, and XVI, show the performance of the network as the number of hidden layer nodes is varied.

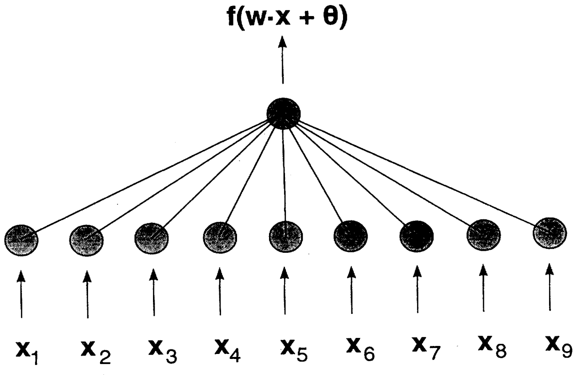

FIG. 6. A two-layer feed-forward network.

TABLE XI. Performance of the two-layer back propagation neural network for varying number of inputs. All jets have 4 GeV $\leq M \leq$ 8 GeV and 1600 GeV^2 $\leq \hat{s} \leq$ 2500 GeV^2.

Input layer	e^+e^-		$p\bar{p}$	
nodes	Quark (%)	Gluon (%)	Quark (%)	Gluon (%)
	(i) *Simple case:* $\eta = 0.01, \epsilon = 0.0$			
9	71.7	71.7	65.7	70.0
8	71.3	71.5	66.4	69.8
7	72.6	71.5	66.4	69.5
6	72.2	70.9	66.9	69.0
5	74.2	69.4	69.0	67.2
4	70.0	73.4	64.8	70.9
3	71.6	71.2	65.7	69.6
2	69.7	72.1	65.5	70.0

TABLE XII. Performance of the two-layer back propagation neural network for varying number of inputs. All jets have 4 GeV $\leq M \leq$ 8 GeV and 1600 GeV^2 $\leq \hat{s} \leq$ 2500 GeV^2.

Input layer	e^+e^-		$p\bar{p}$	
nodes	Quark (%)	Gluon (%)	Quark (%)	Gluon (%)
	(ii) *Weight decay:* $\eta = 0.01, \epsilon = 0.0001$			
9	73.2	69.8	67.6	67.6
8	70.1	73.2	65.3	70.8
7	71.1	71.0	66.7	69.0
6	72.5	70.5	67.1	68.7
5	69.1	74.3	63.9	72.2
4	70.8	71.5	65.8	69.5
3	70.3	72.5	65.0	69.4
2	68.3	72.9	64.6	69.8

TABLE XIII. Performance of the two-layer back propagation neural network for varying number of inputs. All jets have 4 GeV $\leq M \leq$ 8 GeV and 1600 GeV^2 $\leq \hat{s} \leq$ 2500 GeV^2.

Input layer	e^+e^-		$p\bar{p}$	
nodes	Quark (%)	Gluon (%)	Quark (%)	Gluon (%)
	(iii) *Decreasing learning parameter:* $\eta = 0.01$,[a] $\epsilon = 0.0$			
9	72.5	71.8	65.8	69.9
8	72.1	71.5	66.0	69.7
7	71.8	71.6	66.0	69.8
6	71.8	71.5	66.0	69.9
5	71.3	71.7	66.0	69.8
4	72.5	71.9	65.8	69.7
3	71.4	71.3	65.7	69.8
2	69.8	72.1	65.5	70.0

[a] Decreased by a factor of 0.1 in 100 cycles.

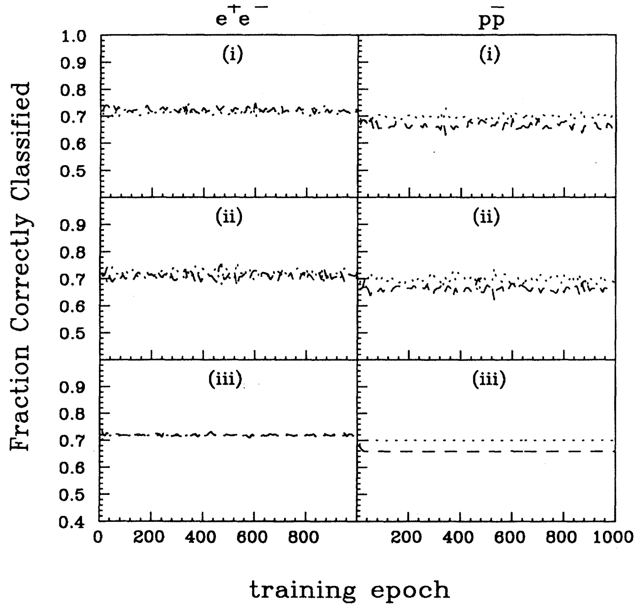

FIG. 7. A plot of performance during training for a two-layer feed-forward network. The network was trained on e^+e^- jets and tested on both e^+e^- jets and $p\bar{p}$ jets every tenth training epoch. (i) *Simple case,* $\eta = 0.01$; (ii) *weight decay,* $\eta = 0.01$, $\epsilon = 0.0001$; and (iii) *decreasing learning parameter,* $\eta = 0.1$, and it is gradually reduced by a factor of 0.1 over 100 training epochs. Gluon jets (dotted lines) and quark jets (dashed lines).

We find that the performance of the network is insensitive to the number of nodes in the hidden layer. Further, it appears that the three-layer network is unable to achieve better quark-gluon separation than the simple linear discriminant function (compare Fig. 10 with Fig. 7).

Inspection of the connection weights from the network configurations (i) and (iii) does not shed any light on the interpretation of the network (Fig. 11 and Table XVII).

It is only by introducing weight decay that we began to see systematic and reproducible weights resulting from the training.

When weight decay was implemented, symmetries in the weights became apparent. The use of a weight decay term pushes a connection weight to its lowest acceptable value. Presumably, there are many solutions to the problem, and weight decay selects a unique solution in weight space (we consider all permutations of hidden layer nodes to be the same solution as long as the connection weights for the nodes are also likewise permuted). Figure 12 and Table XVIII show the connection weights from the input layer to the hidden layer for a network with ten nodes in the hidden layer; some other cases are given in Tables XIX and XX.

Consider, for example, the network with two hidden layer nodes. We find that, when weight decay is implemented, the connection weights from the inputs to hidden layer node 1 are approximately equal in magnitude and opposite in sign to the connection weights from the input layer to hidden layer node 2 (Tables XX). The thresholds and the connection weights from the hidden layer nodes to the output layer node also display this feature. We plot in Fig. 13 the outputs from the hidden layer nodes for the quarks and gluons. As expected, each node gives us output distributions like that for the two-layer case, but with a different offset θ. When we increase the number of hidden layer nodes, we still observe the network following this trend, that is, producing connection weights and thresholds such that there is a "positive" node and a "negative" node and all other nodes are repeats of these two basic types of nodes. The results in Table XVIII and Fig. 12 show this trend to a remarkable degree.

This can be understood geometrically. The output from each node forms a hyperplane decision boundary in the space of the inputs (i.e., the space of the invariant longitudinal momentum fractions) [10]. Two nodes that have connection weights and threshold that are equal in magnitude and opposite in sign will intersect the input space with the same hyperplane decision boundary. Therefore, the two nodes in the hidden layer are produc-

TABLE XIV. Performance of the three-layer back propagation neural network for a varying number of nodes in the hidden layer. None refers to the network with no hidden layer. All jets have 4 GeV $\leq M \leq$ 8 GeV and 1600 GeV^2 $\leq \hat{s} \leq$ 2500 GeV^2.

Hidden layer	e^+e^-		$p\bar{p}$	
nodes	Quark (%)	Gluon (%)	Quark (%)	Gluon (%)
		(i) *Simple case:* $\eta = 0.01, \epsilon = 0.0$		
30	70.7	73.5	64.6	71.4
15	72.3	73.0	65.3	71.0
10	69.1	76.8	62.7	73.9
9	71.8	72.9	65.7	70.9
8	71.4	72.9	65.0	71.0
7	71.0	74.8	63.6	72.3
6	71.2	73.3	64.6	71.6
5	72.8	72.2	66.4	70.3
4	69.5	76.6	62.9	73.5
3	70.7	75.3	63.6	73.1
2	69.4	75.0	62.7	73.1

TABLE XV. Performance of the three-layer back propagation neural network for a varying number of nodes in the hidden layer. None refers to the network with no hidden layer. All jets have 4 GeV $\leq M \leq$ 8 GeV and 1600 GeV^2 $\leq \hat{s} \leq$ 2500 GeV^2.

Hidden layer	e^+e^-		$p\bar{p}$	
nodes	Quark (%)	Gluon (%)	Quark (%)	Gluon (%)
	(ii) *Weight decay:* $\eta = 0.01,^{a}$ $\epsilon = 0.0001$			
30	73.7	70.3	64.1	72.1
15	72.7	70.6	67.6	68.5
10	70.4	72.9	65.3	71.6
9	75.6	66.5	71.0	64.0
8	71.4	72.7	65.8	70.7
7	77.2	67.0	60.2	74.8
6	73.2	70.2	67.8	68.3
5	72.9	70.8	67.4	68.8
4	69.5	74.4	62.9	72.9
3	69.6	76.2	63.0	73.6
2	66.1	79.0	58.6	76.4

[a] Decreased by a factor of 0.1 in 100 cycles.

TABLE XVI. Performance of the three-layer back-propagation neural network for a varying number of nodes in the hidden layer. All jets have 4 GeV $\leq M \leq$ 8 GeV and 1600 GeV^2 $\leq \hat{s} \leq$ 2500 GeV^2.

Hidden layer	e^+e^-		$p\bar{p}$	
nodes	Quark (%)	Gluon (%)	Quark (%)	Gluon (%)
	(iii) *Decreasing learning parameter:* $\eta = 0.01,^{a}$ $\epsilon = 0.0$			
30	71.9	71.8	65.7	69.8
15	71.8	72.7	65.1	70.5
10	71.4	73.0	64.6	71.0
9	70.4	73.3	64.4	71.3
8	70.9	72.5	65.1	70.5
7	70.0	74.1	63.4	72.6
6	71.3	72.8	64.8	71.0
5	69.6	74.0	63.2	72.7
4	69.5	74.7	62.7	73.0
3	70.1	74.1	63.4	72.1
2	69.6	74.5	62.9	73.2

[a] Decreased by a factor of 0.1 in 100 cycles.

TABLE XVII. Values of the weights for the simple network [case (i)] with no weight decay and ten nodes in the hidden layer. The data in this table are represented graphically in Fig. 11.

Hidden	1	2	3	4	5	6	7	8	9
1	3.94	2.00	-2.27	-2.10	0.0303	-0.578	0.0695	-0.490	-0.818
2	2.38	2.71	7.37	4.61	0.340	0.854	1.02	0.726	0.178
3	-1.97	-0.703	-0.431	-0.888	-0.859	-0.985	-0.106	0.360	0.0604
4	-1.61	-0.634	-1.06	0.283	-0.814	-0.389	-0.0456	0.255	0.780
5	-0.764	-0.909	-0.138	-0.489	-0.417	0.554	-0.933	0.813	-0.752
6	-2.58	-1.72	-1.37	0.459	-1.08	-0.962	0.672	-0.0404	0.349
7	3.09	6.07	0.354	-0.496	1.32	1.31	0.0820	-0.0865	0.524
8	0.655	-1.12	-0.729	-0.428	0.899	-0.0275	0.940	-0.512	-0.685
9	1.27	1.03	-0.560	1.06	1.03	0.0202	-0.636	-0.584	-0.406
10	-0.103	0.292	-0.670	-0.752	0.931	-0.609	-0.704	0.161	0.810

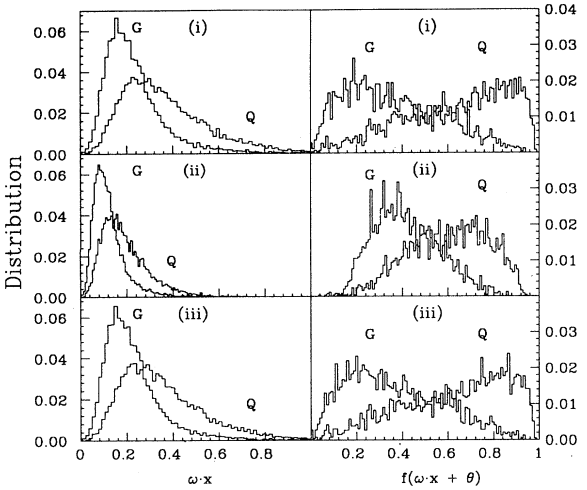

FIG. 8. Plots of the projection of the feature vector along the weight vector and plots of output $f(\omega \cdot \mathbf{x} + \theta)$ from the two-layer network, for testing on the same three cases as in Fig. 7 (e^+e^- events were used). The left peaks are the output for patterns we know to be gluon jets and the right peaks are for patterns we know to be quark jets. Those jets that fall to the left of $f(\omega \cdot \mathbf{x} + \theta) = 0.5$ are classified as gluon jets and those that fall to the right are classified as quark jets.

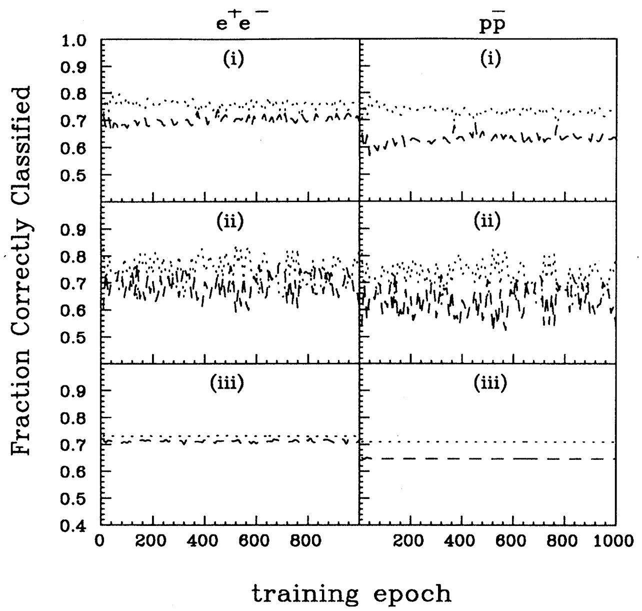

FIG. 10. A plot of performance during training for a three-layer feed-forward network with ten nodes in the hidden layer. The network was trained on e^+e^- jets and tested on both e^+e^- jets and $p\bar{p}$ jets every tenth training epoch. (i) *Simple case*, $\eta = 0.01$; (ii) *weights decay*, $\eta = 0.01$, $\epsilon = 0.0001$; and (iii) *decreasing learning parameter*, $\eta = 0.1$, and it is gradually reduced by a factor of 0.1 over 100 training epochs. Gluon jets (dotted lines) and quark jets (dashed lines).

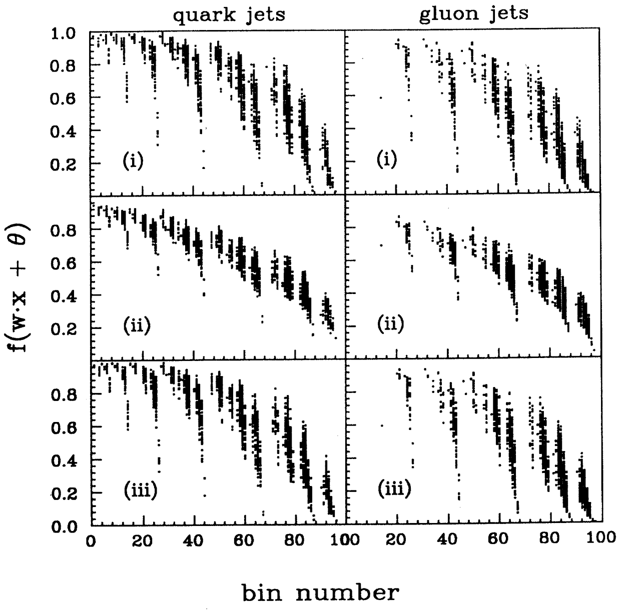

FIG. 9. Plot of the output $f(\omega \cdot \mathbf{x}+\theta)$ for a given jet versus the bin number of the jet, from the two-layer network, testing on e^+e^- jets. The three cases are (i) *simple case*, $\eta = 0.01$; (ii) *weight decay*, $\eta = 0.01$, $\epsilon = 0.0001$; and (iii) *decreasing learning parameter:* $\eta = 0.1$, and it is gradually reduced by a factor of 0.1 over 100 training epochs. On the left column are plots for quark jets and in the right column are plots for gluon jets.

ing almost exactly the same decision boundary.

Occasionally, we observe a network that, during training, develops one node that does not take on connection weights that are approximately equal to the "positive" or the "negative node." The output from this node appears to be shifted slightly from the other peaks.

Consider one such network that develops this third node. Table XIX gives the weights of a three-layer network with five hidden layer nodes. This network was trained with a learning parameter of $\eta = 0.01$ and a weight decay parameter of $\epsilon = 0.0001$ [a case (ii) network]. Nodes 2 and 4 form what we referred to as "positive" nodes and nodes 1 and 3 form the "negative" nodes, while node 5 does not correspond to either. However, notice that the connection weights to node 5 are proportional to the connection weights to node 4. In addition, the threshold for node 5 and the weight connecting node 5 to the output have the same proportionality to the corresponding quantities for node 4.

Geometrically, hidden layer node 5 produces a plane that intersects the input space with the same hyperplane as does node 4. Therefore, this node does not produce any refinement in the solution. In fact, the performance of this network is about the same as the networks that do not have this third type of node.

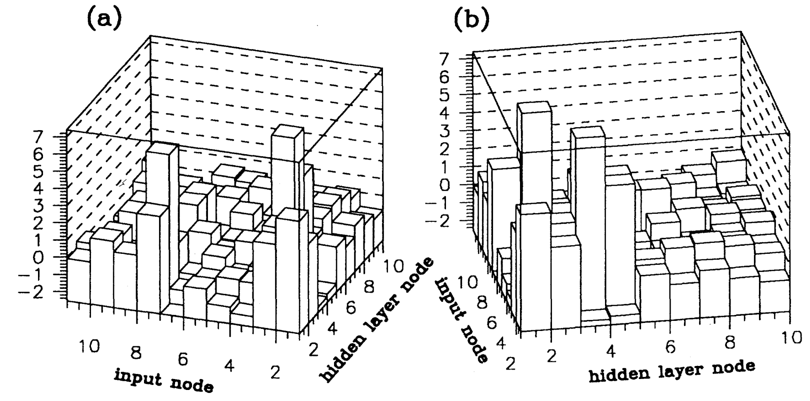

FIG. 11. A lego plot of the connection weights to the hidden layer for a three-layer network with ten nodes in the hidden layer and nine inputs (a) and (b) are the same plot from two views. These values, listed in Table XVII, were obtained for a simple network [case (i)] with no weight decay.

IX. CONCLUSIONS

We find that it is important to group jets according to jet mass and hard-process subenergy, prior to attempting statistical comparisons.

Once jets are classified according to these parameters, quark (gluon) jets from one reaction are statistically similar to quark (gluon) jets from another reaction. This allows use of the jets from $e^+e^- \rightarrow Z^0 \rightarrow 3$ jets in formulating strategies which can discriminate between quarks and gluons in other reactions.

Three forms of quark-gluon discrimination were tested: cutting on a single variable (the jet bin number or z_1), a two-layer neural network (linear discriminant function), and a three-layer neural network. Both networks were trained by the feed-forward back propagation approach.

Cutting on a single variable was only slightly less successful than the neural network approach. The multilayer network was not noticeably superior to the simpler two-layer networks.

These results are simultaneously discouraging and encouraging. Our "best" result, 73% correctly identified, is comparable to results which I. Csabai, F. Czakó, and Z. Fodor [Nucl. Phys. **B374**, 303 (1992)] found using similar information. The fact that only 73% of the jets can be correctly identified by using information intrinsic to the jet may prove to be a handicap in applications. Higher levels of discrimination will have to rely on knowledge of the matrix elements which might produce the jets. This is harder in $p\bar{p}$ than in e^+e^-.

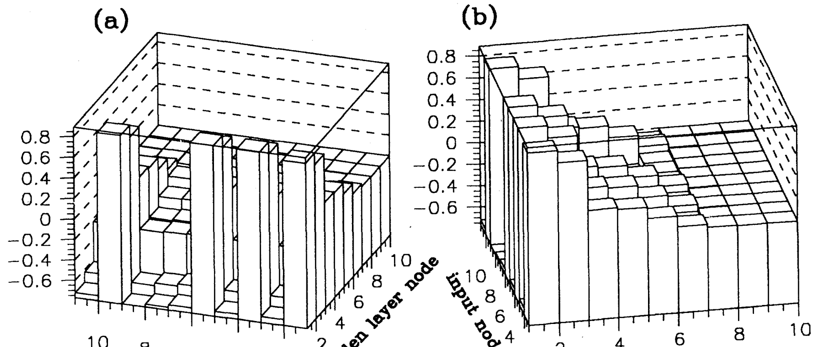

FIG. 12. A lego plot of the connection weights to the hidden layer for a three-layer network with ten nodes in the hidden layer and nine inputs. The network was trained with a weight decay parameter of $\epsilon = 0.0001$. (a) and (b) are two views of the same lego plot. Notice from (a) that there are four nodes that take on positive and approximately equal connection weights. In (b) we see that the connection weights from input i (corresponding to the longitudinal momentum fraction of hadron i) take on values according to the relative importance of the hadron in the jet. For instance, the magnitudes of the weights connecting node 1 (and therefore the leading hadron in the jet) to the hidden layer nodes is larger than the magnitudes of weights connecting all other inputs to the hidden layer nodes. Compare parts (a) and (b) from Fig. 11 with parts (a) and (b) from this figure to see the effect that weight decay produces.

TABLE XVIII. Values of the weights for the network with weight decay and ten nodes in the hidden layer. The data in this table are represented graphically in Fig. 12.

Hidden	1	2	3	4	5	6	7	8	9
1	0.838	0.726	0.258	0.241	0.133	0.0339	0.00362	0.0	0.0
2	-0.711	-0.611	-0.227	-0.211	-0.114	-0.0273	-0.00298	0.0	0.0
3	0.885	0.769	0.274	0.257	0.142	0.0360	0.00383	0.0	0.0
4	-0.720	-0.619	-0.230	-0.214	-0.116	-0.0277	-0.00302	0.0	0.0
5	0.889	0.772	0.275	0.258	0.142	0.0362	0.00384	0.0	0.0
6	-0.718	-0.617	-0.230	-0.215	-0.115	-0.0276	-0.00301	0.0	0.0
7	-0.741	-0.638	-0.237	-0.221	-0.119	-0.0287	-0.00312	0.0	0.0
8	-0.768	-0.663	-0.247	-0.230	-0.124	-0.0299	-0.00325	0.0	0.0
9	0.869	0.754	0.268	0.251	0.137	0.0353	0.00375	0.0	0.0
10	-0.721	-0.620	-0.231	-0.215	-0.116	-0.0278	-0.00303	0.0	0.0

TABLE XIX. Values of the weights for the network with weight decay and five nodes in the hidden layer.

Hidden	1	2	3	4	5	6	7	8	9
1	-1.35	-1.18	-0.439	-0.418	-0.227	-0.0530	-0.00616	0.0	0.0
2	1.05	0.899	0.319	0.299	0.165	0.0404	0.00480	0.0	0.0
3	-1.34	-1.17	-0.435	-0.414	0.225	-0.0525	-0.00612	0.0	0.0
4	1.01	0.867	0.306	0.288	0.159	0.0389	0.00464	0.0	0.0
5	0.611	0.515	0.177	0.165	0.0920	0.023	0.0028	0.0	0.0

TABLE XX. Values of the weights for the network with weight decay and two nodes in the hidden layer.

Hidden	1	2	3	4	5	6	7	8	9
				(i) *Simple case:* $\eta = 0.01, \epsilon = 0.0$					
1	5.52	6.12	1.61	2.13	1.38	0.175	-0.337	0.612	-0.101
2	-4.04	-3.51	-1.97	-1.79	-1.35	-0.680	-0.221	0.950	0.924
				(ii) *Weight decay:* $\eta = 0.01, \epsilon = 0.0001$					
1	1.74	1.50	0.525	0.494	0.283	0.0793	0.00772	0.0	0.0
2	-1.74	-1.50	-0.537	-0.511	-0.291	-0.0791	-0.00779	0.0	0.0
				(iii) *Decreasing learning parameter:* $\eta = 0.01,^{a} \epsilon = 0.0$					
1	0.700	0.859	0.757	0.234	0.0959	0.630	-0.658	0.638	-0.276
2	4.51	4.67	1.72	1.93	1.26	0.224	0.344	-0.564	0.398

[a] Decreased by a factor of 0.1 in 100 cycles.

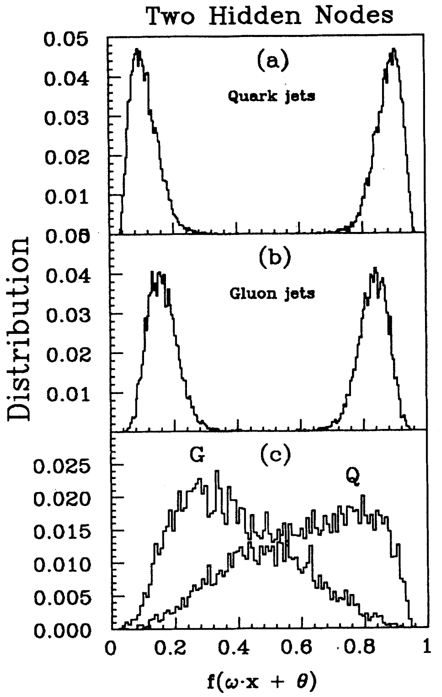

FIG. 13. (a) The output from the hidden layer nodes for quark jets and (b) the output from the hidden layer nodes for gluon jets, for a three-layer network with two units in the hidden layer. The rightmost peak corresponds to the output from one node and the left peak corresponds to the output from the other node. Each jet contributes to both peaks. (c) The output from the final layer of the network. It is the offset of the quark distribution from the gluon distribution that leads to two separate peaks at the final layer.

On the other hand, as physicists with a desire to "understand" the discrimination process, we are pleased about the following.

It is not necessary to have a large number of hidden nodes in a three-layer network to obtain good results. By use of weight decay, we have found that use of a large number of nodes in the hidden layer just leads to high degeneracy of the weights. All trials resulted in only two or three different weight patterns.

The three-layer network is not much better than the two-layer network.

The two-layer network amounts to simply projecting events on a single weight vector and using that to classify.

This is a simple cut in the space of longitudinal momenta of the particles.

The fitting of the two-layer network is thus a finding of the "best" cut on the space of z_i. It is a bit better than cutting on z_1 or on the correlation bin number.

Because there is a fairly simple cut which does the trick, we can have some confidence that enough data will exist to apply this technique to any desired slices of M and $\hat{s}$ (only ten parameters need to be fit in each slice). This should speed up the phenomenology considerably.

Finally, parameters trained on the e^+e^- jets worked slightly less well on the $p\bar{p}$ jets in all cases. We do not have a simple understanding of this phenomenon at present.

ACKNOWLEDGMENTS

We would like to thank L. Lönnblad for providing us with the JETNET program and B.R. Webber for the HERWIG program, as well as helpful discussions concerning its use. We also thank J. Beda for assistance in generating the $p\bar{p}$ events used in this paper. We wish to thank the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign for providing supercomputer time. This work was supported in part by the U.S. Department of Energy under Contract No. DOE FG02 91ER40677-Task P.

APPENDIX A: JET-FINDING ALGORITHM

All "jet-finding" calculations were carried out in the center-of-mass frame of the hard process. This ensures good physical separation of the jets in the two- and three-jet events studied here (the three-jet events all have thrust less than 0.9).

To find a jet, we began by choosing the most energetic hadron as the leading particle in the jet. For each jet, the jet axis was initially taken to be in the direction of the momentum of this hadron. The next most energetic hadron that fell within a cone of half-angle 41° of this initial axis was then assigned to the jet, and the jet axis was then taken to be in the direction of the sum of the momenta of these two hadrons. A third particle was then selected as the most energetic hadron among the particles yet to be associated with a jet that fell within a cone of half-angle 41° about the new jet axis. This procedure was repeated until all the hadrons were assigned to a jet.

Obviously, for two-jet events, this algorithm will yield essentially the same jets as other common jet finders (see Ref. [4]), since the two jets are back to back in the hard-process center of mass. For three-jet events, one might worry that the angle condition in our algorithm could result in inaccurate jet assignments of some of the hadrons originating from the two closest jets. Indeed, for general three-jet events, that is a valid concern and suggests that further work might use a different approach, such as that of the JADE Collaboration [4].

In the specific events studied here, however, we are confident that the jets found are correct because (a) the thrust cut on our three-jet sample resulted in a situation such that most jets were separated by more than 70° and all were separated by more than 66°, (b) the quark jets found in e^+e^- three-jet events are the same as those found in e^+e^- two jet events [5], and (c) as demonstrated in Tables V, VI, and VII, the jets thus found in $e^+e^- \to 3$ jets are identical in detail to those found in $p\bar{p} \to 2$ jets.

Because jets found in the two-jet events are relatively insensitive to the exact details of jet finding, and because the jets found in the three-jet events are identical to those found in the two-jet case, we conclude that the specific samples we are using for our jet-identification studies are appropriate.

APPENDIX B: STATISTICAL ANALYSIS

In order to compare the binned data, we chose to use the Kolmogorov-Smirnov two sample test [13]. To use this test, one assumes as the null hypothesis H_0 that the two sets of data are derived from the same distribution. In our case, the data depend on the discrete variable k, the bin number. If the null hypothesis is true, then the difference between the two sets of data at any value of k should be small. The Kolmogorov-Smirnov test uses a statistic based on this difference.

For each data set, a monotonically increasing function of bin number, $S_M(k)$, is defined as the "cumulative distribution function," a function which gives the fraction of data points to the left of a given bin number k:

$$S_M(k) = \sum_{i=1}^{k} \frac{m_i}{M}, \qquad 1 \leq k \leq k_{\max},$$

where m_i is the number of data points in the ith bin and M is the total number of data points in the set. $S_M(k)$ has end point values of $S_M(1) = 0$ and $S_M(k_{\max}) = 1$.

The statistic D_{MN} is defined as the maximum of the absolute value of all deviations:

$$D_{MN} = \sqrt{\frac{MN}{M+N}} \max_{1<k<k_{\max}} \mid S_M(k) - S_N(k) \mid,$$

where $S_M(k)$ and $S_N(k)$ correspond to two data sets with M and N total events.

If one compares many pairs of histograms generated from the same distribution, an empirical distribution for these values of D_{MN} can be computed. Then the significance level for disproving the hypothesis that the two data sets are derived from the same distribution function, for a given statistic, D_{MN}, is [14]

$$\text{prob}(D_{MN} > \text{observed}) = f(D_{MN}).$$

The rationale here is that a large value of D_{MN} is not very likely if the two histograms came from the same formula, and so one can use a "cutoff" value of D_{MN} beyond which the probability that both histograms came from the same formula is negligible.

For our purposes, the useful feature of this test is that the function $f(z)$ is independent of the actual formula used to generate the histograms. We tried pairs of histograms generated from uniform distributions, pairs of histograms generated from linear distributions, and pairs of histograms created by placing particles randomly in longitudinal phase space and projecting on the 97 bins of Table I; see [7]. A sample such curve is shown as the dashed line in Fig. 14.

When a large sample of histograms generated from different distributions is compared pairwise, the $f(z)$ curve shows no resemblance to the consistent set of curves obtained by comparing histograms from the same distribution. We conclude that indeed $f(z)$ is approximately a universal curve which we can usefully apply even in the case where we are comparing the rather peculiar projection of multiparticle phase space onto one dimension given by the bins of Table I.

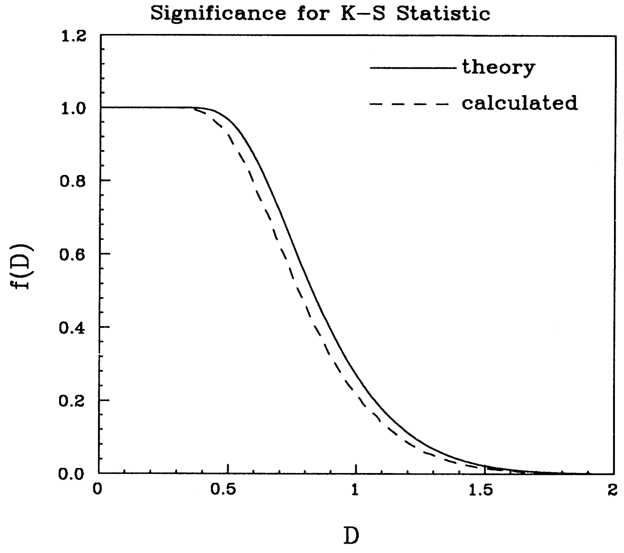

FIG. 14. Plot of statistical significance for the Kolmogorov-Smirnov statistic. Horizontal axis is the statistic (defined in Appendix B); vertical axis is the probability that, for two curves generated from the same formula, the KS statistic will be larger than the abscissa. Solid curve is generated from the "asymptotic formula" given in the literature; dashed curve is typical of our calculated curves using many different formulas, an average of 100 events/bin (in 100 bin histograms), and up to 10^6 pairs of histograms. Note that the probability of KS values larger than 2 arising from a pair of histograms generated from the same formula is extremely small.

In the literature it is claimed that (in the asymptotic limit where the number of data points becomes large) the distribution of the statistic for two data sets drawn from the same distribution can be calculated as

$$f(z) = 2\sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2z^2},$$

where the end point values are $f(0) = 1$ and $f(k_{\max}) = 0$. Figure 14 shows this "theoretical" curve compared with our experimental one. Our experimentally determined $f(z)$ curve is similar to this one, although it is systematically smaller at small values of z.

Press *et al.* [14] use this theoretical curve to recommend a value $f = 0.1$ ($D_{MN} = 1.2$) as a strong significance. Thus, if the significance is 0.1 or less, the hypothesis is disproved, and the data sets are not derived from the same distribution. We can be even more conservative and choose a cutoff of around 2 for D_{MN}. Values greater than 2 for the KS statistic almost never are achieved when the two distributions are generated from the same formula. Using the theoretical formula this amounts to a probability of 0.0007 (for our experimental curve the probability of a KS statistic of 2 or higher is 0.0003).

[1] L. Lönnblad *et al.*, Comput. Phys. Commun. **67**, 193 (1991); L. Lönnblad *et al.*, Nucl. Phys. **B349**, 675 (1991); L. Lönnblad *et al.*, Phys. Rev. Lett. **65**, 1321 (1990).

[2] I. Csabai *et al.*, Nucl. Phys. **B374**, 288 (1992); Z. Fodor, Phys. Lett. B **263**, 305 (1991); I. Csabai *et al.*, Phys. Rev. D **44**, 1905 (1991); Z. Fodor, *ibid.* **41**, 1726 (1990).

[3] G. Marchesini and B. R. Webber, HERWIG, version 5.3, 1990; G. Marchesini *et al.*, Comput. Phys. Commun. **67**, 465 (1992).

[4] S. Bethke *et al.*, Phys. Lett. B **213**, 235 (1988); S. Bethke *et al.*, Nucl. Phys. **B370**, 310 (1992); S. D. Ellis and D. E. Soper, Phys. Rev. D **48**, 3160 (1993); S. Catani *et al.*, Phys. Lett. B **269**, 432 (1991).

[5] M. A. Graham, Ph.D. thesis, University of Illinois, 1994.

[6] L. M. Jones, Phys. Rev. D **42**, 811 (1990).

[7] M. A. Graham, L. M. Jones, and P. R. Daumerie, Phys. Rev. D **46**, 222 (1992).

[8] Leif Lönnblad, JETNET, version 2.0, University of Lund, Lund, Sweden, 1991; L. Lönnblad, C. Peterson, and T. Rögnvaldsson, Comput. Phys. Commun. **70**, 167 (1992).

[9] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Parallel Distributed Processing* (MIT, Cambridge, MA, 1986), Vol. 1.

[10] R. P. Lippmann, IEEE ASSP Mag. **4** (2) (1987).

[11] R. Duda and P. Hart, *Pattern Classification and Scene Analysis* (Wiley, New York, 1973).

[12] Important parameter settings include the momentum parameter $\alpha = 0.5$ and width of initial weights $\omega = 1.0$.

[13] Richard von Mises, *Mathematical Theory of Probability and Statistics* (Academic, New York, 1964).

[14] William H. Press *et al.*, *Numerical Recipes: The Art of Scientific Computing* (Cambridge University Press, New York, 1987).