Supplement:
Algebraic Ground Truth Inference for Possibly Independent Binary Classifiers via Data Moments

Anonymous Authors[1]

This supplement proves theorems 1, 2 and 3 and provides more details on our experimental results. We start by providing a definition of all the sample statistics that we consider in the paper. The theorems are then proved along with some comments about our notion of classifier error independence. We follow with details on the exemplar experiments as well as some statistics on the root mean square error of repeated runs of our `two-norm` experiments. We close with additional comments about the relation of this work to that of Platanois *et al.* (Platanios et al., 2016), and Jaffe *et al.* (Jaffe et al., 2016).

This paper is about binary classifiers. But many of the algebraic concepts and equations we will discuss can be trivially extended to the multi-class case. They can also be extended beyond point statistics. One may be interested in sample sequence statistics of a DNA sample, for example. When appropriate, we will point out how a class of variables or equations generalize to the multi-class, sequence case.

1. Definitions of sample statistics

This paper is concerned with point statistics related to the label decisions of an ensemble of binary classifiers. All of them can be calculated if we are given a map that gives us integer counts given the true label and the voting pattern for the ensemble members,

$$\text{label} \longrightarrow \{\ell_1, \ldots, \ell_n\} \longrightarrow n.$$

An example of this data structure is shown in Table 2. An equivalent formulation is to say that we have knowledge of the integer counts for every voting pattern given the true label, $n_{\ell_1,\ldots,\ell_n;\ell_{\text{true}}}$

1.1. Environmental ground truth statistics

The environmental ground truth statistics we consider are the two prevalences, ϕ_αȧnd ϕ_β. They are given by the usual ratio of label instances,

$$\phi_\alpha = \frac{n_\alpha}{n_\alpha + n_\beta} \tag{1}$$

$$\phi_\beta = \frac{n_\beta}{n_\alpha + n_\beta} \tag{2}$$

The total label instance counts can be written in terms of the $n_{\ell_1,\ldots,\ell_n;\ell_{\text{true}}}$,

$$n_\alpha = \sum_{\ell_1,\ldots,\ell_n} n_{\ell_1,\ldots,\ell_n;\alpha} \tag{3}$$

$$n_\beta = \sum_{\ell_1,\ldots,\ell_n} n_{\ell_1,\ldots,\ell_n;\beta} \tag{4}$$

$$ \tag{5}$$

1.2. Observable voting patterns

Each of the 2^n (L^n in the case of L labels) observable voting patterns for the n members of an ensemble are given by,

$$f_{\ell_1,\ldots,\ell_n} = n_{\ell_1,\ldots,\ell_n;\alpha} + n_{\ell_1,\ldots,\ell_n;\beta}. \tag{6}$$

The generalization to L labels is obvious. This follows from the assumption of a ground truth existence. Each instance can only have one of the labels. We get to observe the number of times a given voting pattern appears and our approach is to see how far purely algebraic considerations can take us in estimating the ground truth statistics.

Generalizing to events with given labels is straightforward also. In that case one can still enumerate all possible events given an ensemble of size n.. For example, statistics related to accuracy of consecutive labels would have 4^n observable decision events.

1.3. Classifiers marginal accuracies

Another useful way to encode our knowledge of the true labels, if we had it, would be to define an indicator functions for the decision of a classifier on an ℓ_{true}sample instance,

$$\mathbb{1}_{d,\ell_i,\ell_{\text{true}}}. \tag{7}$$

This allows us to define a sample statistic for the percentage of times a classifier made a label decision. By the assumption of ground truth existence, for a given ℓ_{true}sample instance d, all the indicator functions related to it would have to sum to one,

$$\sum_{\ell_i} \mathbb{1}_{d,\ell_i,\ell_{\text{true}}} = 1 \tag{8}$$

We define the frequencies of a label decision by,

$$\phi_{i,\ell_j,\ell_{\text{true}}} = \frac{\sum \mathbb{1}_{d,\ell_j,\ell_{\text{true}}}}{n_{\ell \text{true}}} \tag{9}$$

By Equation 8 these ground truth statistics must also sum to one. We can proceed forward with all the ϕ variables and add the normalization condition to our set of equations or we can reduce the dimension of the space by one and not carry around extra normalization conditions. We choose the later.

For the case of binary classification, we have the general conditions,

$$\phi_{i,\alpha,\alpha} + \phi_{i,\beta,\alpha} = 1 \tag{10}$$
$$\phi_{i,\alpha,\beta} + \phi_{i,\beta,\beta} = 1 \tag{11}$$
$$\tag{12}$$

Our variable elimination choice for the rest of the supplement is,

$$\phi_{i,\beta,\alpha} = 1 - \phi_{i,\alpha,\alpha} \tag{13}$$
$$\phi_{i,\alpha,\beta} = 1 - \phi_{i,\beta,\beta}. \tag{14}$$
$$\tag{15}$$

And to lighten notation, we drop the repeated labels,

$$\phi_{i,\alpha} = \phi_{i,\alpha,\alpha} \tag{16}$$
$$\phi_{i,\beta} = \phi_{i,\beta,\beta}. \tag{17}$$
$$\tag{18}$$

Note that these are marginal counts divided by an integer. Their value can be computed for a given sample if we had the ground truth knowledge of the labels. Its value is always be an integer ratio. These assumptions do not depend on any knowledge of how the classifier made its decisions. In particular, we do not need to assume that the instances are i.i.d., for example. Whether the samples were so produced, the marginal count has some integer value. No probability is needed to define it. How well can we measure such a sample statistic without probability assumptions?

1.4. Sample correlation statistics

There are different ways to define correlation between the classifiers. The one we use here for an m-way correlation measure is,

$$\Gamma_{i_1,\ldots,i_m,\ell_{\text{true}}} = \frac{\sum(\mathbb{1}_{d,\ell_{i_1},\ell_{\text{true}}} - \phi_{i_1,\ell_{\text{true}}})\ldots(\mathbb{1}_{d,\ell_{i_m},\ell_{\text{true}}} - \phi_{i_m,\ell_{\text{true}}})}{n_{\ell \text{true}}}. \tag{19}$$

These correlation measures generalize to a matrix for three or more classes. It just so happens that all possible correlations that one can define in the binary case are equal to these definitions or their negative. So a correlation for giving out the wrong label introduces a negative factor depending on whether it appears an even or odd amount of times. Since these sample statistics are rational functions of integer ratios, their value is also an integer ratio.

1.5. The definition of sample error independence

Independence between classifiers is also defined in terms of sample statistics in this paper. We consider an ensemble of n binary classifiers *independent* if all $\Gamma_{i,\ldots,m,\ell_{\text{true}}}$ are zero. This notion is not the same as that of distributional independence or any other notion based on probability. A finite sample from an independent distribution is most certainly likely to be non-independent by our sample based notion. Our notion of independence is meant to assist the task of evaluation. Nonetheless, it should be intuitively clear that, for large sample sizes, this sample based definition is close to a distributional one. **The question posed and partially answered by the paper is that our notion of sample independence can be tested in an unsupervised manner.**

2. Theorem proofs

Theorem 1 *Any of the voting pattern frequencies, $f_{\ell_1,\ldots,\ell_n}$, for n binary classifiers can be written as a polynomial in the variables ϕ_α, the marginal accuracies and correlation statistics*

This theorem is somewhat trivial given our definitions. We include it here for completeness and would not be surprised if an earlier published reference exists. The theorem follows from straightforward manipulations of Equation 6.

$$n_s f_{\ell_1,\ldots,\ell_n} = n_\alpha \frac{n_{\ell_1,\ldots,\ell_n;\alpha}}{n_\alpha} + n_\beta \frac{n_{\ell_1,\ldots,\ell_n;\beta}}{n_\beta} \tag{20}$$

$$f_{\ell_1,\ldots,\ell_n} = \frac{n_\alpha}{n_s}\frac{n_{\ell_1,\ldots,\ell_n;\alpha}}{n_\alpha} + \frac{n_\beta}{n_s}\frac{n_{\ell_1,\ldots,\ell_n;\beta}}{n_\beta} \tag{21}$$

$$f_{\ell_1,\ldots,\ell_n} = \phi_\alpha \frac{n_{\ell_1,\ldots,\ell_n;\alpha}}{n_\alpha} + (1-\phi_\alpha)\frac{n_{\ell_1,\ldots,\ell_n;\beta}}{n_\beta} \tag{22}$$

So far we have shown any frequency is a linear function of ϕ_α. To complete the proof, we need to show that the terms,

$$\frac{n_{\ell_1,\ldots,\ell_n;\ell_{\text{true}}}}{n_{\ell \text{true}}} \tag{23}$$

are themselves polynomials. To that end, consider the denominator. It can be expressed as a product of the indicator functions. We have to be somewhat careful with how we express the product because of our variable choices. In general, this will be a polynomial in powers of $\mathbb{1}_{i,\ell_i,\ell_{\text{true}}}$. For example, the indicator function for vote pattern (α, β) when the true label was α is given by,

$$\mathbb{1}_{i,\alpha,\alpha}(1-\mathbb{1}_{i,\beta,\alpha}). \tag{24}$$

But if the true label was β, the indicator polynomial counting those instances would be,

$$(1-\mathbb{1}_{i,\alpha,\beta})\mathbb{1}_{i,\beta,\beta}. \tag{25}$$

In any case, the proof now reduces to rewriting the label average of any product of the indicator functions. We detail now how the sample correlation statistics allow us to describe that space.

Consider the pair correlation statistics, $\Gamma_{i,j,\ell_{\text{true}}}$. By its definition we have,

$$\Gamma_{i,j,\ell_{\text{true}}} = \frac{1}{n_{\ell \text{true}}} \sum \mathbb{1}_{i,\ell_i,\ell_{\text{true}}} \mathbb{1}_{j,\ell_j,\ell_{\text{true}}} - \phi_{i,\ell_{\text{true}}}\phi_{j,\ell_{\text{true}}}. \tag{26}$$

So we can write the average product of two classifier indicators in terms of their marginals and the pair correlations,

$$\frac{1}{n_{\ell \text{true}}} \sum \mathbb{1}_{i,\ell_i,\ell_{\text{true}}} \mathbb{1}_{j,\ell_j,\ell_{\text{true}}} = \phi_{i,\ell_{\text{true}}}\phi_{j,\ell_{\text{true}}} + \Gamma_{i,j,\ell_{\text{true}}}. \tag{27}$$

We have written the pair indicator product as polynomial in the pair correlation statistic and the marginals. This pattern holds in general. We can write any product of three indicator functions as a polynomial in the marginals, the pair correlations the 3-way correlation statistics. It follows that any indicator polynomial can be written in terms of these statistics alone.

Theorem 2 *The variety of the polynomial system for three independent binary classifiers consists of two point solutions, one of them the ground truth values*

The polynomial system for independent classifiers written out in the paper follows from using Theorem 1 and setting all Γ variables to zero. The correlation variables in effect parametrize when we can factor averages of indicator products as products of marginals. By construction then, the set of points that solves the polynomial system must contain the ground truth values. This set of points is the variety.

The proof that the variety of the independent classifiers system is zero-dimensional and only contains two points is more involved. There are two ways we could proceed here, one algebraic, the other geometric. The algebraic is less intuitive but fits better the algebraic nature of the rest of the paper. We proceed with the algebraic approach and briefly mention the geometric one at the end.

The algebraic proof proceeds by application of Buchberger's algorithm to the independent classifiers polynomial system. It yields seven polynomials in what is called an elimination chain. Their algebraic structure makes clear they define two point solutions. They are complicated polynomials in the voting pattern frequencies. But they are simple algebraic functions of the ground truth statistics,

$$
\begin{aligned}
a\phi_{\alpha}^2 + b\phi_{\alpha} + c &= 0 && (28)\\
d_{11} + e_{11}\phi_{\alpha} + g_{11}\phi_{1,\alpha} &= 0 && (29)\\
d_{12} + e_{12}\phi_{\alpha} + g_{12}\phi_{1,\beta} &= 0 && (30)\\
d_{21} + e_{21}\phi_{\alpha} + g_{21}\phi_{2,\alpha} &= 0 && (31)\\
d_{22} + e_{22}\phi_{\alpha} + g_{22}\phi_{2,\beta} &= 0 && (32)\\
d_{31} + e_{31}\phi_{\alpha} + g_{31}\phi_{3,\alpha} &= 0 && (33)\\
d_{32} + e_{32}\phi_{\alpha} + g_{32}\phi_{3,\beta} &= 0, && (34)\\
& && (35)
\end{aligned}
$$

where the coefficients are the complicated polynomials we have abstracted away. It follows from this algebraic structure that there are only two point solutions in the variety since the equation for $\phi_{\alpha}$is a quadratic and all the other equations are linear in ϕ_{α}.

This completes the proof. The geometrical proof consists of calculating the dimension of the variety for n classifiers. As shown in the paper, the 3 dimensional space for one classifier has a variety of dimension 2. The variety of two independent classifiers exists in a five dimensional space. It has dimension three. Finally, the variety for three independent binary classifiers exists in a 7 dimensional space, and as we have shown, it consists of just two points.

Theorem 3 *The quadratic polynomial for ϕ_{α}obtained when we assume the classifiers are independent contains an irreducible discriminant for arbitrarily correlated classifiers*

The proof is basically complete as given in the paper. Here we just illustrate how this works out algebraically with one specific case - all correlation factors zero except one pair correlation term.

Recall that the solution to the quadratic $ax^2 + bx + c$ is given by the quadratic formula,

$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \tag{36}$$

It so happens in this case that the correct root to choose to obtain the same value for the prevalence is the minus one. The plus one returns $1 - \phi_{\alpha}$. In the case of independent classifiers the discriminant term is equal to,

$$(1 - 2\phi_{\alpha})^2 (\phi_{\alpha} - 1)^4 \phi_{\alpha}^4 (\phi_{1,\alpha} + \phi_{1,\beta} - 1)^4 (\phi_{2,\alpha} + \phi_{2,\beta} - 1)^4 (\phi_{3,\alpha} + \phi_{3,\beta} - 1)^4. \tag{37}$$

If we only turn on $\Gamma_{1,2,\alpha}$, we get

$$
\begin{aligned}
&(\phi_{\alpha} - 1)^4 \phi_{\alpha}^4 (\phi_{3,\alpha} + \phi_{3,\beta} - 1)^4 . (\phi_{1,\alpha} + \phi_{1,\beta} - 1)^2 (\phi_{2,\alpha} + \phi_{2,\beta} - 1)^2\\
&\qquad \left((1 - 2\phi_{\alpha})^2 (\phi_{1,\alpha} + \phi_{1,\beta} - 1) (\phi_{2,\alpha} + \phi_{2,\beta} - 1) + \Gamma_{1,2,\alpha}\right)\\
&\qquad\qquad ((\phi_{1,\alpha} + \phi_{1,\beta} - 1) (\phi_{2,\alpha} + \phi_{2,\beta} - 1) - \Gamma_{1,2,\alpha})
\end{aligned} \tag{38}
$$

This is unlikely to factor into a perfect square - a probabilistic argument! One purpose for showing this very simple case is to show that we can, nonetheless, Taylor expand the independent solution about a zero value. We have not done so here. Nonetheless, the polynomial nature of the solution means that we can always find such an expansion close enough to the independent errors point. Getting the Taylor expansion would then allow us to quantify the expected consistency between nearly independent classifiers.

3. Experimental Results

3.1. The two-norm experiment

The classifiers used, along with the features used (1-indexed), are given in Table 1

algorithm	feature set
NeuralNetwork, with NetworkDepth $= 5$	1, 6, 8, 10, 11
GradientBoostedTrees	2, 3, 5, 9, 12
NaiveBayes	4, 13, 15, 19, 20
LogisticRegression	7, 14, 16, 17, 18

Table 1. Algorithms and features used by the four classifiers in the twonorm exemplar experiment

The ground truth counts are shown in Table 2.

decision event	0 label	1 label
{0,0,0,0}	4	1330
{0,0,0,1}	7	237
{0,0,1,0}	18	271
{0,0,1,1}	58	48
{0,1,0,0}	15	285
{0,1,0,1}	58	42
{0,1,1,0}	50	58
{0,1,1,1}	268	7
{1,0,0,0}	11	258
{1,0,0,1}	52	38
{1,0,1,0}	42	58
{1,0,1,1}	247	8
{1,1,0,0}	56	38
{1,1,0,1}	284	7
{1,1,1,0}	245	5
{1,1,1,1}	1172	1

Table 2. Observed decision event counts by true label for the ensemble of four classifiers in the twonorm exemplar experiment

3.1.1. Marginal accuracy estimate variances

The goal of our experimental results has been to demonstrate exemplars of the different ways we can see consistency in the trio estimates. This experiment, tt two-norm, demonstrates that you can find find datasets and classifiers independent enough to get estimates within a few percent of the correct value for both labels. But how hard is it to find this condition? We briefly address that by considering repeated applications of our experimental protocol.

Our 1st repeated application retained the same classifier algorithms, the same feature sets, and test set - only changing the training data used by each classifier (750 out of 1000 samples for each label). One hundred experimental runs had an 1.2% RMSE for the prevalence. The average RMSE for the α label was about 0.9% for the alpha accuracies, and about 1.0% on the β label. These ranged from 0.5% to 1.7%. The figures (1 and 2) below show the error deltas for these two extreme cases.

Our 2nd set of runs varied the feature set partitions and training data alon. We still partitioned the twenty features for two-norm into four disjoint sets of five features each. These runs had a higher RMSE, averaging about 3% as shown in

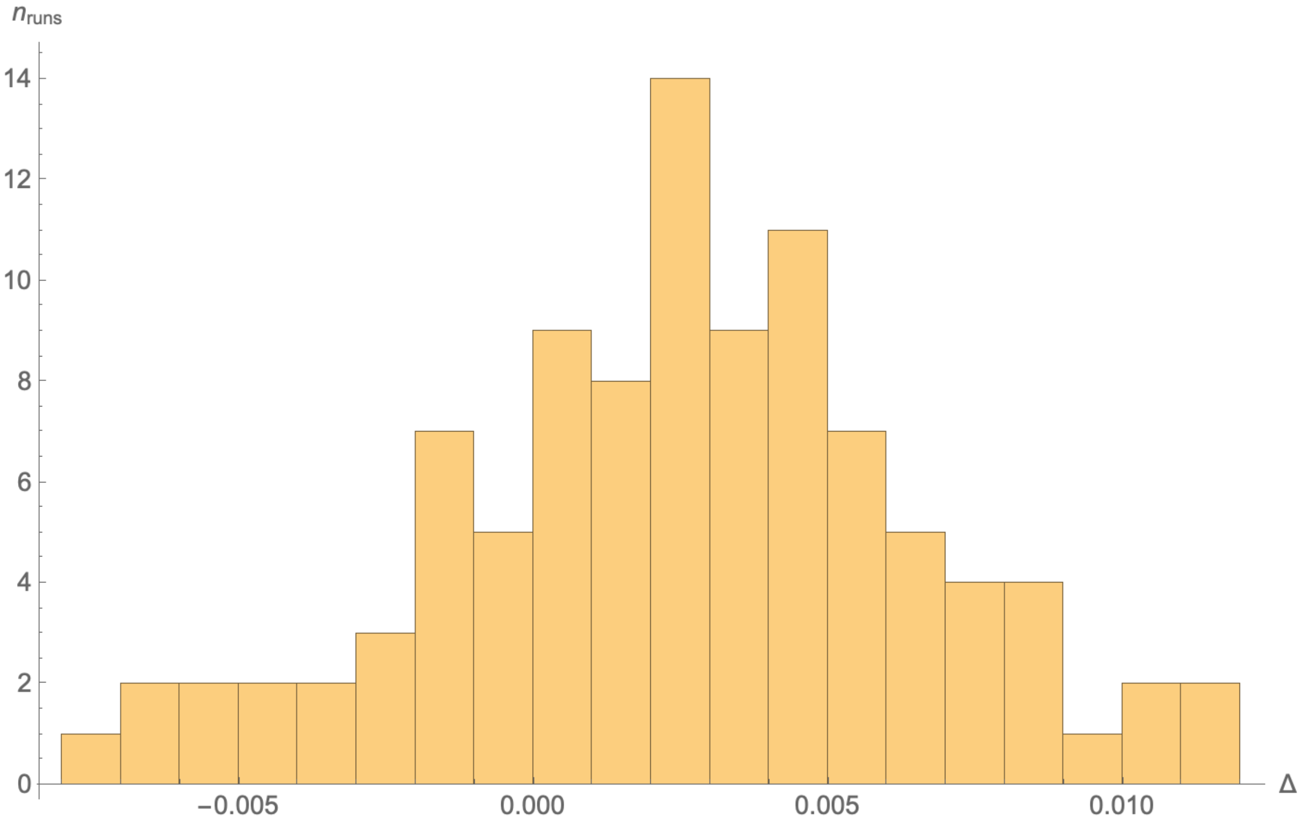

Figure 1. Histogram of the error deltas (Δ) for the least varying classifier accuracy ($\phi_{2,\alpha}$) in the 1st protocol for 100 repeated runs

the corresponding figures (and).

3.2. `spambase` experiment

The classifier algorithms and features used in the `spambase` experiment are shown in Table 3. The ground truth counts for the point statistics are shown in Table 4.

algorithm	feature set
`NeuralNetwork`, with NetworkDepth $= 4$	2, 11, 13, 18, 21, 25, 33, 35, 44, 56
`SupportVectorMachine`, KernelType $=$ Polynomial, and PolynomialDegree $= 3$	1, 3, 4, 14, 28, 38, 39, 43, 54, 57
`DecisionTree`, DistributionSmoothing $= 5$	6, 15, 16, 27, 30, 31, 42, 45, 46, 53
`NaiveBayes`	8, 10, 12, 24, 29, 32, 40, 41, 50, 55

Table 3. Algorithms and features used by the four classifiers in the `spambase` exemplar experiment

The ground truth counts are shown in Table 4.

3.3. `mushroom` experiment

The classifier algorithms and features used in the `mushroom` experiment are shown in Table 5. The ground truth counts for the point statistics are shown in Table 6.

The ground truth counts are shown in Table 6.

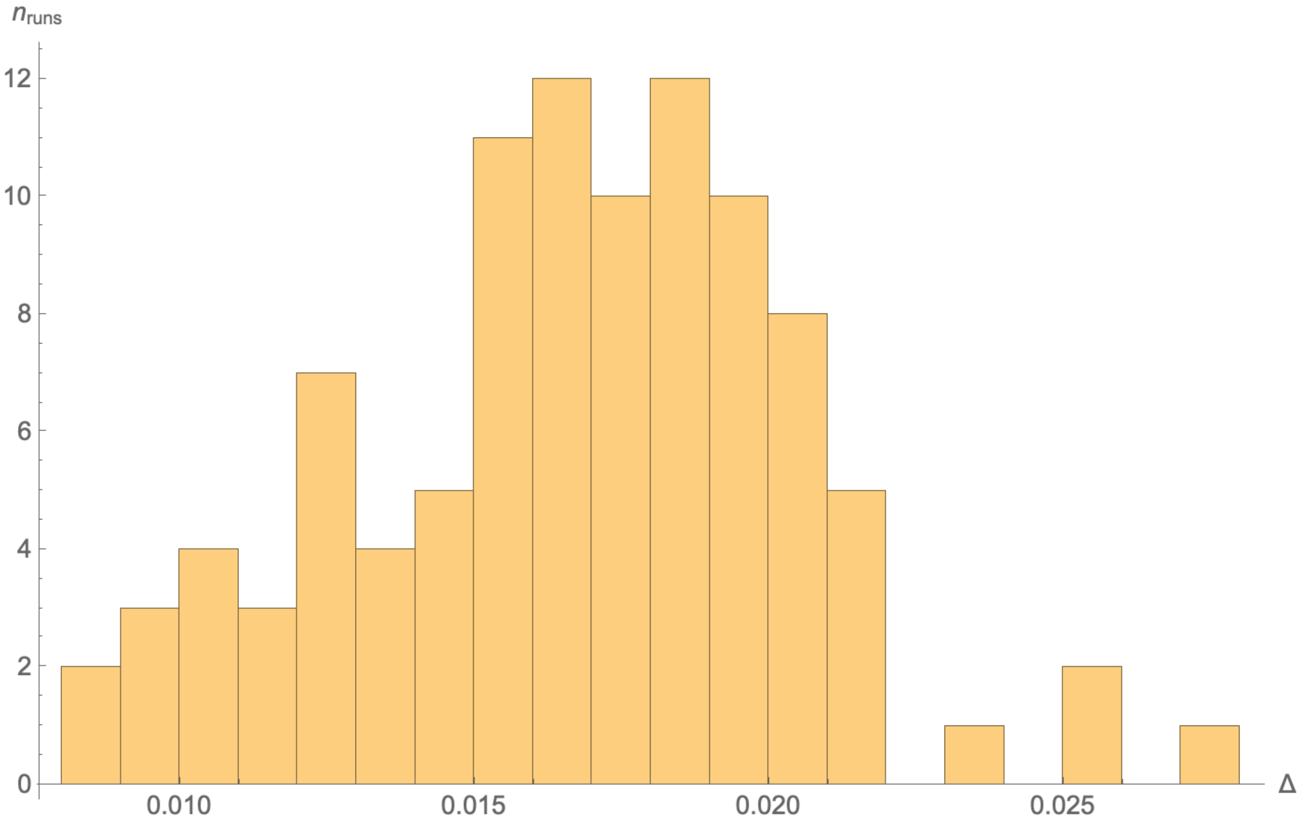

Figure 2. Histogram of the error deltas (Δ) for the most varying classifier accuracy ($\phi_{2,\beta}$) in the 1st protocol for 100 repeated runs

4. Comparison to previous work

In this section we make some quick comparisons between equations in our work and some of the previous literature we mentioned in the paper.

4.1. Platanois, Dubey, and Mitchell (Platanios et al., 2016)

Platanios *et al.* have a result very close to our purely algebraic one for independent classifiers. They define the average error rate for classifiers and are able to obtain a closed solution of a linear system they construct (equation 7, (Platanios et al., 2016)). Does this linear solution, considering only pair agreement/disagreement statistics solve the full problem where we have the $\phi_\alpha$estimate as well as the classifier margins? No. This can be stated mathematically as follows.

Their average error for a classifier is given, with our sample statistics, by,

$$e_{\{i\}} = \phi_\alpha(1 - \phi_{i,\alpha}) + (1 - \phi_\alpha)(1 - \phi_{i,\beta}. \tag{39}$$

So point solutions of these equations, even if we knew the left hand sides perfectly, would require solving for 7 variables with 3 equations. It cannot be done.

4.2. Jaffe, Fetaya, Nadler, Jian, and Kluger (Jaffe et al., 2016)

Jaffe *et al.* (Jaffe et al., 2016) consider how to detect clusters of pair correlated classifiers. To this end, they define a quantity T_{ij} sample statistic that can be written as a polynomial of the four voting patterns for two classifiers. The four polynomials one would write with the general moments theorem, 1, define an algebraic ideal, as mentioned before. The sample statistic version of the T_{ij} can be shown to belong to the same ideal. This is technically proven by showing that the quantity T_{ij} can be written as a linear combination of the polynomials in the Gröbner basis for the Theorem 1 polynomial system.

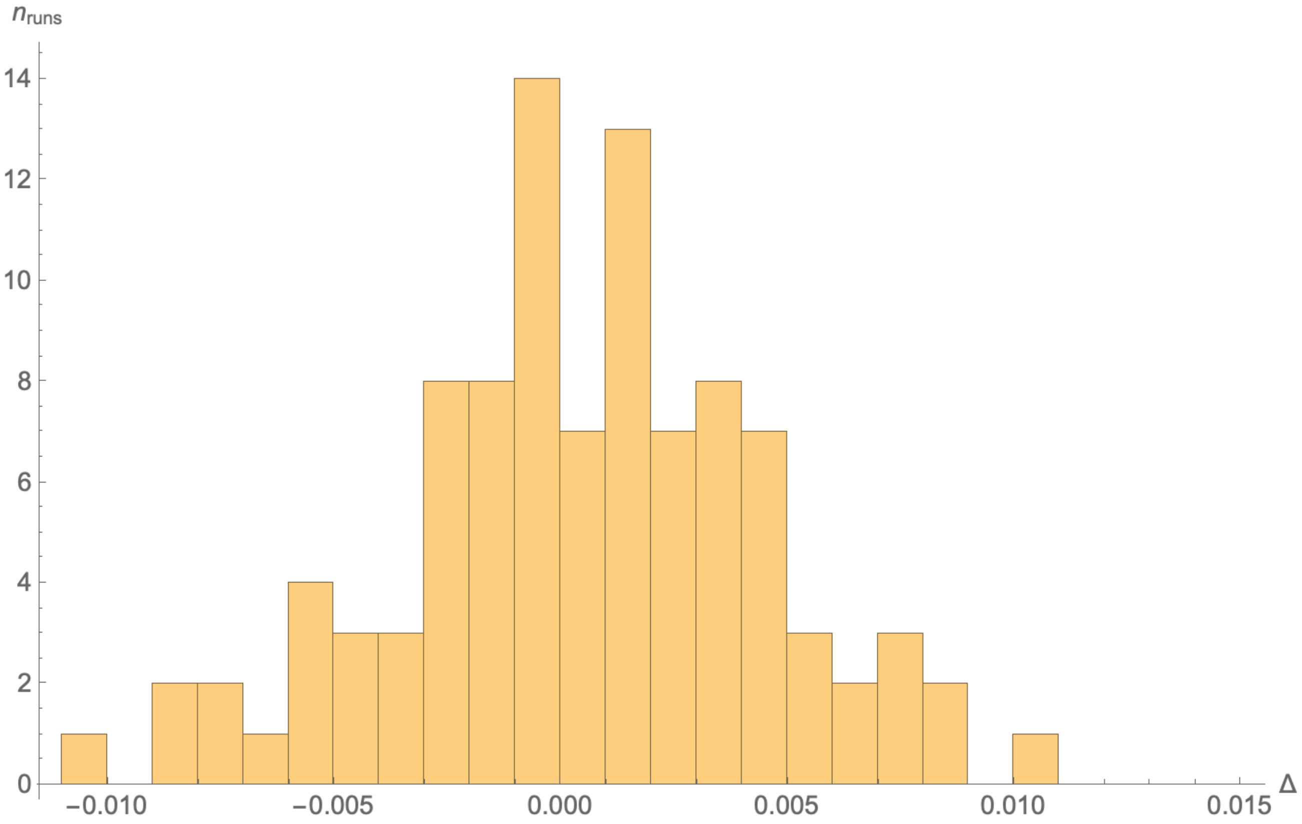

Figure 3. Histogram of the error deltas (Δ) for the least varying classifier accuracy ($\phi_{2,\alpha}$) in the 2nd protocol for 100 repeated runs

One can ask, more generally, how many classifiers would be needed to measure the amount of pair correlation between two classifiers? The polynomials shown here cannot resolve pair correlations until we use the moments of three or more classifiers. This is a purely algebraic result, the use of additional assumptions could circumvent this limit.

Another question not explored here for dependent classifiers is whether a given set of voting pattern frequencies could be the output of purely binary classification decisions. The frequencies are integer ratios by definition, but are all bounded integer ratios that sum to one explainable by the general system? If not, this could be used to detect certain cases of spoofed counts.

References

Jaffe, A., Fetaya, E., Nadler, B., Jiang, T., and Kluger, Y. Unsupervised ensemble learning with dependent classifiers. In Gretton, A. and Robert, C. C. (eds.), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, pp. 351–360, Cadiz, Spain, 2016. PMLR.

Platanios, E. A., Dubey, A., and Mitchell, T. Estimating accuracy from unlabeled data: A bayesian approach. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1416–1425, New York, New York, USA, 2016.

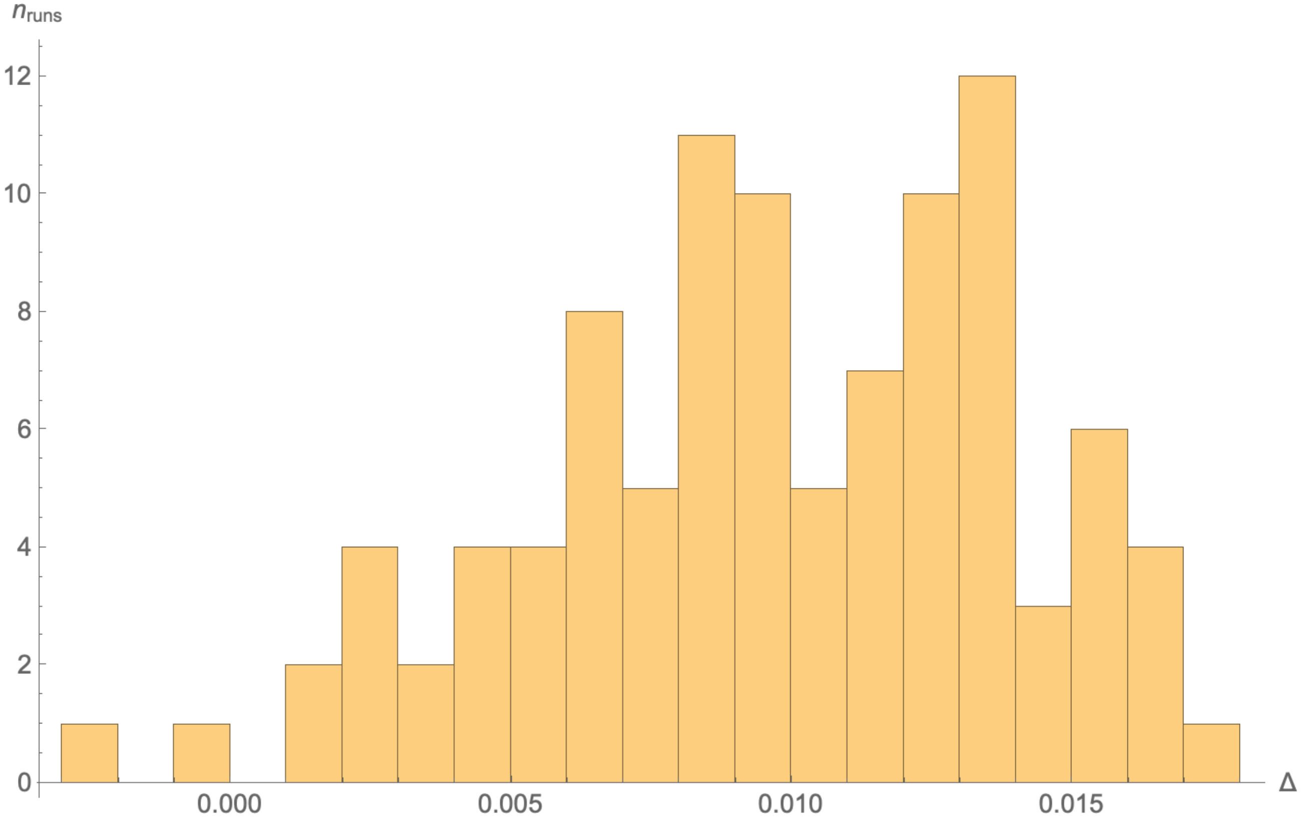

Figure 4. Histogram of the error deltas (Δ) for the most varying classifier accuracy ($\phi_{3,\beta}$) in the 2nd protocol for 100 repeated runs

decision event	0 label	1 label
{0,0,0,0}	1827	185
{0,0,0,1}	53	25
{0,0,1,0}	145	153
{0,0,1,1}	13	30
{0,1,0,0}	121	14
{0,1,0,1}	17	12
{0,1,1,0}	9	14
{0,1,1,1}	3	10
{1,0,0,0}	223	151
{1,0,0,1}	10	90
{1,0,1,0}	45	182
{1,0,1,1}	5	268
{1,1,0,0}	26	22
{1,1,0,1}	5	66
{1,1,1,0}	5	65
{1,1,1,1}	2	345

Table 4. Observed decision event counts by true label for the ensemble of four classifiers in the `spambase` exemplar experiment

algorithm	feature set
`DecisionTree`, with DistributionSmoothing $= 5$	2, 9, 11, 13, 18, 19
`NaiveBayes`	6, 7, 12, 14, 20, 21
`NeuralNetwork`, with NetworkDepth $= 4$	4, 8, 10, 15, 16, 22
`SupportVectorMachine`, with KernelType = Polynomial, and PolynomialDegree $= 3$	1, 3, 5, 8, 17

Table 5. Algorithms and features used by the four classifiers in the `mushroom` exemplar experiment

decision event	0 label	1 label
{0,0,0,0}	2929	0
{0,0,0,1}	75	0
{0,0,1,0}	70	28
{0,0,1,1}	45	266
{0,1,0,0}	135	35
{0,1,0,1}	0	0
{0,1,1,0}	16	14
{0,1,1,1}	42	174
{1,0,0,0}	310	0
{1,0,0,1}	5	0
{1,0,1,0}	110	29
{1,0,1,1}	10	106
{1,1,0,0}	20	29
{1,1,0,1}	0	129
{1,1,1,0}	20	0
{1,1,1,1}	0	2714

Table 6. Observed decision event counts by true label for the ensemble of four classifiers in the `mushroom` exemplar experiment