A STUDENT'S PROBABILITY OF FINDING A BASIS

C. Bryan Dawson

A few years ago, I was teaching a course called Vector Spaces (our second linear algebra course, for seniors or beginning graduate students), and having done some review at the beginning of the semester, I decided to put an "easy" question on the test:

"Give a basis for $\mathbb{R}^4$ other than the canonical unit vector basis."

Of course, that task is simply enough done, but it will probably come as no surprise to the reader that two of the seven students, including one of the best of the group, answered the question incorrectly. It is the purpose of this paper to explore issues relating to the questions, "What in the world were they thinking?" and "How likely is it that a student will guess the answer?"

We will assume that the students who answered incorrectly were unaware of the various methods for quickly writing down an answer to the question. Also, none of the students checked their answer by using Gaussian elimination (which, of course, was not necessary for the students who knew what they were doing). We must, therefore, assume that the students were trying some method of guessing by writing down four vectors (which, thankfully, they all did) and somehow applying "naive" checks to make it "look right."

The first idea that comes to mind is to determine the probability of finding a basis by randomly choosing four vectors from $\mathbb{R}^4$. If one interprets this as randomly choosing four vectors of unit length using a "uniform" distribution, then a simple geometric argument shows that this method determines a basis with probability 1. Thus, either the students were having an unbelievably bad day or they were not using this method. The latter is the case. In fact, most (if not all) students gave vectors whose entries consisted only of 0's and 1's.

We now have a new question, namely, "What is the probability of obtaining a basis by randomly choosing vectors consisting only of 0's and 1's?" We must be careful, though, to restrict our randomness according to certain simple conditions which the students appeared to have checked: no student included the zero vector in their attempt at a basis, no one used two identical vectors, and, if memory serves me correctly, no two vectors added to give a third. In other words, the students at least knew something. The following theorem makes me grateful that I asked the question in dimension four.

Theorem 1. Consider the following list of conditions:
(1) No vector is the zero vector.
(2) No two vectors are identical.
(3) No two vectors add to give a third.
Let $k \in \{1, 2, 3\}$ and $n \geq k$. Then a set of k vectors in $\mathbb{R}^n$, all of whose components are from the set $\{0, 1\}$, is linearly independent if conditions (1)–(k) hold.

The proof is simple for the cases $k = 1$ and $k = 2$. The proof for the $k = 3$ case lends some insight into later work, so we sketch it here. Suppose our three conditions are met and $\alpha x + \beta y = z$ for our three vectors x, y, and z. By the $k = 2$ case, α, $\beta \neq 0$. As x and y must be different, there is a row in which one of them (WLOG x) has component 0 and the other (y) has component 1. Then β must be the component of z in that same row. Since β cannot be 0, we have $\beta = 1$. Since x is not the zero vector, there is some row in which its component is 1. Then the following cases result (each row here is a possible row where x has component 1):

x	y	z	result
1	0	0	$\alpha = 0$
1	0	1	$\alpha = 1$
1	1	0	$\alpha = -1$
1	1	1	$\alpha = 0$

Since $\alpha = 0$ contradicts, we have either $x + y = z$, which violates condition 3, or $-x + y = z$, which yields $y = x + z$ and again contradicts condition 3.

The above theorem shows that a student would be able to successfully answer my test question for $\mathbb{R}^3$ applying only these three conditions to their vectors of 0's and 1's. It is also not unreasonable to think that a student might check these natural conditions. The question that arises is then whether or not there are additional natural conditions guaranteeing success for four vectors. Adding "no three add to give a fourth" is not sufficient as seen by

$$\frac{1}{2}\begin{bmatrix}0\\1\\0\\1\end{bmatrix} + \frac{1}{2}\begin{bmatrix}1\\0\\1\\1\end{bmatrix} + \frac{1}{2}\begin{bmatrix}1\\1\\1\\0\end{bmatrix} = \begin{bmatrix}1\\1\\1\\1\end{bmatrix}.$$

Checking the three conditions for rows as well as columns also fails:

$$\begin{bmatrix}0\\1\\0\\1\end{bmatrix}+\begin{bmatrix}1\\0\\1\\0\end{bmatrix}-\begin{bmatrix}1\\1\\0\\0\end{bmatrix}=\begin{bmatrix}0\\0\\1\\1\end{bmatrix}.$$

In fact, we offer the following theorem, whose proof, which is similar to that of the $k = 3$ case above, is omitted as it was approximately 4.5 handwritten pages of straightforward, but tedious, work.

Theorem 2. If conditions (1)–(3) of Theorem 1 hold, $\alpha x + \beta y + \gamma z = w$, and the components of x, y, z, and w are only 0's and 1's, then α, β, and γ can be permuted into one of the following triples:

$$\begin{array}{lll}
1, & 1, & 1\\
1, & 1, & -1\\
1, & -1, & -1\\
\frac{1}{2}, & \frac{1}{2}, & \frac{1}{2}\\
\frac{1}{2}, & \frac{1}{2}, & -\frac{1}{2}\\
2, & 1, & -1\\
-2, & 1, & 1\\
2, & -1, & -1
\end{array}$$

This seems to be an unreasonable list of conditions for a student to check on an exam. Indeed, it is an unreasonable list of conditions to expect the average student to figure out on their own in the first place. Also, no other natural condition jumps out from the above. We may therefore conclude that $\mathbb{R}^4$ is a much better space than $\mathbb{R}^3$ in which to ask the test question.

We now return to our original two questions. It is possible that the students were "randomly" writing vectors of 0's and 1's and checking the three conditions. What then, is their probability of success? We attack this question next.

Although there are $2^{16} = 65536$ possible ordered sets of four vectors of four 0's and 1's, only $15 \cdot 14 \cdot 13 \cdot 12 = 32760$ meet conditions (1) and (2). Now, since no two vectors are the same, they may be rearranged in order from smallest to largest, where the ordering is accomplished by viewing the vectors as integers base 2. Hence, each possible different set of vectors is listed 24 times above, leaving only 1365 cases to check, not all of which meet condition (3). We also have a systematic way of listing these cases. Note that to check condition 3, we need only check addition of smaller vectors yielding a larger one. Similarly, in the interest of using

Theorem 2, we need only check to see if the larger vector is a combination of the smaller ones using the coefficients from rows 1, 2, 4, 6, or 7 of that theorem.

We should add one more condition before finding our probability. In addition to checking conditions (1)–(3), we assume the student will check to see that

(4) No row consists of only 0's.

Then the list of possible student answers is comprised of those sets of four vectors of four 0's and 1's meeting these four conditions. Just for the fun of it, we check to see which of these fails to be a basis by Theorem 2, as reconfigured in the preceding paragraph. This can be done by hand, in only a few hours, on only 14 pages. In order not to spoil the reader's fun, we omit the details so that the reader may enjoy doing it also. It turns out that 1038 of the 1365 cases meet the four conditions, and only 80 of these fail to be a basis. Hence, we have

Theorem 3. The probability that a randomly selected set of four vectors from $\mathbb{R}^4$ whose entries consist only of 0's and 1's forms a basis, given that conditions (1)–(4) hold, is $\frac{958}{1038} \approx .923$. If we are only given that (1) and (2) hold, the probability is $\frac{958}{1365} \approx .702$.

The original test question excluded one basis, and hence, the probability of success given conditions (1)–(4) is $\frac{957}{1037}$. If all seven students guessed in the same manner, the probability of two or more failures is approximately .096. Since at least some students did not guess, it is not likely, but still somewhat plausible, that the students checked conditions (1)–(4). Two more likely scenarios remain. One is that only conditions (1) and (2) were checked, which seems a very reasonable possibility (especially if it was the case that one of them failed condition (3)). The other is that in order to make their answers "look right," they gravitated toward areas of higher probability of failure. The fact that there are such areas of greater failure rates can be seen by doing the analysis of Theorem 3 by hand.

Exploring the two questions of this paper can be interesting and enjoyable. Unfortunately, we will likely never know the answer to the first of the questions. Considering some of what we have discovered, we may not want to know.

C. Bryan Dawson
Division of Mathematics and Computer Science
Emporia State University
Emporia, KS 66801
email: dawsonbr@esumail.emporia.edu