Notes on Discrete Probability

The following notes cover, mostly without proofs, the basic notions and results of discrete probability. Two important topics (the *birthday paradox* and *universal hash functions*) are not covered, and they will be the subject of future notes.

1 Basic Definitions

In cryptography we typically want to prove that an adversary that tries to break a certain protocol has only minuscule (technically, we say "negligible") probability of succeeding. In order to prove such results, we need some formalism to talk about the probability that certain events happen, and also some techniques to make computations about such probabilities.

In order to model a probabilistic system we are interested in, we have to define a *sample space* and a *probability distribution*. The sample space is the set of all possible *elementary events*, i.e. things that can happen. A probability distribution is a function that assigns a non-negative number to each elementary event, this number being the *probability* that the event happen. We want probabilities to sum to 1. Formally,

Definition 1.1 *For a finite sample space* Ω *and a function* $\mathbf{Pr} : \Omega \to \mathbf{R}$*, we say that* $\mathbf{Pr}$ *is a* probability distribution *if*

1. $\mathbf{Pr}(a) \geq 0$ *for every* $a \in \Omega$*;*
2. $\sum_{a\in\Omega} \mathbf{Pr}(a) = 1$*.*

For example, if we want to model a sequence of three coin flips, our sample space will be $\{Head, Tail\}^3$ (or, equivalently, $\{0,1\}^3$) and the probability distribution will assign 1/8 to each element of the sample space (since each outcome is equally likely).

If we model an algorithm that first chooses at random a number in the range $1, \ldots, 10^{200}$ and then does some computation, our sample space will be the set $\{1, 2, \ldots, 10^{200}\}$, and each element of the sample space will have probability $1/10^{200}$.

We will always restrict ourselves to *finite* sample spaces, so we will not remark it each time. *Discrete probability* is the restriction of probability theory to finite sample spaces. Things are much more complicated when the sample space can be infinite.

An *event* is a subset $A \subseteq \Omega$ of the sample space. The probability of an event is defined in the intuitive way

$$\mathbf{Pr}[A] = \sum_{a \in A} \mathbf{Pr}(a)$$

(Conventially, we set $\mathbf{Pr}[\emptyset] = 0$.)

We use square brackets to remind us that now we are considering a different function: while $\mathbf{Pr}(\cdot)$ is a function whose inputs are *elements* of the sample space, $\mathbf{Pr}[\cdot]$ is a function whose inputs are *subsets* of the sample space.

For example, suppose that we want to ask what is the probability that, when flipping three coins, we get two heads. Then $\Omega = \{0,1\}^3$, $\mathbf{Pr}(a) = 1/8$ for every $a \in \Omega$, we define A as the subset of $\{0,1\}^3$ containing strings with exactly two 1s, and we ask what is $\mathbf{Pr}[A]$. As it turns out, A has 3 elements, that is $011, 101, 110$, and so $\mathbf{Pr}[A] = 3/8$. Very often, as in this example, computing the probability of an event reduces to counting the number of elements of a set.

When $\mathbf{Pr}(\cdot)$ assigns the same value $1/|\Omega|$ to all the elements of the sample space, then it is called the *uniform distribution over* Ω.

The following distribution arises very often, and it is good to know about it. Consider the situation where you flip n biased coins, and the probability that each coin turns out head is p (where $0 \leq p \leq 1$ is some fixed parameter). The outcome of each flip is independent of all the other outcomes. Then the sample space is $\Omega = \{0,1\}^n$, identifying heads with 1s; the probability distribution is

$$\mathbf{Pr}(a) = p^k (1-p)^{n-k} \text{ where } k \text{ is the number of 1s in } a$$

When $p = 1/2$ then we have the uniform distribution over $\{0,1\}^n$. If $p = 0$ then all a have probability zero, except $00 \cdots 0$, which has probability one. (Similarly if $p = 1$.) The other cases are more interesting.

These distributions are called *Bernoulli* distributions or *binomial* distributions.

If we have a binomial distribution with parameter p, and we ask what is the probability of the event A_k that we get a string with k ones, then such a probability is

$$\mathbf{Pr}[A_k] = \binom{n}{k} p^k (1-p)^{n-k}$$

2 Random Variables and Expectation

Very often, when studying a probabilistic system (say, a randomized algorithm) we are interested in some values that depend on the elementary event that takes place. For example, when we play dice, we are interested in the probabilistic system where two dice are rolled, and the sample space is $\{1, 2, \ldots, 6\}^2$, with the uniform distribution over the 36 elements of the sample space, and we are interested in the *sum* of the outcomes of the two dice. Similarly, when we study a randomized algorithm that makes some internal random choices, we are interested in the *running time* of the algorithm, or in its *output*. The notion of a *random variable* gives a tool to formalize questions of this kind.

A *random variable* X is a function $X : \Omega \to V$ where Ω is a sample space and V is some arbitrary set (V is called the *range* of the random variable). One should think of a random variable as an algorithm that on input an elementary event returns some output. Typically, V will either be a subset of the set of real numbers or of the set of binary strings of a certain length.

Let Ω be a sample space, $\mathbf{Pr}$ a probability distribution on Ω and X be a random variable on Ω. If v is in the range of X, then the expression $X = v$ denotes an event, namely the event $\{a \in \Omega : X(a) = v\}$, and thus the expression $\mathbf{Pr}[X = v]$ is well defined, and it is something interesting to try to compute.

Let's look at the example of dice. In that case, $\Omega = \{1, \dots, 6\}^2$, for every $(a, b) \in \Omega$ we have $\mathbf{Pr}(a, b) = 1/36$. Let us define X as the random variable that associates $a + b$ to an elementary event (a, b). Then the range of X is $\{2, 3, \dots, 12\}$. For every element of the range we can compute the probability that X take such value. By counting the number of elementary events in each event we get

$$\mathbf{Pr}[X = 2] = 1/36\ ,\ \mathbf{Pr}[X = 3] = 2/36\ ,\ \mathbf{Pr}[X = 4] = 3/36$$

$$\mathbf{Pr}[X = 5] = 4/36\ ,\ \mathbf{Pr}[X = 6] = 5/36\ ,\ \mathbf{Pr}[X = 7] = 6/36$$

and the other probabilities can be computed by observing that

$$\mathbf{Pr}[X = v] = \mathbf{Pr}[X = 14 - v]$$

It is possible to define more than one random variable over the same sample space, and consider expressions more complicated than equalities.

When the range of a random variable X is a subset of the real numbers (e.g. if X is the running time of an algorithm — in which case the range is even a subset of the integers) then we can define the *expectation* of X. The expectation of a random variable is a number defined as follows.

$$\mathbf{E}[X] = \sum_{v \in V} v\mathbf{Pr}[X = v]$$

where V is the range of X. We can assume without loss of generality that V is finite, so that the expression above is well defined (if it were an infinite series, it could diverge or even be undefined).

Expectations can be understood in terms of betting. Say that I am playing some game where I have a probability $2/3$ of winning, a probability $1/6$ of losing and a probability $1/6$ of a draw. If I win, I win \$ 1; if I lose I lose \$ 2; if there is a draw I do not win or lose anything. We can model this situation by having a sample space $\{L, D, W\}$ with probabilities defined as above, and a random variable X that specifies my wins/losses. Specifically $X(L) = -2$, $X(D) = 0$ and $X(W) = 1$. The expectation of X is

$$\mathbf{E}[X] = \frac{1}{6} \cdot (-2) + \frac{1}{6} \cdot 0 + \frac{2}{3} \cdot 1 = \frac{1}{3}$$

so if I play this game I "expect" to win \$ 1/3. The game is more than fair on my side.

When we analyze a randomized algorithm, the running time of the algorithm typically depends on its internal random choices. A complete analysis of the algorithm would be a specification of the running time of the algorithm for *each* possible sequence of internal choices. This is clearly impractical. If we can at least analyse the *expected* running time of the algorithm, then this will be just a single value, and it will give useful information about the typical behavior of the algorithm (see Section 4 below).

Here is a very useful property of expectation.

Theorem 2.1 (Linearity of Expectation) *Let X be a random variable and a be real; then $\mathbf{E}[aX] = a\mathbf{E}[X]$. Let $X_1, \ldots, X_n$ be random variables over the same sample space; then $\mathbf{E}[X_1 + \cdots + X_n] = \mathbf{E}[X_1] + \cdots \mathbf{E}[X_n]$.*

Example 1 Consider the following question: if we flip a coin n times, what is the expected number of heads? If we try to answer this question without using the linearity of expectation we have to do a lot of work. Define $\Omega = \{0,1\}^n$ and let $\mathbf{Pr}$ be the uniform distribution; let X be the random variable such that $X(a) =$ the number of 1s in $a \in \Omega$. Then we have, as a special case of Bernoulli distribution, that

$$\mathbf{Pr}[X = k] = \binom{n}{k} 2^{-n}$$

In order to compute the average of X, we have to compute the sum

$$\sum_{k=0}^{n} \binom{n}{k} k 2^{-n} \tag{1}$$

which requires quite a bit of ingenuity. We now show how to solve Expression (1) just to see how much work can be saved by using the linearity of expectation. An inspection of Expression (1) shows that it looks a bit like the expressions that one gets out of the Binomial Theorem, except for the presence of k. In fact it looks pretty much like the *derivative* of an expression coming from the Binomial Theorem (this is a standard trick). Consider $(1/2 + x)^n$ (we have in mind to substitute $x = 1/2$ at some later point), then we have

$$\left(\frac{1}{2} + x\right)^n = \sum_{k=0}^{n} \binom{n}{k} 2^{-(n-k)} x^k$$

and then

$$\frac{d((1/2 + x)^n)}{dx} = \sum_{k=0}^{n} \binom{n}{k} 2^{-(n-k)} k x^{k-1}$$

but also

$$\frac{d((1/2+x)^n)}{dx} = n\left(\frac{1}{2}+x\right)^{n-1}$$

and putting together

$$\sum_{k=0}^{n}\binom{n}{k}2^{-(n-k)}kx^{k-1} = n\left(\frac{1}{2}+x\right)^{n-1} .$$

Now we substitute $x = 1/2$, and we have

$$\sum_{k=0}^{n}\binom{n}{k}2^{-(n-k)}k2^{-(k-1)} = n .$$

Here we are: dividing by 2 we get

$$\sum_{k=0}^{n}\binom{n}{k}k2^{-n} = \frac{n}{2} .$$

So much for the definition of average. Here is a better route: we can view X as the sum of n random variables $X_1, \ldots, X_n$, where X_i is 1 if the i-th coin flip is 1 and X_i is 0 otherwise. Clearly, for every i, $\mathbf{E}[X_i] = \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 1 = \frac{1}{2}$, and so

$$\mathbf{E}[X] = \mathbf{E}[X_1 + \cdots + X_n] = \mathbf{E}[X_1] + \cdots + \mathbf{E}[X_n] = \frac{n}{2} .$$

■

3 Independence

3.1 Conditioning and Mutual Independence

Suppose I toss two coins, without letting you see the outcome, and I tell you that at least one of the coins came up heads, what is the probability that both coin are heads?

In order to answer to this question (I will give it away that the answer is 1/3), one needs some tools to reason about the probability that a certain event holds *given* (or *conditioned* on the fact) that a certain other event holds.

Fix a sample space Ω and a probability distribution $\mathbf{Pr}$. Suppose we are given that a certain event $A \subseteq \Omega$ holds. Then the probability of an elementary event a given the fact that A holds (written $\mathbf{Pr}(a|A)$) is defined as follows: if $a \notin A$, then it is impossible that a holds, and so $\mathbf{Pr}(a|A) = 0$; otherwise, if $a \in A$, then $\mathbf{Pr}(a|A)$ has a value that is proportional to $\mathbf{Pr}(a)$. One realizes that the factor of proportionality has to be $1/\mathbf{Pr}[A]$, so that probabilities sum to 1 again. Our definition of conditional probability of an elementary event is then

$$\mathbf{Pr}(a|A) = \begin{cases} 0 & \text{If } a \notin A \\ \frac{\mathbf{Pr}(a)}{\mathbf{Pr}[A]} & \text{Otherwise} \end{cases}$$

The above formula already lets us solve the question asked at the beginning of this section. Notice that probabilities conditioned on an event A such that $\mathbf{Pr}[A] = 0$ are undefined.

Then we extend the definition to arbitrary events, and we say that for an event B

$$\mathbf{Pr}[B|A] = \sum_{b \in B} \mathbf{Pr}(b|A)$$

One should check that the following (more standard) definition is equivalent

$$\mathbf{Pr}[B|A] = \frac{\mathbf{Pr}[A \cap B]}{\mathbf{Pr}[A]}$$

Definition 3.1 *Two events A and B are independent if*

$$\mathbf{Pr}[A \cap B] = \mathbf{Pr}[A] \cdot \mathbf{Pr}[B]$$

If A and B are independent, and $\mathbf{Pr}[A] > 0$, then we have $\mathbf{Pr}[B|A] = \mathbf{Pr}[B]$. Similarly, if A and B are independent, and $\mathbf{Pr}[B] > 0$, then we have $\mathbf{Pr}[A|B] = \mathbf{Pr}[A]$. This motivates the use of the term "independence." If A and B are independent, then whether A holds or not is not influenced by the knowledge that B holds or not.

When we have several events, we can define a generalized notion of independence.

Definition 3.2 *Let $A_1, \ldots, A_n \subseteq \Omega$ be events is a sample space Ω; we say that such events are* mutually *independent if for every subset of indices $I \subseteq \{1, \ldots n\}$, $I \neq \emptyset$, we have*

$$\mathbf{Pr}[\bigcap_{i \in I} A_i] = \prod_{i \in I} \mathbf{Pr}[A_i]$$

All this stuff was just to prepare for the definition of independence for random variables, which is a very important and useful notion.

Definition 3.3 *If X and Y are random variables over the same sample space, then we say that X and Y are independent if for any two values v, w, the event $(X = v)$ and $(Y = w)$ are independent.*

Therefore, if X and Y are independent, knowing the value of X, no matter which value it is, does not tell us noting about the distribution of Y (and vice versa).

Theorem 3.4 *If X and Y are independent, then $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$.*

This generalizes to several random variables

Definition 3.5 *Let $X_1, \ldots, X_n$ be random variables over the same sample space, then we say that they are* mutually *independent if for any sequence of values $v_1, \ldots, v_n$, the events $(X_1 = v_1)$, $\ldots$, $(X_n = v_n)$ are mutually independent.*

Theorem 3.6 *If $X_1, \ldots, X_n$ are mutually independent random variables, then*

$$\mathbf{E}[X_1 \cdot X_2 \cdots X_n] = \mathbf{E}[X_1] \cdot \mathbf{E}[X_2] \cdots \mathbf{E}[X_n]$$

3.2 Pairwise Independence

It is also possible to define a weaker notion of independence.

Definition 3.7 *Let $X_1, \ldots, X_n$ be random variables over the same sample space, then we say that they are* pairwise *independent if for every $i, j \in \{1, \ldots, n\}$, $i \neq j$, we have that X_i and X_j are independent.*

It is important to note that a collection of random variables can be pairwise independent without being mutually independent. (But a collection of mutually independent random variables is always pairwise independent for a stronger reason).

Example 2 Consider the following probabilistic system: we toss 2 coins, and we let the random vaiables X, Y, Z be, respectively, the outcome of the first coin, the outcome of the second coin, and the XOR of the outcomes of the two coins (as usual, we interpret outcomes of coins as 0/1 values). Then X, Y, Z are not mutually independent, for example

$$\mathbf{Pr}[Z = 0 | X = 0, Y = 0] = 1$$

while

$$\mathbf{Pr}[Z = 0] = 1/2$$

in fact, intuitively, since the value of Z is *totally determined* by the values of X and Y, the three variables cannot be mutually independent. On the other hand, we will now show that X, Y, Z are pairwise independent. By definition, X and Y are independent, so we have to focus of X and Z and on Y and Z. Let us prove that X and Z are independent (the proof for Y and Z is identical). We have to show that for each choice of two values $v, w \in \{0, 1\}$, we have

$$\mathbf{Pr}[X = v, Z = w] = \mathbf{Pr}[X = v]\mathbf{Pr}[Z = w] = \frac{1}{4}$$

and this is true, since, in order to have $Z = w$ and $X = v$, we must have $Y = w \oplus v$, and the event that $X = v$ and $Y = w \oplus v$ happens with probability 1/4. ■

Let us see two additional, more involved, examples.

Example 3 Suppose we flip k coins, whose outcomes be $a_1, \ldots, a_n \in \{0, 1\}^k$. Then for every non-empty subset $I \subseteq \{0, 1\}^k$ we define a random variable X_I, whose value is $\bigoplus_{i \in I} a_i$. It is possible to show that $\{X_I\}_{I \subseteq \{0,1\}^k, I \neq \emptyset}$ is a pairwise independent collection of random variables. Notice that we have $2^k - 1$ random variables defined over a sample space of only 2^k points. ■

Example 4 Let p be a prime number; suppose we pick at random two elements $a, b \in \mathbf{Z}_p$ — that is, our sample space is the set of pairs $(a, b) \in \mathbf{Z}_p \times \mathbf{Z}_p = \Omega$, and we consider the uniform distribution over this sample space. For every $z \in \mathbf{Z}_p$, we define one random variable X_z whose value is $az + b \pmod p$. Thus we have a collection of p random variables. It is possible to show that such random variables are pairwise independent. ■

4 Deviation from the Expectation

4.1 Markov's Inequality

Say that X is the random variable expressing the running time in seconds of an algorithm on inputs of a certain size, and that we computed $\mathbf{E}[X] = 10$. Since this is the order of magnitude of the time that we expect to spend while running the algorithm, it would be devastating if it happened that, say, $X \geq 1,000,000$ (i.e. more than 11 days) with large probability. However, we quickly realize that if $\mathbf{E}[X] = 10$, then it must be $\mathbf{Pr}[X \geq 1,000,000] \leq 1/100,000$, as otherwise the contribution to the expectation of the only events where $X \geq 1,000,000$ would already exceed the value 10. This reasoning can be generalized as follows.

Theorem 4.1 (Markov's Inequality) *If X is a non-negative random variable then*

$$\mathbf{Pr}[X \geq k] \leq \frac{\mathbf{E}[X]}{k}$$

Sometimes the bound given by Markov's inequality are extremely bad, but the bound is as strong as possible if the only information that we have is the expectation of X.

For example, suppose that X counts the number of heads in a sequence of n coin flips. Formally, Ω is $\{0,1\}^n$ with the uniform distribution, and X is the number of ones in the string. Then $\mathbf{E}[X] = n/2$. Suppose we want to get an upper bound on $\mathbf{Pr}[X \geq n]$ using Markov. Then we get

$$\mathbf{Pr}[X \geq n] \leq \frac{\mathbf{E}[X]}{n} = \frac{1}{2}$$

This is ridiculous! The right value is 2^{-n}, and the upper bound given by Markov's inequality is totally off, and it does not even depend on n.

However, consider now the experiment where we flip n coins that are *glued* together, so that the only possible outcomes are n heads (with probability 1/2) and n tails (with probability 1/2). Define X again as the number of heads. We still have that $\mathbf{E}[X] = n/2$, and we can apply Markov's inequality as before to get

$$\mathbf{Pr}[X \geq n] \leq \frac{\mathbf{E}[X]}{n} = \frac{1}{2}$$

But, now, the above inequality is tight, becuase $\mathbf{Pr}[X \geq n]$ is precisely 1/2.

The moral is that Markov's inequality is very useful because it applies to every non-negative random variables having a certain expectation, so we can use it without having to study our random variable too much. On the other hand, the inequality will be accurate when applied to a random variable that typically deviates a lot from its expectation (say, the number of heads that we get when we toss n glued coins) and the inequality will be very bad when we apply it to a random variable that is concentrated around its expectation (say, the number of heads that we get in n independent coin tosses). In the latter case, if we want accurate estimations we have to use more powerful methods. One such method is described below.

4.2 Variance

For a random variable X, the random variable

$$X' = |X - \mathbf{E}[X]|$$

gives all the information that we need in order to decide whether X is likely to deviate a lot from its expectation or not. All we need to do is to prove that X' is typically small. However this idea does not lead us very far (analysing X' does not seem to be any easier than analysing X).

Here is a better tool. Consider

$$(X - \mathbf{E}[X])^2$$

This is again a random variable that tells us how much X deviates from its expectation. In particular, if the *expectation* of such an auxliary random variable is small, then we expect X to be typically close to its expectation. The *variance* of X is defined as

$$\mathbf{Var}(X) = \mathbf{E}[(X - \mathbf{E}[X])^2]$$

Here is an equivalent epxression (we use linearity of expectation in the derivation of the final result):

$$\begin{aligned}
\mathbf{Var}(X) &= \mathbf{E}[(X - \mathbf{E}[X])^2] \\
&= \mathbf{E}[X^2 - 2X\mathbf{E}[X] + (\mathbf{E}[X])^2] \\
&= \mathbf{E}[X^2] - 2\mathbf{E}[X\mathbf{E}[X]] + (\mathbf{E}[X])^2 \\
&= \mathbf{E}[X^2] - 2\mathbf{E}[X]\mathbf{E}[X] + (\mathbf{E}[X])^2 \\
&= \mathbf{E}[X^2] - (\mathbf{E}[X])^2
\end{aligned}$$

The variance is a useful notion for two reasons: it is often easy to compute and it gives rise to sometimes strong estimations on the probability that a random variable deviates from its expectation.

Theorem 4.2 (Chebyshev's Inequality)

$$\mathbf{Pr}[|X - \mathbf{E}[X]| \geq k] \leq \frac{\mathbf{Var}(X)}{k^2}$$

The proof uses Markov's inequality and a bit of ingenuity.

$$\begin{aligned}\mathbf{Pr}[|X-\mathbf{E}[X]| \geq k] &= \mathbf{Pr}[(X-\mathbf{E}[X])^2 \geq k^2] \\ &\leq \frac{\mathbf{E}[(X-\mathbf{E}[X])^2]}{k^2} \\ &= \frac{\mathbf{Var}(X)}{k^2}\end{aligned}$$

The nice idea is in the first step. The second step is just an application of Markov's inequality and the last step uses the definition of variance.

The value $\sigma(X) = \sqrt{\mathbf{Var}(X)}$ is called the *standard deviation* of X. One expects the value of a random variable X to be around the interval $\mathbf{E}[X] \pm \sigma(X)$. We can restate Chebyshev's Inequality in terms of standard deviation

Theorem 4.3 (Chebyshev's Inequality, Alternative Form)

$$\mathbf{Pr}[|X-\mathbf{E}[X]| \geq c \cdot \sigma(X)] \leq \frac{1}{c^2}$$

Let Y be a random variable that is equal to 0 with probability 1/2 and to 1 with probability 1/2. Then $\mathbf{E}[Y] = 1/2$, $Y = Y^2$, and

$$\mathbf{Var}(Y) = \mathbf{E}[Y^2] - (\mathbf{E}[Y])^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

Let X the random variable that counts the number of heads in a sequence of n independent coin flips. We have seen that $\mathbf{E}[X] = n/2$. Computing the variance according to the definition would be painful. We are fortunate that the following result holds.

Lemma 4.4 (Tools to Compute Variance)

1. *Let X be a random variable, a, b be reals, then*

$$\mathbf{Var}(aX + b) = a^2 \mathbf{Var}(X)$$

2. *Let $X_1, \ldots, X_n$ be* pairwise *independent random variables on the same sample space. Then*

$$\mathbf{Var}(X_1 + \cdots + X_n) = \mathbf{Var}(X_1) + \cdots + \mathbf{Var}(X_n)$$

Then we can view X as $X_1 + \cdots + X_n$ where X_i are mutually independent random variables such that for each i X_i takes value 1 with probability 1/2 and value 0 with probability 1/2. As computed before, $\mathbf{Var}(X_i) = 1/4$. Therefore $\mathbf{Var}(X) = n/4$ and the standard deviation is $\sqrt{n}/2$. This means that when we flip n coins we expect to get about $n \pm \sqrt{n}$ heads.

Let us test Chebyshev's inequality on the same example of the previous subsection. Let X be a random variable defined over $\Omega = \{0,1\}^n$, where $\mathbf{Pr}$ is uniform, and X counts the number of 1s in the elementary event: suppose we want to compute $\mathbf{Pr}[X \geq n]$. As computed above, $\mathbf{Var}(X) = n/4$, so

$$\mathbf{Pr}[X \geq n] \leq \mathbf{Pr}[|X - \mathbf{E}[X]| \geq n/2] \leq \frac{1}{n}$$

This is still much less than the correct value 2^{-n}, but at least it is a value that decreases with n. It is also possible to show that Chebyshev's inequality is as strong as possible given its assumption.

Let $n = 2^k - 1$ for some integer k and let $X_1, \ldots, X_n$ be the collection of pairwise independent random variables as defined in Example 3. Let $X = X_1 + \ldots + X_n$. Suppose we want to compute $\mathbf{Pr}[X = 0]$. Since each X_i has variance $1/4$, we have that X has variance $n/4$, and so

$$\mathbf{Pr}[X = 0] \leq \mathbf{Pr}[|X - \mathbf{E}[X]| \geq n/2] \leq \frac{1}{n}$$

which is almost the right value: the right value is $2^{-k} = 1/(n+1)$.

A Appendix

A.1 Some Combinatorial Facts

Consider a set Ω with n elements. Ω has 2^n subsets (including the empty set and Ω itself).

For every $0 \leq k \leq n$, Ω has $\binom{n}{k}$ subsets of k elements. The symbol $\binom{n}{k}$ is read "n choose k" and is defined as

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Then we must have

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n \tag{2}$$

which is a special case of the following result

Theorem A.1 (Binomial Theorem) *For every two reals a, b and non-negative integer n,*

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

We can see that Equation (2) follows from the Binomial Theorem by simply substituting $a = 1$ and $b = 1$.

Sometimes we have to deal with summations of the form $1 + 1/2 + 1/3 + \ldots + 1/n$. It's good to know that $\sum_{k=1}^{n} 1/k \approx \ln n$. More precisely

Theorem A.2 $\lim_{n\to\infty} \frac{\sum_{k=1}^{n} 1/k}{\ln n} = 1$.

In particular, $\sum_{k=1}^{n} 1/k \leq 1 + \ln n$ for every n, and $\sum_{k=1}^{n} 1/k \geq \ln n$ for sufficiently large n.

The following inequality is exceedingly useful in computing upper bounds of probabilities of events:

$$1 + x \leq e^x \tag{3}$$

This is easy to prove by looking at the Taylor series of e^x:

$$e^x = 1 + x + \frac{1}{2}x^2 + \ldots + \frac{1}{k!}x^k + \ldots$$

Observe that Equation (3) is true for *every* real x, not necessarily positive (but it becomes trivial for $x < -1$).

Here is a typical application of Equation (3). We have a randomized algorithm that has a probability ϵ over its internal coin tosses of succeeding in doing something (and when it succeeds, we notice that it does, say because the algorithm is trying to invert a one-way function, and when it succeeds then we can check it efficiently); how many times do we have to run the algorithm before we have probability at least $3/4$ that the algorithm succeeds?

The probability that it never succeeds in k runs is

$$(1 - \epsilon)^k \leq e^{-\epsilon k}$$

If we choose $k = 2/\epsilon$, the probability of k consecutive failures is less than $e^{-2} < 1/4$, and so the probability of succeeding (at least once) is at least $3/4$.

A.2 Examples of Analysis of Error Probability of Algorithms

Example 5 Suppose that we have an algorithm whose worst-case running time (on inputs of a certain length) is bounded by a random variable T (whose sample space is the set of random choices made by the algorithm). For concreteness, suppose that we are considering the randomized algorithm that given a prime p and an element $a \in \mathbf{Z}^*{}_p$ decides whether a is a quadratic residue or not. Suppose that we are given $t = \mathbf{E}[T]$ but no additional information on the algorithm, and we would like to know how much time we have to wait in oder to have a probability at least $1 - 10^{-6}$ that the algorithm terminates. If we only know $\mathbf{E}[T]$, then we can just use Markov's inequality and say that

$$\mathbf{Pr}[T \geq kt] \leq \frac{1}{k}$$

and if we choose $k = 10^6$ we have that

$$\mathbf{Pr}[T \geq 10^6 t] \leq 10^{-6} .$$

However there is a much faster way of guaranteeing termination with high probability. We let the program run for $2t$ time. There is a probability $1/2$ that the algorithm will stop before that time. If so we are happy. If not, we terminate the computation, and start it over (in the second iteration, we let the algorithm use independent random bits). If the second computation does not terminate within $2t$ time, we reset it once more, and so on.[1] Let T' be the random variable that gives the time taken by this new version of the algorithm (with the stop and reset actions). Now we have that the probability that we use more than $2kt$ time is equal to the probability that for k consecutive (independent) times the algorithm takes more than $2t$ time. Each of these events happen with probability at most $1/2$, and so

$$\mathbf{Pr}[T' \geq 2kt] \leq 2^{-k}$$

and if take $k = 20$, the probability is less than 10^{-6}, and the time is only $40t$ rather than $1,000,000t$.

Suppose that $t = t(n)$ is the average running time of our algorithm on inputs of length n, and that we want to find another algorithm that finishes always in time $t'(n)$ and that reports a failure only with negligible probability, say with probability at most $n^{-\log n}$. How large do we have choose t', and what the new algorithm should be like?

If we just put a timeout t' on the original algorithm, then we can use Markov's inequality to say that $t'(n) = n^{\log n} t(n)$ will suffice, but now t' is not polynomial in n (even if t was). Using the second method, we can put a timeout $2t$ and repeat the algorithm $(\log n)^2$ times. Then the failure probability will be as requested and $t'(n) = 2(\log n)^2 t(n)$. ■

If we know how the algorithm works, then we can make a more direct analysis.

Example 6 Suppose that our goal is, given n, to find a number $2 \leq a \leq n-1$ such that $\gcd(a, n) = 1$. To simplify notation, let $l = ||n|| \approx \log n$ be the number of digits of n in binary notation (in a concrete application, l would be a few hundreds). Our algorithm will be as follows:

- Repeat no more than k times:

    1. Pick uniformly at random $a \in \{2, \ldots, n-1\}$;
    2. Use Euclid's algorithm to test whether $\gcd(a, n) = 1$.

[1] This is reminescent of the way one works with Windows'98.

3. If $\gcd(a, n) = 1$ then output a and halt.

- Output "failure".

We would like to find a value of k such that the probability that the algorithm reports a failure is negligible in the size of the input (i.e. in l).

At each iteration, the probability that algorithm finds an element that is coprime with n is

$$\frac{\phi(n)}{n-2} \geq \frac{1}{6 \log \log n} = \frac{1}{6 \log l}$$

So the probability that there is a failure in one iteration is at most

$$\left(1 - \frac{1}{6 \log l}\right)$$

and the probability of k consecutive independent failures is at most

$$\left(1 - \frac{1}{6 \log l}\right)^k \leq e^{-k/6 \log l}$$

if we set $k = (\log l)^3$ then the probability of k consecutive failures is at most

$$e^{-(\log l)^3/6 \log l} = l^{-(\log l)/6}$$

that is negligible in l.

$\blacksquare$