Basics of Probability Theory
(for Theory of Computation courses)

Oded Goldreich
Department of Computer Science
Weizmann Institute of Science
Rehovot, ISRAEL.
oded.goldreich@weizmann.ac.il

November 24, 2008

Preface. Probability theory plays a central role in many areas of computer science, and specifically in cryptography and complexity theory. In this text, we present the basic probabilistic notions and notations that are used in various courses in the theory of computation. Specifically, we refer to the notions of discrete probability space and random variables, and to corresponding notations. Also included are overviews of three useful probabilistic inequalities: Markov's Inequality, Chebyshev's Inequality, and Chernoff Bound.

1 The Very Basics

Formally speaking, probability theory is merely a quantitative study of the relation between the sizes of various sets. In the simple case, which underlies our applications, these sets are finite and the "probability of an event" is merely a shorthand for the density of a corresponding set with respect to an underlying (finite) universe. For example, when we talk of the probability that a coin flip yields the outcome `heads`, the universe is the set of the two possible outcomes (i.e., `heads` and `tails`) and the event we refer to is a subset of this universe (i.e., the singleton set `heads`). In general, the universe corresponds to all possible "basic" situations (which are assumed or postulated to be equally likely), and events correspond to specific sets of some of these situations. Thus, one may claim that probability theory is just combinatorics; however, as is often the case, good notations (let alone notions) are instrumental to more complex studies.

Throughout the entire text we refer only to *discrete* probability distributions. Such probability distributions refer to a finite universe, called the probability space (or the sample space), and to subsets of this universe. Specifically, the underlying probability (or sample) space consists of a *finite* set, denoted Ω, and events correspond to subsets of Ω. The probability of an event $A \subseteq \Omega$ is defined as $|A|/|\Omega|$. Indeed, one should think of probabilities by referring to a (mental) experiment in which an element in the space Ω is selected with uniform probability distribution (i.e., each element is selected with equal probability, $1/|\Omega|$).

Random variables. Traditionally, random variables are defined as functions from the sample space to the reals. (For example, for any $A \subseteq \Omega$, we may consider a random variable $X : \Omega \to \mathsf{R}$ such that $X(e) = 1$ of $e \in A$ and $X(e) = 0$ otherwise.) The probability that X is assigned a particular value v is denoted $\Pr[X=v]$ and is defined as the probability of the event $\{e \in \Omega : X(e)\}$; that is, $\Pr[X\!=\!v] = |\{e \in \Omega : X(e)\!=\!v\}|/|\Omega|$. The support of a random variable is the (finite) set of values that are assigned positive probability; that is, the support of X is the set $\{X(e) : e \in \Omega\}$.

The expectation of a random variable X, defined over Ω, is defined as the average value of X when the underlying sample is selected uniformly in Ω (i.e., $\Omega|^{-1} \cdot \sum_{e\in\Omega} X(e)$). We denote this expectation by $\mathsf{E}[X]$. A key observation, referred to as *linearity of expectation*, is that for any two random variables $X, Y : \Omega \to \mathsf{R}$ it holds that $\mathsf{E}[X+Y] = \mathsf{E}[X] + \mathsf{E}[Y]$. In particular, for any constants a and b, it holds that $\mathsf{E}[aX+b] = a\mathsf{E}[X] + b$.

The variance of a random variable X, denoted $\mathsf{Var}[X]$, is defined as $\mathsf{E}[(X - \mathsf{E}[X])^2]$. Note that $\mu \stackrel{\rm def}{=} \mathsf{E}[X]$ is a constant, and thus

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

which is upper bounded by $\mathsf{E}[X^2]$. Note that a random variable has positive variance if and only if it is not a constant (i.e., $\mathsf{Var}[X] > 0$ if and only if X assumes at least two different values).

It is common practice to talk about random variables without specifying the underlying probability space. In these cases a random varibale is viewed as a discrete probability distribution over the reals; that is, we fix the probability that this random variable is assigned any value. Assuming that this probabilities are all rationales, a corresponding probability space always exists. For example, when we consider a random variable X such that $\Pr[X\!=\!1] = 1/3$ and $\Pr[X\!=\!2] = 2/3$, a possible corresponding probability space is $\{1,2,3\}$ such that $X(1) = 1$ and $X(2) = X(3) = 2$.

Statistical difference. The statistical distance (a.k.a variation distance) between the random variables X and Y is defined as

$$\frac{1}{2} \cdot \sum_v |\Pr[X=v] - \Pr[Y=v]| \;=\; \max_S\{\Pr[X \in S] - \Pr[Y \in S]\}. \tag{1}$$

We say that X is δ-close (resp., δ-far) to Y if the statistical distance between them is at most (resp., at least) δ.

Our Notational Conventions

Typically, in our courses, the underlying probability space will consist of the set of all strings of a certain length ℓ. That is, the sample space is the set of all ℓ-bit long strings, and each such string is assigned probability measure $2^{-\ell}$.

Abusing the traditional terminology, we use the term random variable also when referring to functions mapping the sample space into the set of binary strings. We often do not specify the probability space, but rather talk directly about random variables. For example, we may say that X is a random variable assigned values in the set of all strings such that $\Pr[X=00] = \frac{1}{4}$ and $\Pr[X\!=\!111] = \frac{3}{4}$. (Such a random variable may be defined over the sample space $\{0,1\}^2$, so that $X(11) = 00$ and $X(00) = X(01) = X(10) = 111$.) One important case of a random variable is the output of a randomized process (e.g., a probabilistic polynomial-time algorithm).

All our probabilistic statements refer to random variables that are defined beforehand. Typically, we may write $\Pr[f(X)\!=\!1]$, where X is a random variable defined beforehand (and f is a function). An important convention is that *all occurrences of the same symbol in a probabilistic statement refer to the same* (unique) *random variable.* Hence, if $B(\cdot,\cdot)$ is a Boolean expression depending on two variables, and X is a random variable then $\Pr[B(X,X)]$ denotes the probability that $B(x,x)$ holds when x is chosen with probability $\Pr[X=x]$. For example, for every random variable X, we have $\Pr[X=X]=1$. We stress that if we wish to discuss the probability that $B(x,y)$ holds when x and y are chosen independently with identical probability distribution, then we will define *two* independent random variables each with the same probability distribution.[1] Hence, if X and Y are two independent random variables, then $\Pr[B(X,Y)]$ denotes the probability that $B(x,y)$ holds when the pair (x,y) is chosen with probability $\Pr[X=x]\cdot\Pr[Y=y]$. For example, for every two independent random variables, X and Y, we have $\Pr[X=Y]=1$ only if both X and Y are trivial (i.e., assign the entire probability mass to a single string).

We will often use U_n to denote a random variable uniformly distributed over the set of all strings of length n. Namely, $\Pr[U_n=\alpha]$ equals 2^{-n} if $\alpha\in\{0,1\}^n$ and equals 0 otherwise. We often refer to the distribution of U_n as the uniform distribution (neglecting to qualify that it is uniform over $\{0,1\}^n$). In addition, we occasionally use random variables (arbitrarily) distributed over $\{0,1\}^n$ or $\{0,1\}^{\ell(n)}$, for some function $\ell:\mathsf{N}\rightarrow\mathsf{N}$. Such random variables are typically denoted by X_n, Y_n, Z_n, etc. We stress that in some cases X_n is distributed over $\{0,1\}^n$, whereas in other cases it is distributed over $\{0,1\}^{\ell(n)}$, for some function ℓ (which is typically a polynomial). We often talk about probability ensembles, which are infinite sequence of random variables $\{X_n\}_{n\in\mathsf{N}}$ such that each X_n ranges over strings of length bounded by a polynomial in n.

2 Three Inequalities

The following probabilistic inequalities are very useful. These inequalities refer to random variables that are assigned real values and provide upper-bounds on the probability that the random variable deviates from its expectation.

2.1 Markov's Inequality

The most basic inequality is Markov's Inequality that applies to any random variable with bounded maximum or minimum value. For simplicity, this inequality is stated for random variables that are lower-bounded by zero, and reads as follows: *Let X be a non-negative random variable and v be a non-negative real number. Then*

$$\Pr\left[X\geq v\right]\leq\frac{\mathsf{E}(X)}{v} \tag{2}$$

Equivalently, $\Pr[X\geq r\cdot\mathsf{E}(X)]\leq\frac{1}{r}$. The proof amounts to the following sequence:

$$\begin{aligned}\mathsf{E}(X) &= \sum_x \Pr[X=x]\cdot x\\ &\geq \sum_{x<v}\Pr[X=x]\cdot 0+\sum_{x\geq v}\Pr[X=x]\cdot v\\ &= \Pr[X\geq v]\cdot v\end{aligned}$$

[1] Two random variables, X and Y, are called independent if for every pair of possible values (x,y) it holds that $\Pr[(X,Y)=(x,y)]=\Pr[X=x]\cdot\Pr[Y=y]$.

2.2 Chebyshev's Inequality

Using Markov's inequality, one gets a potentially stronger bound on the deviation of a random variable from its expectation. This bound, called Chebyshev's inequality, is useful when having additional information concerning the random variable (specifically, a good upper bound on its variance). For a random variable X of finite expectation, we denote by $\mathsf{Var}(X) \stackrel{\rm def}{=} \mathsf{E}[(X-\mathsf{E}(X))^2]$ the variance of X, and observe that $\mathsf{Var}(X) = \mathsf{E}(X^2) - \mathsf{E}(X)^2$. Chebyshev's Inequality then reads as follows: *Let X be a random variable, and $\delta > 0$. Then*

$$\Pr\left[|X-\mathsf{E}(X)| \geq \delta\right] \;\leq\; \frac{\mathsf{Var}(X)}{\delta^2}. \tag{3}$$

Proof: We define a random variable $Y \stackrel{\rm def}{=} (X-\mathsf{E}(X))^2$, and apply Markov's inequality. We get

$$\begin{aligned}\Pr\left[|X-\mathsf{E}(X)| \geq \delta\right] &= \Pr\left[(X-\mathsf{E}(X))^2 \geq \delta^2\right] \\ &\leq \frac{\mathsf{E}[(X-\mathsf{E}(X))^2]}{\delta^2}\end{aligned}$$

and the claim follows. ■

Corollary (Pairwise Independent Sampling)**:** Chebyshev's inequality is particularly useful in the analysis of the error probability of approximation via repeated sampling. It suffices to assume that the samples are picked in a pairwise independent manner, where $X_1, X_2, ..., X_n$ are pairwise independent if for every $i \neq j$ and every α, β it holds that $\Pr[X_i = \alpha \wedge X_j = \beta] = \Pr[X_i = \alpha] \cdot \Pr[X_j = \beta]$. The corollary reads as follows: *Let $X_1, X_2, ..., X_n$ be pairwise independent random variables with identical expectation, denoted μ, and identical variance, denoted σ^2. Then, for every $\epsilon > 0$, it holds that*

$$\Pr\left[\left|\frac{\sum_{i=1}^n X_i}{n} - \mu\right| \geq \epsilon\right] \leq \frac{\sigma^2}{\epsilon^2 n}. \tag{4}$$

Proof: Define the random variables $\overline{X}_i \stackrel{\rm def}{=} X_i - \mathsf{E}(X_i)$. Note that the $\overline{X}_i$'s are pairwise independent, and each has zero expectation. Applying Chebyshev's inequality to the random variable $\sum_{i=1}^n \frac{X_i}{n}$, and using the linearity of the expectation operator, we get

$$\begin{aligned}\Pr\left[\left|\sum_{i=1}^n \frac{X_i}{n} - \mu\right| \geq \epsilon\right] &\leq \frac{\mathsf{Var}\left[\sum_{i=1}^n \frac{X_i}{n}\right]}{\epsilon^2} \\ &= \frac{\mathsf{E}\left[\left(\sum_{i=1}^n \overline{X}_i\right)^2\right]}{\epsilon^2 \cdot n^2}\end{aligned}$$

Now (again using the linearity of expectation)

$$\mathsf{E}\left[\left(\sum_{i=1}^n \overline{X}_i\right)^2\right] = \sum_{i=1}^n \mathsf{E}\left[\overline{X}_i^2\right] + \sum_{1 \leq i \neq j \leq n} \mathsf{E}\left[\overline{X}_i \overline{X}_j\right]$$

By the pairwise independence of the $\overline{X}_i$'s, we get $\mathsf{E}[\overline{X}_i \overline{X}_j] = \mathsf{E}[\overline{X}_i] \cdot \mathsf{E}[\overline{X}_j]$, and using $\mathsf{E}[\overline{X}_i] = 0$, we get

$$\mathsf{E}\left[\left(\sum_{i=1}^n \overline{X}_i\right)^2\right] = n \cdot \sigma^2$$

The corollary follows. ■

2.3 Chernoff Bound

When using pairwise independent sample points, the error probability in the approximation decreases linearly with the number of sample points (see Eq. (4)). When using totally independent sample points, the error probability in the approximation can be shown to decrease exponentially with the number of sample points. (Recall that the random variables $X_1, X_2, ..., X_n$ are said to be totally independent if for every sequence $a_1, a_2, ..., a_n$ it holds that $\Pr[\wedge_{i=1}^n X_i = a_i] = \prod_{i=1}^n \Pr[X_i = a_i]$.) Probability bounds supporting the foregoing statement are given next. The first bound, commonly referred to as Chernoff Bound, concerns 0-1 random variables (i.e., random variables that are assigned as values either 0 or 1), and asserts the following. *Let $p \leq \frac{1}{2}$, and $X_1, X_2, ..., X_n$ be independent 0-1 random variables such that $\Pr[X_i = 1] = p$, for each i. Then, for every $\epsilon \in (0, p]$, it holds that*

$$\Pr\left[\left|\frac{\sum_{i=1}^n X_i}{n} - p\right| > \epsilon\right] \; < \; 2 \cdot e^{-c \cdot \epsilon^2 \cdot n}, \text{ where } c = \max(2, \tfrac{1}{3p}). \tag{5}$$

The more common formulation sets $c = 2$, but the case $c = 1/3p$ is very useful when p is small and one cares about a *multiplicative* deviation (e.g., $\epsilon = p/2$).

Proof Sketch: We upper-bound $\Pr[\sum_{i=1}^n X_i - pn > \epsilon n]$, and $\Pr[pn - \sum_{i=1}^n X_i > \epsilon n]$ is bounded similarly. Letting $\overline{X}_i \overset{\text{def}}{=} X_i - \mathsf{E}(X_i)$, we apply Markov's inequality to the random variable $e^{\lambda \sum_{i=1}^n \overline{X}_i}$, where $\lambda \in (0, 1]$ will be determined to optimize the expressions that we derive. Thus, $\Pr[\sum_{i=1}^n \overline{X}_i > \epsilon n]$ is upper-bounded by

$$\frac{\mathsf{E}[e^{\lambda \sum_{i=1}^n \overline{X}_i}]}{e^{\lambda \epsilon n}} = e^{-\lambda \epsilon n} \cdot \prod_{i=1}^n \mathsf{E}[e^{\lambda \overline{X}_i}]$$

where the equality is due to the independence of the random variables. To simplify the rest of the proof, we establish a sub-optimal bound as follows. Using a Taylor expansion of e^x (e.g., $e^x < 1 + x + x^2$ for $|x| \leq 1$) and observing that $\mathsf{E}[\overline{X}_i] = 0$, we get $\mathsf{E}[e^{\lambda \overline{X}_i}] < 1 + \lambda^2 \mathsf{E}[\overline{X}_i^2]$, which equals $1 + \lambda^2 p(1-p)$. Thus, $\Pr[\sum_{i=1}^n X_i - pn > \epsilon n]$ is upper-bounded by $e^{-\lambda \epsilon n} \cdot (1 + \lambda^2 p(1-p))^n < \exp(-\lambda \epsilon n + \lambda^2 p(1-p)n)$, which is optimized at $\lambda = \epsilon/(2p(1-p))$ yielding $\exp(-\frac{\epsilon^2}{4p(1-p)} \cdot n) \leq \exp(-\epsilon^2 \cdot n)$. $\square$

The foregoing proof strategy can be applied in more general settings.[2] A more general bound, which refers to independent random variables that are each bounded but are not necessarily identical, is given next (and is commonly referred to as Hoefding Inequality). *Let $X_1, X_2, ..., X_n$ be n independent random variables, each ranging in the (real) interval $[a, b]$, and let $\mu \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \mathsf{E}(X_i)$ denote the average expected value of these variables. Then, for every $\epsilon > 0$,*

$$\Pr\left[\left|\frac{\sum_{i=1}^n X_i}{n} - \mu\right| > \epsilon\right] < 2 \cdot e^{-\frac{2\epsilon^2}{(b-a)^2} \cdot n} \tag{6}$$

The special case (of Eq. (6)) that refers to identically distributed random variables is easy to derive from the foregoing Chernoff Bound (by recalling Footnote 2 and using a linear mapping of the interval $[a, b]$ to the interval $[0, 1]$). This special case is useful in estimating the average value of a (bounded) function defined over a large domain, especially when the desired error probability needs to be negligible (i.e., decrease faster than any polynomial in the number of samples). Such an estimate can be obtained provided that we can sample the function's domain (and evaluate the function).

[2] For example, verify that the current proof actually applies to the case that $X_i \in [0, 1]$ rather than $X_i \in \{0, 1\}$, by noting that $\mathsf{Var}[X_i] \leq p(1-p)$ still holds.

2.4 Pairwise independent versus totally independent sampling

In Sections 2.2 and 2.3 we considered two "Laws of Large Numbers" that assert that, when sufficiently many trials are made, the average value obtained in these *actual* trials typically approaches the *expected* value of a trial. In Section 2.2 these trials were performed based on pairwise independent samples, whereas in Section 2.3 these trials were performed based on totally independent samples. In this section we shall see that the *amount of deviation* (of the average from the expectation) is approximately the same in both cases, but the *probability of deviation* is much smaller in the latter case.

To demonstrate the difference between the sampling bounds provided in Sections 2.2 and 2.3, we consider the problem of estimating the average value of a function $f : \Omega \to [0,1]$. In general, we say that a random variable Z provides an (ϵ, δ)-approximation of a value v if $\Pr[|Z - v| > \epsilon] \ \leq \ \delta$.

- By Eq. (6), the average value of f evaluated at $n = O((\epsilon^{-2} \cdot \log(1/\delta))$ *independent* samples (selected uniformly in Ω) yield an (ϵ, δ)-approximation of $\mu = \sum_{x \in \Omega} f(x)/|\Omega|$. Thus, the number of sample points is polynomially related to ϵ^{-1} and logarithmically related to δ^{-1}.
- In contrast, by Eq. (4), an (ϵ, δ)-approximation by n *pairwise independent* samples calls for setting $n = O(\epsilon^{-2} \cdot \delta^{-1})$.

We stress that, *in both cases the number of samples is polynomially related to the desired accuracy of the estimation* (i.e., ϵ). *The only advantage of totally independent samples over pairwise independent ones is in the dependency of the number of samples on the error probability* (i.e., δ).