Probabilistic Preliminaries for Lecture Notes on Property Testing

Oded Goldreich*

October 10, 2015

Summary: This appendix presents background from probability theory, which will be used extensively in the lecture notes. This background and preliminaries include conventions regarding random variables, basic notions and facts, and three useful probabilistic inequalities (i.e., Markov's Inequality, Chebyshev's Inequality, and Chernoff Bound).

1 Notational Conventions

We assume that the reader is familiar with the basic notions of probability theory. In this section, we merely present the probabilistic notations that will be used extensively in the lecture notes.

Throughout the entire text we refer only to *discrete* probability distributions. Specifically, the underlying probability space consists of the set of all strings of a certain length ℓ, taken with uniform probability distribution. That is, the sample space is the set of all ℓ-bit long strings, and each such string is assigned probability measure $2^{-\ell}$. Traditionally, *random variables* are defined as functions from the sample space to the reals. 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. One important case of a random variable is the output of a randomized process (e.g., a probabilistic oracle machine).

We often do not specify the probability space, but rather talk directly about random variables. For example, we may say that X is a 0-1 random variable such that $\mathbf{Pr}[X=0] = \frac{1}{4}$ and $\mathbf{Pr}[X = 1] = \frac{3}{4}$, without specifying the underlying probability space. Indeed, this random variable may be defined over the sample space $\{0,1\}^2$, such that $X(11) = 0$ and $X(00) = X(01) = X(10) = 1$.)

Many probabilistic statements refer to random variables that are defined beforehand. Typically, we may write $\mathbf{Pr}[\chi(X)]$, where X is a random variable defined beforehand and χ is a predicate (e.g., we may write $f(X) = v$, when $f:\mathbb{R}\to\mathbb{R}$ is a function and $v \in \mathbb{R}$). In other cases, we may write $\mathbf{Pr}_{x\sim D}[\chi(x)]$, meaning that x is drawn according to a predetermined distribution D. In case D is the uniform distribution over some finite set S, we may write $\mathbf{Pr}_{x\in S}[\chi(x)]$ instead of $\mathbf{Pr}_{x\sim D}[\chi(x)]$.

2 Some basic notions and facts

We shall often use the following notions and facts.

*Department of Computer Science, Weizmann Institute of Science, Rehovot, Israel.

Union bound. An obvious fact regarding finite sets is that the size of their union is upper-bounded by the sum of their sizes; that is, if $S_1, ..., S_t$ are finite sets, then $|\cup_{i\in[t]} S_i| \leq \sum_{i\in[t]} |S_i|$. It follows that

$$\mathbf{Pr}_{r\in U}[r \in \cup_{i\in[t]} S_i] \leq \sum_{i\in[t]} \mathbf{Pr}_{r\in U}[r \in S_i],$$

where $S_1, ..., S_t \subseteq U$. Recalling that events over a probability space are merely subsets of that space, and considering the events $E_1, ..., E_t$, it holds that $\mathbf{Pr}[\vee_{i\in[t]} E_i] \leq \sum_{i\in[t]} \mathbf{Pr}[E_i]$.

Independent random variables. A sequence of random variables, $X_1, ..., X_n$, is called independent if for every $x_1, ..., x_n$ it holds that

$$\mathbf{Pr}[(X_1, ..., X_n) = (x_1, ..., x_n)] = \prod_{i\in[n]} \mathbf{Pr}[X_i = x_i].$$

This is often written in terms of conditional probabilities; namely, by writing $\mathbf{Pr}[X_1 = x_1|(X_2, ..., X_n) = (x_2, ..., x_n)] = \mathbf{Pr}[X_1 = x_1]$, which implies $\mathbf{Pr}[(X_2, ..., X_n) = (x_2, ..., x_n)|X_1 = x_1] = \mathbf{Pr}[(X_2, ..., X_n) = (x_2, ..., x_n)|X_1 = x_1]$. The latter assertion is based on Bayes' Law, which asserts that

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

since both sides equal $\frac{\mathbf{Pr}[A\&B]}{\mathbf{Pr}[B]}$.

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

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

(The equality can be verified by considering the set $S = \{v : \mathbf{Pr}[X = v] > \mathbf{Pr}[Y = v]\}$.) 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. A useful fact is that statistical distance may only decease when the same function (or even the same random process) is applied to both random variables.

Claim 1 (statistical distance is non-increasing): *Let X and Y be random variables X and Y, and A be an arbitrary randomized algorithm. Then, the statistical distance between $A(X)$ and $A(Y)$ is upper-bounded by the statistical distance between X and Y.*

Proof: We first prove the claim for a deterministic algorithm or rather any function, denoted f. In that case

$$\begin{aligned}
\sum_v |\mathbf{Pr}[f(X) = v] - \mathbf{Pr}[f(Y) = v]| &= \sum_v \left| \sum_{z\in f^{-1}(v)} \mathbf{Pr}[X = z] - \sum_{z\in f^{-1}(v)} \mathbf{Pr}[Y = z] \right| \\
&\leq \sum_v \sum_{z\in f^{-1}(v)} |\mathbf{Pr}[X = z] - \mathbf{Pr}[Y = z]| \\
&= \sum_z |\mathbf{Pr}[X = z] - \mathbf{Pr}[Y = z]|
\end{aligned}$$

We next observe that the statistical distance is preserved when appending an independent random variable to a given pair of random variables; that is, let Z be a random variable independent of both X and Y, then

$$\begin{aligned}\sum_{v,w} |\mathbf{Pr}[(X,Z)=(v,w)] - \mathbf{Pr}[(Y,Z)=(v,w)]| &= \sum_{v,w} |\mathbf{Pr}[X=v]\cdot\mathbf{Pr}[Z=w] - \mathbf{Pr}[Y=v]\cdot\mathbf{Pr}[Z=w]| \\ &= \sum_{v,w} \mathbf{Pr}[Z=w]\cdot|\mathbf{Pr}[X=v]-\mathbf{Pr}[Y=v]| \\ &= \sum_{v} |\mathbf{Pr}[X=v]-\mathbf{Pr}[Y=v]|\end{aligned}$$

Finally, letting $f(z,r)$ denote the output of a randomized algorithm A on input z when using internal coins r, we observe that the random variable $A(z)$ is represented by $f(z,R)$, where R is a random variable representing the internal coin tosses of A. Denoting the statistical distance by Δ, we have

$$\begin{aligned}\Delta(A(X),A(Y)) &= \Delta(f(X,R),f(Y,R)) \\ &\leq \Delta((X,R),(Y,R)) \\ &= \Delta(X,Y)\end{aligned}$$

establishing the claim. ■

3 Basic facts regarding expectation and variance

Throughout the rest of this appendix, we refer to discrete random variables that are assigned real values. We first recall these two standard notions.

Definition 2 (expectation and variance): *The* expectation *of a random variable $X\in\mathbb{R}$, denoted $\mathbf{E}[X]$, is defined as $\sum_{x\in\mathbb{R}}\mathbf{Pr}[X=x]\cdot x$, and its* variance*, denoted $\mathbf{V}[X]$, is defined as $\mathbf{E}[(X-\mathbf{E}[X])^2]$.*

Note that since we confine ourselves to discrete (and so finite) probability distributions, the expectation and variance can always be defined. This is best seen by replacing the summation over $\mathbb{R}$ by a summation over the support of X (i.e., the set of values v such that $\mathbf{Pr}[X=v]>0$). Three useful facts that we often use without reference follow.

Fact 1: Linearity of expectation. For every sequence of (possibly dependent) random variables, $X_1,...,X_n$, it holds that

$$\mathbf{E}\left[\sum_{i\in[n]} X_i\right] = \sum_{i\in[n]}\mathbf{E}[X_i].$$

This holds by commutativity of summation.

Fact 2: Variance and the expectation of the square. $\mathbf{V}[X]=\mathbf{E}[X^2]-\mathbf{E}[X]^2$.

This follows by $\mathbf{E}[(X-\mathbf{E}[X])^2]=\mathbf{E}[X^2-2\cdot\mathbf{E}[X]\cdot X+\mathbf{E}[X]^2]$ and linearity of expectation.

Fact 3: Functions of independent random variables are independent. If $X_1, ..., X_n$ are independent random variables, then for every sequence of functions $f_1, ..., f_n : \mathbb{R} \to \mathbb{R}$ it holds that $f_1(X_1), ..., f_n(X_n)$ are independent random variables.

This holds by definition. Specifically, for every $(v_1, ..., v_n)$, consider all $(x_1, ..., x_n) \in (f_1^{-1}(v_1) \times \cdots \times f_n^{-1}(v_n))$.

The following two additional facts will be used in this appendix, but we shall not use them outside it.

Fact 4: The expactation of the product of independent random variables. For every sequence of independent random variables $X_1, ..., X_n$, it holds that

$$\mathbf{E}\left[\prod_{i\in[n]} X_i\right] = \prod_{i\in[n]} \mathbf{E}[X_i].$$

This holds by distributivity of multiplication.

Fact 5: Linearity of the variance of independent random variables. For every sequence of independent random variables $X_1, ..., X_n$, it holds that

$$\mathbf{V}\left[\sum_{i\in[n]} X_i\right] = \sum_{i\in[n]} \mathbf{V}[X_i].$$

This can be shown by letting $\overline{X}_i = X_i - \mathbf{E}[X_i]$, and using

$$\begin{aligned}
\mathbf{V}\left[\textstyle\sum_{i\in[n]} X_i\right] &= \mathbf{E}\left[\left(\textstyle\sum_{i\in[n]} \overline{X}_i\right)^2\right] && \text{[Fact 1]}\\
&= \mathbf{E}\left[\textstyle\sum_{i,j\in[n]} \overline{X}_i\overline{X}_j\right]\\
&= \textstyle\sum_{i,j\in[n]} \mathbf{E}[\overline{X}_i\overline{X}_j] && \text{[Fact 1]}\\
&= \textstyle\sum_{i\in[n]} \mathbf{E}[\overline{X}_i^2] + \sum_{i,j\in[n]:i\neq j} \mathbf{E}[\overline{X}_i\overline{X}_j]\\
&= \textstyle\sum_{i\in[n]} \mathbf{V}[X_i] + \sum_{i,j\in[n]:i\neq j} \mathbf{E}[\overline{X}_i]\cdot\mathbf{E}[\overline{X}_j] && \text{[Fact 4]}\\
&= \textstyle\sum_{i\in[n]} \mathbf{V}[X_i]
\end{aligned}$$

where the last equality holds since $\mathbf{E}[\overline{X}_i] = 0$ and the one before it is due to the fact that X_i and X_j are independent.

The trick of considering $\overline{X}_i = X_i - \mathbf{E}[X_i]$ is good to bear in mind. We also observe that the latter proof only relied on the fact that each two variables are independent. For sake of future reference, let us state the consequence of this fact.

Claim 3 (linearity of the variance of pairwise independent random variables): *Let $X_1, ..., X_n$ be a sequence of random variables such that every two variables in the sequence are independent; that is, for every $i \neq j$ and every y, z it holds that $\mathbf{Pr}[(X_i, X_j) = (y, z)] = \mathbf{Pr}[X_i = y] \cdot \mathbf{Pr}[X_j = z]$. Then,*

$$\mathbf{V}\left[\sum_{i\in[n]} X_i\right] = \sum_{i\in[n]} \mathbf{V}[X_i].$$

Indeed, a sequence as in the hypothesis of Claim 3 is called pairwise independent.

4 Three Inequalities

The following probabilistic inequalities are very useful. These inequalities provide upper-bounds on the probability that a random variable deviates from its expectation.

4.1 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:

Theorem 4 (Markov's Inequality): *Let X be a non-negative random variable and v be a non-negative real number. Then*

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

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

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

4.2 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).

Theorem 5 (Chebyshev's Inequality): *Let X be a random variable, and $\delta > 0$. Then*

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

Proof: Defining a random variable $Y \stackrel{\text{def}}{=} (X - \mathbf{E}(X))^2$, and applying Markov's inequality to it, we get

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

and the claim follows. ■

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 $\mathbf{Pr}[X_i = \alpha \wedge X_j = \beta] = \mathbf{Pr}[X_i = \alpha] \cdot \mathbf{Pr}[X_j = \beta]$. Then, as a corollary to Chebyshev's inequality, we get

Corollary 6 (pairwise independent sampling): *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*

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

Using $\epsilon = \gamma \cdot \mu$ and $m = n \cdot \mu$, and assuming that $\sigma^2 \leq \mu$ (which always holds when $X_i \in [0,1]$), we obtain a ("multiplicative") bound of the form

$$\mathbf{Pr}\left[\left|\sum_{i\in[n]} X_i - m\right| \geq \gamma \cdot m\right] \leq \frac{1}{\gamma^2 m}. \tag{5}$$

Proof: Combining Chebyshev's inequality with Claim 3, we get

$$\begin{aligned}\mathbf{Pr}\left[\left|\sum_{i\in[n]} X_i - n\cdot\mu\right| \geq n\cdot\epsilon\right] &\leq \frac{\mathbf{V}\left[\sum_{i\in[n]} X_i\right]}{(n\epsilon)^2} \\ &= \frac{\sum_{i\in[n]} \mathbf{V}[X_i]}{(n\epsilon)^2} \\ &= \frac{n\sigma^2}{n^2\epsilon^2}\end{aligned}$$

and the claim follows. ■

Sampling by t-wise independent points: A sequence of random variables is called t-wise independent if every t variables in it are totally independent. While we shall not use the following result in this text, we find it useful in many other setting and believes that its derivation highlights the ideas that underly the proof of Corollary 6. For simplicity, we consider the case that the random variable range over $[0,1]$; a generalization to other bounded ranges can be derived similarly to the way this is done in the proof of Theorem 11 (in next section). Note that for $X \in [0,1]$, it holds that $\mathbf{E}[X^2] \leq \mathbf{E}[X]$, and thus $\mathbf{V}[X] \leq \mathbf{E}[X]$.

Theorem 7 ($2k$-wise independent sampling): *Let $X_1, X_2, ..., X_n \in [0,1]$ be $2k$-wise independent random variables and $\mu = \sum_{i\in[n]} \mathbf{E}[X_i]/n$. Suppose that $\mathbf{V}[X_i] \leq \beta$ for every $i \in [n]$. Then, for every $\epsilon > 0$, it holds that*

$$\mathbf{Pr}\left[\left|\frac{\sum_{i\in[n]} X_i}{n} - \mu\right| \geq \epsilon\right] < \left(\frac{3k\beta}{n\epsilon^2}\right)^k \tag{6}$$

Recall that for any random variable Z ranging in $[0,1]$, it holds that $\mathbf{V}[Z] \leq \mathbf{E}[Z]$. Hence, if the X_i's have identical expectation (which equals μ), then we may use $\beta = \mu$.

Proof: Define the random variables $\overline{X}_i \stackrel{\text{def}}{=} X_i - \mathbf{E}(X_i)$. Note that the $\overline{X}_i$'s are $2k$-wise independent, and each has zero expectation. Mimicking the proof of Chebyshev's inequality, we have

$$\mathbf{Pr}\left[\left|\sum_{i\in[n]}\frac{X_i}{n}-\mu\right|\geq\epsilon\right]\leq\frac{\mathbf{E}\left[\left(\sum_{i\in[n]}\overline{X}_i\right)^{2k}\right]}{\epsilon^{2k}\cdot n^{2k}} \tag{7}$$

The rest of the proof is devoted to upper-bounding the numerator in the r.h.s of Eq. (7). Generalizing the proof of Claim 3, we have

$$\begin{aligned}\mathbf{E}\left[\left(\sum_{i\in[n]}\overline{X}_i\right)^{2k}\right] &= \mathbf{E}\left[\sum_{i_1,\ldots,i_{2k}\in[n]}\prod_{j\in[2k]}\overline{X}_{i_j}\right]\\ &= \sum_{i_1,\ldots,i_{2k}\in[n]}\mathbf{E}\left[\prod_{j\in[2k]}\overline{X}_{i_j}\right]\end{aligned}$$

Now, the key observation is that each term in this sum that has some random variable appear in it with multiplicity 1 equals zero. More generally, for each sequence $\overline{i}=(i_1,...,i_{2k})$ and $j\in[n]$, denoting by $m_j(\overline{i})$ the multiplicity of j in $\overline{i}$, we have

$$\begin{aligned}\mathbf{E}\left[\prod_{j\in[2k]}\overline{X}_{i_j}\right] &= \mathbf{E}\left[\prod_{j\in[n]}\overline{X}_j^{m_j(\overline{i})}\right]\\ &= \prod_{j\in[n]}\mathbf{E}\left[\overline{X}_j^{m_j(\overline{i})}\right]\end{aligned}$$

where the last equality is due to the $2k$-wise independence of the random variables $\overline{X}_{i_1},...,\overline{X}_{i_{2k}}$. Denoting by S the set of $2k$-long sequences over $[n]$ in which no element appears with multiplicity 1 (and recalling that $\mathbf{E}[\overline{X}_j]=0$), we get

$$\mathbf{E}\left[\left(\sum_{i\in[n]}\overline{X}_i\right)^{2k}\right]=\sum_{(i_1,\ldots,i_{2k})\in S}\prod_{j\in[n]}\mathbf{E}\left[\overline{X}_j^{m_j(i_1,\ldots,i_{2k})}\right]. \tag{8}$$

Indeed, the maximum number of elements that may appear in any sequence $(i_1,...,i_{2k})\in S$ in at most k, since each element that appears in $(i_1,...,i_{2k})$ must appear in it with multiplicity at least 2. This already yields an upper bound of $|S|<\binom{n}{k}\cdot k^{2k}<(nk^2)^k$ on Eq. (8). A better upper bound can be obtained by partitioning S into $(S_1,...,S_k)$ such that $S_t\subset S$ contains all sequences such that each sequence contains exactly t elements.

$$\begin{aligned}\sum_{(i_1,\ldots,i_{2k})\in S}\prod_{j\in[n]}\mathbf{E}\left[\overline{X}_j^{m_j(i_1,\ldots,i_{2k})}\right] &= \sum_{t\in[k]}\sum_{(i_1,\ldots,i_{2k})\in S_t}\prod_{j\in[n]}\mathbf{E}\left[\overline{X}_j^{m_j(i_1,\ldots,i_{2k})}\right]\\ &\leq \sum_{t\in[k]}|S_t|\cdot\mathbf{E}\left[\overline{X}_1^2\right]^t \qquad (9)\\ &< \sum_{t\in[k]}(en/k)^t\cdot t^{2k}\cdot\beta^t\end{aligned}$$

where the first inequality uses the fact that for every $m > 2$ and $Z \in [-1,1]$ it holds that $\mathbf{E}[Z^m] \leq \mathbf{E}[Z^2]$, and the last inequality uses $\binom{n}{t} < (en/t)^t$ (for $t \leq n/2$). Combining Eq. (7)&(8)&(9), we get

$$\begin{aligned}
\mathbf{Pr}\left[\left|\sum_{i\in[n]} \frac{X_i}{n} - \mu\right| \geq \epsilon\right] &\leq \frac{\mathbf{E}\left[\left(\sum_{i\in[n]} \overline{X}_i\right)^{2k}\right]}{\epsilon^{2k} \cdot n^{2k}} \\
&< \frac{\sum_{t\in[k]} (en/k)^t \cdot t^{2k} \cdot \beta^t}{\epsilon^{2k} \cdot n^{2k}} \\
&< \frac{k^{2k} \cdot \sum_{t\in[k]} (\beta en/k)^t}{\epsilon^{2k} \cdot n^{2k}} \\
&< \frac{k^{2k} \cdot 2 \cdot (\beta en/k)^k}{\epsilon^{2k} \cdot n^{2k}} \\
&< \left(\frac{3k\beta}{n\epsilon^2}\right)^k
\end{aligned}$$

as required. ■

4.3 Chernoff Bound

When using pairwise independent sample points, the error probability of 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. Probability bounds supporting the foregoing statement are commonly referred to as Chernoff Bounds. We present such a bound next.

The bound that we present first is not the most popular bound, but it is a better starting point for deriving the popular bounds as well as other useful bounds, which we shall do later. In particular, the following bound considers independent random variables ranging arbitrarily in $[0,1]$ (rather than in $\{0,1\}$), where these random variables are not necessarily identical.

Theorem 8 (a Chernoff Bound): *Let $X_1, X_2, ..., X_n$ be independent random variables ranging in $[0,1]$, and $\beta > 0$. Let $\mu = \sum_{i\in[n]} \mathbf{E}[X_i]$ and suppose that $\sum_{i\in[n]} \mathbf{V}[X_i] \leq \beta$. Then, for every $\alpha \in (0, 2\beta]$, it holds that*

$$\mathbf{Pr}\left[\left|\sum_{i\in[n]} X_i - \mu\right| > \alpha\right] < 2 \cdot e^{-\alpha^2/4\beta} \tag{10}$$

Note that $\sum_{i\in[n]} \mathbf{V}[X_i] \leq \sum_{i\in[n]} \mathbf{E}[X_i^2] = \mu$, where the last equality uses the fact that $\mathbf{E}[X^2] \leq \mathbf{E}[X]$ holds for every random variable $X \in [0,1]$. Hence (assuming $\mu > 0$)[1], we can always use $\beta = \mu$, and obtain a meaningful bound whenever $\alpha > 2\sqrt{\mu}$.

Proof: We upper-bound $\mathbf{Pr}[\sum_{i\in[n]} X_i > \mu + \alpha]$, and $\mathbf{Pr}[\sum_{i\in[n]} X_i < \mu - \alpha]$ is bounded similarly (or, alternatively, by letting $Y_i = 1 - X_i$ and using the bound on $\mathbf{Pr}[\sum_{i\in[n]} Y_i > (n-\mu) + \alpha]$). Letting $\overline{X}_i \stackrel{\text{def}}{=} X_i - \mathbf{E}(X_i)$, we apply Markov's inequality to the random variable $e^{\lambda \sum_{i=1}^n \overline{X}_i}$, where

[1] Note that $\mu = 0$ implies that each X_i is identically zero.

$\lambda \in (0,1]$ will be determined to optimize the expression that we derive. Specifically, we get

$$\begin{aligned}
\mathbf{Pr}\left[\sum_{i\in[n]} \overline{X}_i > \alpha\right] &= \mathbf{Pr}\left[e^{\lambda \sum_{i\in[n]} \overline{X}_i} > e^{\lambda\alpha}\right] \\
&\leq \frac{\mathbf{E}\left[e^{\lambda \sum_{i\in[n]} \overline{X}_i}\right]}{e^{\lambda\alpha}} \\
&= e^{-\lambda\alpha} \cdot \mathbf{E}\left[\prod_{i\in[n]} e^{\lambda \overline{X}_i}\right] \\
&= e^{-\lambda\alpha} \cdot \prod_{i\in[n]} \mathbf{E}\left[e^{\lambda \overline{X}_i}\right]
\end{aligned}$$

where the last equality is due to the independence of the random variables. Now, using $e^x \leq 1 + x + x^2$ for every $x \in [-1,1]$, and observing that $\mathbf{E}[\overline{X}_i] = 0$, we get $\mathbf{E}[e^{\lambda \overline{X}_i}] \leq 1 + \lambda^2 \cdot \mathbf{E}[\overline{X}_i^2]$, which equals $1 + \lambda^2 \cdot \mathbf{V}[X_i]$. Hence,

$$\begin{aligned}
\mathbf{Pr}\left[\sum_{i\in[n]} \overline{X}_i > \alpha\right] &\leq e^{-\lambda\alpha} \cdot \prod_{i\in[n]} \mathbf{E}\left[e^{\lambda \overline{X}_i}\right] \\
&\leq e^{-\lambda\alpha} \cdot \prod_{i\in[n]} \left(1 + \lambda^2 \cdot \mathbf{V}[X_i]\right) \\
&\leq e^{-\lambda\alpha} \cdot \prod_{i\in[n]} e^{\lambda^2 \cdot \mathbf{V}[X_i]} \\
&= e^{-\lambda\alpha} \cdot e^{\lambda^2 \cdot \sum_{i\in[n]} \mathbf{V}[X_i]}
\end{aligned}$$

where the last inequality is due to using $1+y \leq e^y$ for every $y \in [0,1]$. Recalling that $\sum_{i\in[n]} \mathbf{V}[X_i] \leq \beta$ and optimizing at $\lambda = \alpha/2\beta \in (0,1]$, we obtain

$$\begin{aligned}
\mathbf{Pr}\left[\sum_{i\in[n]} \overline{X}_i > \alpha\right] &\leq e^{-\lambda\alpha + \lambda^2\beta} \\
&= e^{-\alpha^2/4\beta}
\end{aligned}$$

and the claim follows. ■

The popular Chernoff Bounds. The popular bounds refer to the case that all X_i's are indentical (and range in $[0,1]$). The more popular version refers to an additive deviation of $\epsilon > 0$.

Corollary 9 (a standard ("additive") Chernoff Bound): *Let $X_1, X_2, ..., X_n$ be identical independent random variables ranging in $[0,1]$, and let $p = \mathbf{E}[X_1]$. Then, for every $\epsilon \in (0, 2(1-p)p]$, it holds that*

$$\mathbf{Pr}\left[\left|\frac{1}{n} \cdot \sum_{i\in[n]} X_i - p\right| > \epsilon\right] \;<\; 2 \cdot e^{-\epsilon^2 n/(4p(1-p))} \;<\; 2 \cdot e^{-\epsilon^2 n}. \tag{11}$$

For every $\epsilon \in (0,1]$, it holds that

$$\mathbf{Pr}\left[\left|\frac{1}{n}\cdot\sum_{i\in[n]}X_i - p\right| > \epsilon\right] \;<\; 2\cdot e^{-\epsilon^2 n/4} \tag{12}$$

Proof: We invoke Theorem 8 with $\mu = n\cdot p$ and $\alpha = n\cdot\epsilon$. For Eq. (11) we use $\beta = n\cdot(1-p)p$, while noting that $\mathbf{V}[X_i] \leq \mathbf{E}[X_i] - \mathbf{E}[X_i]^2 = (1-p)p$ (since $X_i \in [0,1]$ implies $\mathbf{E}[X_i^2] \leq \mathbf{E}[X_i]$). For Eq. (11) we use $\beta = n\cdot p$, while assuming without loss of generality that $p \geq 1/2$ (and considering the $1 - X_i$'s otherwise). ■

Corollary 10 (a standard multiplicative Chernoff Bound): *Let $X_1, X_2, ..., X_n$ be identical independent random variables ranging in $[0,1]$, and let $p = \mathbf{E}[X_1]$. Then, for every $\gamma \in (0,2]$, it holds that*

$$\mathbf{Pr}\left[\left|\frac{1}{n}\cdot\sum_{i\in[n]}X_i - p\right| > \gamma\cdot p\right] \;<\; 2\cdot e^{-\gamma^2 pn/4}. \tag{13}$$

Proof: We invoke Theorem 8 with $\mu = n\cdot p$ and $\alpha = \gamma\cdot\mu$, and use $\beta = \mu$ (while relying on $\mathbf{V}[X_i] \leq \mathbf{E}[X_i]$). ■

Generalization to an arbitrary bounded range. The case that the X_i's range in an arbitrary interval can be handled my using a linear transformation that maps this interval to $[0,1]$.

Theorem 11 (Theorem 8, generalized): *Let $X_1, X_2, ..., X_n$ be independent random variables ranging in $[a,b]$, and $\beta > 0$. Let $\mu = \sum_{i\in[n]}\mathbf{E}[X_i]$ and suppose that $\sum_{i\in[n]}\mathbf{V}[X_i] \leq \beta$. Then, for every $\alpha \in (0, 2\beta/(b-a)]$, it holds that*

$$\mathbf{Pr}\left[\left|\sum_{i\in[n]}X_i - \mu\right| > \alpha\right] \;<\; 2\cdot e^{-\alpha^2/4\beta} \tag{14}$$

Note that in this case (i.e., of independent X_i's ranging in $[a,b]$) it holds that $\sum_{i\in[n]}\mathbf{V}[X_i] \leq (b-a)\cdot(\mu - n\cdot a)$, where the inequality uses $\mathbf{V}[X_i] = (b-a)^2\cdot\mathbf{V}[(X_i-a)/(b-a)]$ and the fact that $(X_i-a)/(b-a) \in [0,1]$.[2] Hence, we may use $\beta = (b-a)\cdot(\mu - n\cdot a)$.

Before proving Theorem 11, we note that a multiplicative version of Theorem 11 can be obtained by letting $\gamma = \alpha/(\mu - n\cdot a)$ and using $\beta = (b-a)\cdot(\mu - n\cdot a)$. Hence, for every $\gamma \in (0,2]$, it holds that

$$\mathbf{Pr}\left[\left|\sum_{i\in[n]}X_i - \mu\right| > \gamma\cdot(\mu - n\cdot a)\right] \;<\; 2\cdot e^{-\gamma^2(\mu - n\cdot a)/4(b-a)} \tag{15}$$

For $a = 0$, the bound simplifies to $2\cdot e^{-\gamma^2\mu/4b}$.

Proof: We consider the random variables $X'_1, ..., X'_n$ such that $X'_i = (X_i - a)/(b-a) \in [0,1]$. Let $\alpha' = \alpha/(b-a)$ and $\beta' = \beta/(b-a)^2$, and note that $\sum_{i\in[n]}\mathbf{V}[X'_i] = \sum_{i\in[n]}\mathbf{V}[X_i]/(b-a)^2 \leq \beta'$ and that $\alpha' \in (0, 2\beta')$. Invoking Theorem 8 (with parameters α' and β'), we get

$$\mathbf{Pr}\left[\left|\sum_{i\in[n]}X'_i - \frac{\mu}{b-a}\right| > \frac{\alpha}{b-a}\right] \;<\; 2\cdot e^{-(\alpha/(b-a))^2/4(\beta/(b-a)^2)}$$

[2] Hence $\mathbf{V}[X_i] \leq (b-a)^2\cdot\mathbf{E}[(X_i-a)/(b-a)]$, whereas $\mathbf{E}[(X_i-a)/(b-a)] = (\mathbf{E}[X_i]-a)/(b-a)$.

and the claim follows. ■

4.4 Pairwise independent versus totally independent sampling

To demonstrate the difference between the sampling bounds provided in 4.2 and 4.3, 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 $\mathbf{Pr}[|Z - v| > \epsilon] \leq \delta$. By Chernoff Bound (e.g., Corollary 9), 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 Corollary 6 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).