Appendix A

Appendix

A.1 Probability theory

A random variable is a mapping from a probability space to $\mathbb{R}$. To give an example, the probability space could be that of all $2^n$ possible outcomes of $n$ tosses of a fair coin, and $X_i$ is the random variable that is 1 if the $i$th toss is a head, and is 0 otherwise. An event is a subset of the probability space. The following simple bound —called the union bound—is often used in the book. For every set of events $B_1, B_2, \ldots, B_k$,

$$\Pr[\bigcup_{i=1}^k B_i] \leq \sum_{i=1}^k \Pr[B_i].$$  \hspace{1cm} (A.1)

A.1.1 The averaging argument

We list various versions of the “averaging argument.” Sometimes we give two versions of the same result, one as a fact about numbers and one as a fact about probability spaces.

**Lemma A.1**

If $a_1, a_2, \ldots, a_n$ are some numbers whose average is $c$ then some $a_i \geq c$.

**Lemma A.2** (“The Probabilistic Method”)

If $X$ is a random variable which takes values from a finite set and $E[X] = \mu$ then the event “$X \geq \mu$” has nonzero probability.

**Corollary A.3**

If $Y$ is a real-valued function of two random variables $x, y$ then there is a choice $c$ for $y$ such that $E[Y(x, c)] \geq E[Y(x, y)]$. 

403
Lemma A.4
If $a_1, a_2, \ldots, a_n \geq 0$ are numbers whose average is $c$ then the fraction of $a_i$’s that are greater than (resp., at least) $kc$ is less than (resp, at most) $1/k$.

Lemma A.5 (“Markov’s inequality”)
Any non-negative random variable $X$ satisfies

$$\Pr(X \geq kE[X]) \leq \frac{1}{k}.$$ 

Corollary A.6
If $a_1, a_2, \ldots, a_n \in [0, 1]$ are numbers whose average is $1 - \gamma$ then at least $1 - \sqrt{\gamma}$ fraction of them are at least $1 - \sqrt{\gamma}$.

Can we give any meaningful upperbound on $\Pr[X < c \cdot E[X]]$ where $c < 1$? Yes, if $X$ is bounded.

Lemma A.7
If $a_1, a_2, \ldots, a_n$ are numbers in the interval $[0, 1]$ whose average is $\rho$ then at least $\rho/2$ of the $a_i$’s are at least as large as $\rho/2$.

Proof: Let $\gamma$ be the fraction of $i$’s such that $a_i \geq \rho/2$. Then $\gamma + (1 - \gamma)\rho/2$ must be at least $\rho/2$, so $\gamma \geq \rho/2$. □ More generally, we have

Lemma A.8
If $X \in [0, 1]$ and $E[X] = \mu$ then for any $c < 1$ we have

$$\Pr[X \leq c\mu] \leq \frac{1 - \mu}{1 - c\mu}.$$ 

Example A.1 Suppose you took a lot of exams, each scored from 1 to 100. If your average score was 90 then in at least half the exams you scored at least 80.

----

A.1.2 Deviation upperbounds

Under various conditions, one can give upperbounds on the probability of an event. These upperbounds are usually derived by clever use of Markov’s inequality.

The first bound is Chebyshev’s inequality, useful when only the variance is known.
Lemma A.9 (Chebyshev inequality)
If $X$ is a random variable and $E[X] = \mu$ and $E[(X - \mu)^2] = \sigma^2$ then $\Pr[|X - \mu| > k\sigma] \leq 1/k^2$.

Proof: Apply Markov’s inequality to the random variable $(X - \mu)^2$, noting that by definition of variance, $E[(X - \mu)^2] = \sigma^2$. □

The next inequality has many names, and is widely known in theoretical computer science as Chernoff bounds. It considers scenarios of the following type. Suppose we toss a fair coin $n$ times. The expected number of heads is $n/2$. How tightly is this number concentrated? Should we be very surprised if after 1000 tosses we have 625 heads? The bound we present is more general, and concerns $n$ different coin tosses of possibly different biases (the bias is the probability of obtaining “heads”; for a fair coin this is $1/2$).

A sequence of \{0, 1\} random variables is Poisson, as against Bernoulli, if the expected values can vary between trials.

Theorem A.10 (“Chernoff” bounds)
Let $X_1, X_2, \ldots, X_n$ be independent Poisson trials and let $p_i = E[X_i]$, where $0 < p_i < 1$. Then the sum $X = \sum_{i=1}^n X_i$, which has mean $\mu = \sum_{i=1}^n p_i$, satisfies

$$\Pr[X \geq (1 + \delta)\mu] \leq \left[ \frac{e^\delta}{(1 + \delta)^{1+\delta}} \right]^\mu. \quad (A.2)$$

$$\Pr[X \leq (1 - \delta)\mu] \leq \left[ \frac{e^{-\delta}}{(1 - \delta)^{1-\delta}} \right]^\mu. \quad (A.3)$$

Proof: Surprisingly, these are also proved using the Markov inequality. We only prove the first inequality; a similar proof exists for the second. We introduce a positive dummy variable $t$ and observe that

$$E[\exp(tX)] = E[\exp(t \sum_i X_i)] = E[\prod_i \exp(tX_i)] = \prod_i E[\exp(tX_i)], \quad (A.4)$$

where the last equality holds because the $X_i$ r.v.s are independent. Now,

$$E[\exp(tX_i)] = (1 - p_i) + p_i e^t,$$

therefore,

$$\prod_i E[\exp(tX_i)] = \prod_i [1 + p_i(e^t - 1)] \leq \prod_i \exp(p_i(e^t - 1))$$

$$= \exp(\sum_i p_i(e^t - 1)) = \exp(\mu(e^t - 1)), \quad (A.5)$$
as } 1 + x \leq e^x \). Finally, apply Markov’s inequality to the random variable \( \exp(tX) \), viz.

\[
\Pr[X \geq (1+\delta)\mu] = \Pr[\exp(tX) \geq \exp(t(1+\delta)\mu)] \leq \frac{\mathbb{E}[\exp(tX)]}{\exp(t(1+\delta)\mu)} = \frac{\exp((e^t - 1)\mu)}{\exp(t(1+\delta)\mu)},
\]

using lines (A.4) and (A.5) and the fact that \( t \) is positive. Since \( t \) is a dummy variable, we can choose any positive value we like for it. Simple calculus shows that the right hand side is minimized for \( t = \ln(1 + \delta) \) and this leads to the theorem statement. \( \Box \)

By the way, if all \( n \) coin tosses are fair (Heads has probability 1/2) then the probability of seeing \( N \) heads where \(|N - n/2| > a\sqrt{n}\) is at most \( e^{-a^2/2} \). The chance of seeing at least 625 heads in 1000 tosses of an unbiased coin is less than \( 5 \times 10^{-7} \).

### A.1.3 Sampling methods

Typically, to take \( n \) independent samples the amount of random bits required increases linearly with \( n \). At a few places in the book we encounter pairwise independent sampling. This is very frugal with its use of random bits, since it allows \( n \) samples to be taken using \( 2 \log n \) random bits. The downside is that one can only use Chebyshev’s inequality for deviation bounds, not Chernoff bounds.

**Definition A.1** Random variables \( x_1, \ldots, x_m \) are pairwise independent if

\[
\forall i, j \forall a, b \Pr[x_i = a, x_j = b] = \Pr[x_i = a] \Pr[x_j = b].
\] (A.6)

Pairwise independent random variables are useful because of the Chebyshev inequality. Suppose that \( x_1, \ldots, x_m \) are pairwise independent with \( \mathbb{E}(x_i) = \mu, \text{Var}(x_i) = \sigma^2 \). Then we have

\[
\text{Var}(\sum x_i) = \mathbb{E}[(\sum x_i)^2] - \mathbb{E}[\sum x_i]^2 = \sum_i (\mathbb{E}[x_i^2] - \mathbb{E}[x_i]^2),
\]

where we have used pairwise independence in the last step. Thus \( \text{Var}(\sum x_i) = m\sigma^2 \). By the Chebyshev inequality, \( \Pr[|\sum x_i - m\mu| > k\sqrt{m\sigma}] \leq 1/k^2 \). So, the sum of the variables is somewhat concentrated about the mean. This is in contrast with the case of complete independence, when Chernoff bounds would give an exponentially stronger concentration result (the \( 1/k^2 \) would be replaced by \( \exp(\Theta(-k^2)) \)).
Example A.2 In \( \mathbb{Z}_p \), choose \( a \) and \( b \) randomly and independently. Then the random variables \( a + b, a + 2b, \ldots, a + (p-1)b \) are pairwise independent. Indeed, for any \( t, s, j \neq k \in \mathbb{Z}_p \), \( \Pr[a + jb = t] = 1/p \), and \( \Pr[a + jb = t, a + kb = s] = 1/p^2 \), because this linear system is satisfied by exactly one \((a, b)\)-pair out of \( p^2 \).

Example A.3 Let \( m = 2^k - 1 \). The set \([1 \ldots m]\) is in \( 1-1 \) correspondence with the set \( 2^{[1\ldots k]} \setminus \emptyset \). We will construct \( m \) random variables corresponding to all nonempty subsets of \([1 \ldots k]\). Pick uniformly at random \( k \) binary strings \( t_1, \ldots, t_k \) of length \( n \) and set \( Y_S = \sum_{i \in S} t_i \) (mod 2), where \( S \subset [1 \ldots k], S \neq \emptyset \). For any \( S_1 \neq S_2 \), the random variables \( Y_{S_1} \) and \( Y_{S_2} \) are independent, because one can always find an \( i \) such that \( i \in S_2 \setminus S_1 \), so the difference between \( Y_{S_1} \) and \( Y_{S_2} \) is always a sum of several uniformly distributed random vectors, which is a uniformly distributed random vector itself. That is, even if we fix the value of \( Y_{S_1} \), we still have to toss a coin for each position in \( Y_{S_2} \), and

\[
\Pr[Y_{S_1} = \vec{s}, Y_{S_2} = \vec{t}] = \Pr[Y_{S_1} \oplus Y_{S_2} = \vec{t} \oplus \vec{s} | Y_{S_1} = \vec{s}] = \frac{1}{2^{2n}}.
\]

A.1.4 The random subsum principle

The following fact is used often in the book. Let \( a_1, a_2, \ldots, a_k \) be elements of a finite field \( GF(q) \). We do not know the elements, but we are told that at least one of them is nonzero. Now we pick a subset \( S \) of \( \{1, 2, \ldots, k\} \) uniformly at random. Then the probability that \( \sum_{i \in S} a_i \neq 0 \) is at least \( 1/2 \).

Here is the reason. Suppose that \( a_j \) is nonzero. Then we can think of \( S \) as having been picked as follows. First we pick a random subset \( S' \) of \( \{1, 2, \ldots, k\} \setminus \{j\} \) and then we toss a fair coin to decide whether or not to add \( j \) to it. If \( \sum_{i \in S'} a_i \neq 0 \) then with probability \( 1/2 \) \( S = S' \), in which case \( \sum_{i \in S} a_i \neq 0 \). If \( \sum_{i \in S'} a_i = 0 \) then with probability \( 1/2 \) \( S = S' \cup \{j\} \), in which case \( \sum_{i \in S} a_i \neq 0 \).

A.2 Polynomials

We list some basic facts about univariate polynomials.
Theorem A.11
A nonzero polynomial of degree $d$ has at most $d$ distinct roots.

Proof: Suppose $p(x) = \sum_{i=0}^{d} c_i x^i$ has $d + 1$ distinct roots $\alpha_1, \ldots, \alpha_{d+1}$ in some field $F$. Then

$$\sum_{i=0}^{d} \alpha_j^i \cdot c_i = p(\alpha_j) = 0,$$

for $j = 1, \ldots, d + 1$. This means that the system $A \mathbf{y} = \mathbf{0}$ with

$$A = \begin{pmatrix}
1 & \alpha_1 & \alpha_1^2 & \ldots & \alpha_1^d \\
1 & \alpha_2 & \alpha_2^2 & \ldots & \alpha_2^d \\
\vdots & \vdots & \vdots & & \vdots \\
1 & \alpha_{d+1} & \alpha_{d+1}^2 & \ldots & \alpha_{d+1}^d
\end{pmatrix}$$

has a solution $\mathbf{y} = \mathbf{c}$. The matrix $A$ is a Vandermonde matrix, hence

$$\det A = \prod_{i>j} (\alpha_i - \alpha_j),$$

which is nonzero for distinct $\alpha_i$. Hence $\text{rank } A = d + 1$. The system $A \mathbf{y} = \mathbf{0}$ has therefore only a trivial solution — a contradiction to $\mathbf{c} \neq \mathbf{0}$. □

Theorem A.12
For any set of pairs $(a_1, b_1), \ldots, (a_{d+1}, b_{d+1})$ there exists a unique polynomial $g(x)$ of degree at most $d$ such that $g(a_i) = b_i$ for all $i = 1, 2, \ldots, d + 1$.

Proof: The requirements are satisfied by Lagrange Interpolating Polynomial:

$$\sum_{i=1}^{d+1} b_i \prod_{j \neq i} \frac{x - a_j}{a_i - a_j},$$

If two polynomials $g_1(x), g_2(x)$ satisfy the requirements then their difference $p(x) = g_1(x) - g_2(x)$ is of degree at most $d$, and is zero for $x = a_1, \ldots, a_{d+1}$. Thus, from the previous theorem, polynomial $p(x)$ must be zero and polynomials $g_1(x), g_2(x)$ identical. □

The following elementary result is usually attributed to Schwartz and Zippel in the computer science community, though it was certainly known earlier (see e.g. DeMillo and Lipton).
A.2. POLYNOMIALS

**Theorem A.13 (Schwartz-Zippel)**
If a polynomial \( p(x_1, x_2, \ldots, x_m) \) over \( F = GF(q) \) is nonzero and has total degree at most \( d \), then
\[
\Pr[p(a_1 \ldots a_m) \neq 0] \geq 1 - \frac{d}{q},
\]
where the probability is over all choices of \( a_1 \ldots a_m \in F \).

**Proof:** We use induction on \( m \). If \( m = 1 \) the statement follows from Theorem A.11. Suppose the statement is true when the number of variables is at most \( m - 1 \). Then \( p \) can be written as
\[
p(x_1, x_2, \ldots, x_m) = \sum_{i=0}^{d} x_1^i p_i(x_2, \ldots, x_m),
\]
where \( p_i \) has total degree at most \( d - i \). Since \( p \) is nonzero, at least one of \( p_i \) is nonzero. Let \( k \) be the largest \( i \) such that \( p_i \) is nonzero. Then by the inductive hypothesis,
\[
\Pr[p_1(a_2, a_3, \ldots, a_m) \neq 0] \geq 1 - \frac{d - k}{q}.
\]
Whenever \( p_i(a_2, a_3, \ldots, a_m) \neq 0 \), \( p(x_1, a_2, a_3, \ldots, a_m) \) is a nonzero univariate polynomial of degree \( k \), and hence becomes 0 only for at most \( k \) values of \( x_1 \). Hence
\[
\Pr[p(a_1 \ldots a_m) \neq 0] \geq (1 - \frac{k}{q})(1 - \frac{d - k}{q}) \geq 1 - \frac{d}{q},
\]
and the induction is completed. \( \Box \)

### A.2.1 Representing data by multivariate polynomials

**Lemma A.14 (Polynomial extension lemma)**
For some integer \( h \) let \( H = \{u_0, u_1, \ldots, u_h\} \) be a subset of \( h + 1 \) elements from \( GF(q) \). For each function \( f : H^m \to F \) there exists a polynomial \( \hat{f}(x_1, x_2, \ldots, x_m) \) whose degree in each \( x_i \) is \( h \) and which satisfies \( \hat{f}(v) = f(v) \) for all \( v \in H^m \). This polynomial can be computed in time \( \text{poly}(h^m) \).

**Proof:** The requirements are satisfied by the Lagrange interpolating polynomial:
\[
\hat{f}(x_1, x_2, \ldots, x_m) = \sum_{u \in H^m} f(u) \prod_{i=1}^{m} \frac{\Pi_{j \neq i} x_i - u_j}{\Pi_{j \neq i} u_i - u_j} \quad (A.7)
\]
\( \Box \)
A.3 Eigenvalues and expanders

An eigenvalue of an \( n \times n \) symmetric real matrix \( M \) is a real number \( \lambda \) such that there exists a vector \( x \in \mathbb{R}^n \) such that \( M \cdot x = \lambda x \). It is known that for every such matrix there are \( n \) eigenvalues \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \) and \( n \) corresponding eigenvectors \( u_1, u_2, \ldots, u_n \in \mathbb{R}^n \) that form an orthonormal basis of \( \mathbb{R}^n \). The Courant Fisher theorem characterizes \( \lambda_1 \) as

\[
\lambda_1 = \min_{x \in \mathbb{R}^n} \frac{x^T M x}{|x|^2}
\]  

(A.8)

and more generally \( \lambda_i \) as

\[
\lambda_i = \min_{x \in L_{i-1}} \frac{x^T M x}{|x|^2}
\]  

(A.9)

where \( L_{i-1} \) is the space of vectors orthogonal to \( u_1, u_2, \ldots, u_{i-1} \).

The adjacency matrix of an \( n \)-node undirected graph \( G = (V, E) \) is an \( n \times n \) matrix \( A_G \) whose \((i, j)\) entry is 1 if \( \{i, j\} \) is an edge and 0 otherwise.

In a few places in the book we are interested in constant degree expanders, and these have more than one definition. (The definitions are related.) We define them using eigenvalues. We recall that for a \( d \)-regular graph, the largest eigenvalue of the adjacency matrix is \( d \).

**Definition A.2 (Expander family)** If \( \lambda < d \) then an infinite sequence of graphs is a \((d, \lambda)\)-expander family if each graph in the sequence is \( d \)-regular and its second largest eigenvalue (in absolute value) is at most \( \lambda \). The quantity \( d - \lambda \) is called the eigenvalue gap.

An explicit construction of a \((5, \lambda)\)-expander family is known for some \( \lambda < 5 \) (Margulis, simplified in Gabber-Galil). For every positive integer \( n \), the family contains a graph vertex set with size close to \( n \). In this context, “explicit construction” means that there exists a polynomial-time machine that, given any vertex number \( i \), gives the 5 neighbors of \( i \). (Note that the running time is polynomial in the representation of \( i \), in other words, poly(log \( i \)).) An explicit construction of expanders (for some constant \( d \)) also follows from the algorithm in Section 18.6.

Let \( G \) be any \((d, \lambda)\)-expander on \( n \) nodes, with adjacency matrix \( A \). Let \( M \) be the Markov chain defined by \( G \) (that is, \( M \) is a \( n \times n \) matrix such that in entry \((i, j)\) there is the transition probability from node \( i \) to node \( j \)), hence \( M = \frac{1}{d} A \). Thus the largest eigenvalue is 1 and every other eigenvalue \( \lambda_i \) satisfies \( |\lambda_i| \leq \lambda/5 \).
Then $M$ is ergodic, and the stationary distribution is $\frac{1}{n}\mathbf{1}$.

Let $X$ be some distribution on the vertices of $G$. Then $X$ can be written as:

$$X = \frac{1}{n}\mathbf{1} + \sum_{i=2}^{n} \alpha_i e_i$$

Where the $e_i$'s are the eigenvectors of $M$; they form an orthogonal basis since $M$ is symmetric. The coefficient of the first eigenvector is $\frac{1}{n}$ because

$$\langle X, \mathbf{1} \rangle = \frac{1}{\|\mathbf{1}\|_2} \sum_a \Pr[X = a] = \frac{1}{n}.$$

Therefore, after a random walk of length $t$, the resulting distribution on the vertices is:

$$M^t X = M^{t-1}(MX) = M^{t-1}\left(\frac{1}{n}\mathbf{1} + \sum_{i=2}^{n} \alpha_i \lambda_i e_i\right)$$

$$= \ldots = \frac{1}{n}\mathbf{1} + \sum_{i=2}^{n} \alpha_i \lambda_i^t e_i$$

Hence

$$\|M^t X - \frac{1}{n}\mathbf{1}\|_1 \leq (\frac{\lambda}{5})^t \|X - \frac{1}{n}\mathbf{1}\|_1$$

$$\leq (\frac{\lambda}{5})^t \|X - \frac{1}{n}\mathbf{1}\|_2 \sqrt{n}. \quad (A.10)$$

In particular, since $\lambda/5 < 1$ is some constant independent of $n$, we see that for some large enough constant $c$, taking $t = c \log n$ steps is enough to make the distance of $M^t X$ from the uniform distribution less than $1/n$, regardless of $X$. To give an example, if the random walk starts from a fixed vertex ($X$ is a vector which has one coordinate 1 and the rest 0) then a random walk of length $O(\log n)$ is enough to erase almost all “memory” of the starting vertex.
Bibliography


BIBLIOGRAPHY


BIBLIOGRAPHY


