Contents

12 Appendix

12.1 Asymptotic Notation

We introduce the big O notation here. The motivating example is the analysis of the running time of an algorithm. The running time may be a complicated function of the input length n such as $5n^3+25n^2\ln n-6n+22$. Asymptotic analysis is concerned with the behavior as $n \to \infty$ where the higher order term $5n^3$ dominates. Further, the coefficient 5 of $5n^3$ is not of interest since its value varies depending on the machine model. So we say that the function is $O(n^3)$. The big O notation applies to functions on the positive integers taking on positive real values.

Definition 12.1 *For functions f and g from the natural numbers to the positive reals, $f(n)$ is $O(g(n))$ if there exists a constant $c>0$ such that for all n, $f(n) \leq cg(n)$.* ∎

Thus, $f(n) = 5n^3 + 25n^2 \ln n - 6n + 22$ is $O(n^3)$. The upper bound need not be tight. Not only is $f(n)$, $O(n^3)$, it is also $O(n^4)$. Note $g(n)$ must be strictly greater than 0 for all n.

To say that the function $f(n)$ grows at least as fast as $g(n)$, one uses a notation called omega of n. For positive real valued f and g, $f(n)$ is $\Omega(g(n))$ if there exists a constant $c > 0$ such that for all n, $f(n) \geq cg(n)$. If $f(n)$ is both $O(g(n))$ and $\Omega(g(n))$, then $f(n)$ is $\Theta(g(n))$. Theta of n is used when the two functions have the same asymptotic growth rate.

Many times one wishes to bound the low order terms. To do this, a notation called little o of n is used. We say $f(n)$ is $o(g(n))$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$. Note that $f(n)$ being $O(g(n))$ means that asymptotically $f(n)$ does not grow faster than $g(n)$, whereas $f(n)$ being $o(g(n))$ means that asymptotically $f(n)/g(n)$ goes to zero. If $f(n) = 2n+\sqrt{n}$, then

asymptotic upper bound $f(n)$ is $O(g(n))$ if for all n, $f(n) \leq cg(n)$ for some constant $c > 0$.	$\leq$
asymptotic lower bound $f(n)$ is $\Omega(g(n))$ if for all n, $f(n) \geq cg(n)$ for some constant $c > 0$.	$\geq$
asymptotic equality $f(n)$ is $\Theta(g(n))$ if it is both $O(g(n))$ and $\Omega(g(n))$.	$=$
$f(n)$ is $o(g(n))$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$.	$<$
$f(n) \sim g(n)$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$.	$=$
$f(n)$ is $\omega\left(g\left(n\right)\right)$ **if** $\lim_{n\to\infty} \frac{f(n)}{g(n)} = \infty$.	$>$

$f(n)$ is $O(n)$ but in bounding the lower order term, we write $f(n) = 2n + o(n)$. Finally, we write $f(n) \sim g(n)$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$ and say $f(n)$ is $\omega(g(n))$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = \infty$. The difference between $f(n)$ being $\Theta(g(n))$ and $f(n) \sim g(n)$ is that in the first case $f(n)$ and $g(n)$ may differ by a multiplicative constant factor.

12.2 Useful relations

Summations

$$\sum_{i=0}^{n} a^i = 1 + a + a^2 + \cdots = \frac{1 - a^{n+1}}{1 - a}, \quad a \neq 1$$

$$\sum_{i=0}^{\infty} a^i = 1 + a + a^2 + \cdots = \frac{1}{1 - a}, \quad |a| < 1$$

$$\sum_{i=0}^{\infty} ia^i = a + 2a^2 + 3a^3 \cdots = \frac{a}{(1 - a)^2}, \quad |a| < 1$$

$$\sum_{i=0}^{\infty} i^2a^i = a + 4a^2 + 9a^3 \cdots = \frac{a(1 + a)}{(1 - a)^3}, \quad |a| < 1$$

$$\sum_{i=1}^{n} i = \frac{n(n + 1)}{2}$$

$$\sum_{i=1}^{n} i^2 = \frac{n(n + 1)(2n + 1)}{6}$$

$$\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6}$$

We prove one equality.

$$\sum_{i=0}^{\infty} ia^i = a + 2a^2 + 3a^3 \cdots = \frac{a}{(1 - a)^2}, \text{ provided } |a| < 1.$$

Write $S = \sum_{i=0}^{\infty} ia^i$.

$$aS = \sum_{i=0}^{\infty} ia^{i+1} = \sum_{i=1}^{\infty} (i - 1)a^i.$$

Thus,

$$S - aS = \sum_{i=1}^{\infty} ia^i - \sum_{i=1}^{\infty} (i - 1)a^i = \sum_{i=1}^{\infty} a^i = \frac{a}{1 - a},$$

from which the equality follows. The sum $\sum_i i^2a^i$ can also be done by an extension of this method (left to the reader). Using generating functions, we will see another proof of both

these equalities by derivatives.

$\sum_{i=1}^{\infty} \frac{1}{i} = 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) + \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}\right) + \cdots \geq 1 + \frac{1}{2} + \frac{1}{2} + \cdots$ and thus diverges.

The summation $\sum_{i=1}^{n} \frac{1}{i}$ grows as $\ln n$ since $\sum_{i=1}^{n} \frac{1}{i} \approx \int_{x=1}^{n} \frac{1}{x}\, dx$. In fact, $\lim_{i\to\infty} \left(\sum_{i=1}^{n} \frac{1}{i} - \ln(n) \right) = \gamma$ where $\gamma \cong 0.5772$ is Euler's constant. Thus, $\sum_{i=1}^{n} \frac{1}{i} \cong \ln(n) + \gamma$ for large n.

Truncated Taylor series

If all the derivatives of a function $f(x)$ exist, then we can write

$$f(x) = f(0) + f'(0)x + f''(0)\frac{x^2}{2} + \cdots .$$

The series can be truncated. In fact, there exists some y between 0 and x such that

$$f(x) = f(0) + f'(y)x.$$

Also, there exists some z between 0 and x such that

$$f(x) = f(0) + f'(0)x + f''(z)\frac{x^2}{2}$$

and so on for higher derivatives. This can be used to derive inequalities. For example, if $f(x) = \ln(1 + x)$, then its derivatives are

$$f'(x) = \frac{1}{1+x}\ ;\ f''(x) = -\frac{1}{(1+x)^2}\ ;\ f'''(x) = \frac{2}{(1+x)^3}.$$

For any z, $f''(z) < 0$ and thus for any x, $f(x) \leq f(0) + f'(0)x$ hence, $\ln(1+x) \leq x$, which also follows from the inequality $1 + x \leq e^x$. Also using

$$f(x) = f(0) + f'(0)x + f''(0)\frac{x^2}{2} + f'''(z)\frac{x^3}{3!}$$

for $z > -1$, $f'''(z) > 0$, and so for $x > -1$,

$$\ln(1 + x) > x - \frac{x^2}{2}.$$

Exponentials and logs

$$a^{\log b} = b^{\log a}$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \qquad e = 2.7182 \qquad \frac{1}{e} = 0.3679$$

Setting $x = 1$ in the equation $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$ yields $e = \sum_{i=0}^{\infty} \frac{1}{i!}$.

$$\lim_{n\to\infty}\left(1+\tfrac{a}{n}\right)^n = e^a$$

$$\ln(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \frac{1}{4}x^4 \cdots \quad |x| < 1$$

The above expression with $-x$ substituted for x gives rise to the approximations

$$\ln(1-x) < -x$$

which also follows from $1 - x \leq e^{-x}$, since $\ln(1-x)$ is a monotone function for $x \in (0,1)$.

For $0 < x < 0.69$, $\ln(1-x) > -x - x^2$.

Trigonometric identities

$$\begin{aligned}
&e^{ix} = \cos(x) + i\sin(x)\\
&\cos(x) = \tfrac{1}{2}\left(e^{ix} + e^{-ix}\right)\\
&\sin(x) = \tfrac{1}{2i}\left(e^{ix} - e^{-ix}\right)\\
&\sin(x \pm y) = \sin(x)\cos(y) \pm \cos(x)\sin(y)\\
&\cos(x \pm y) = cos(x)\cos(y) \mp \sin(x)\sin(y)\\
&\cos(2\theta) = \cos^2\theta - \sin^2\theta = 1 - 2\sin^2\theta\\
&\sin(2\theta) = 2\sin\theta\cos\theta\\
&\sin^2\tfrac{\theta}{2} = \tfrac{1}{2}(1 - \cos\theta)\\
&\cos^2\tfrac{\theta}{2} = \tfrac{1}{2}(1 + \cos\theta)
\end{aligned}$$

Gaussian and related integrals

$$\int xe^{ax^2}dx = \frac{1}{2a}e^{ax^2}$$

$$\int \frac{1}{a^2+x^2}dx = \frac{1}{a}\tan^{-1}\frac{x}{a} \text{ thus } \int_{-\infty}^{\infty} \frac{1}{a^2+x^2}dx = \frac{\pi}{a}$$

$$\int_{-\infty}^{\infty} e^{-\frac{a^2x^2}{2}}dx = \frac{\sqrt{2\pi}}{a} \text{ thus } \frac{a}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{a^2x^2}{2}}dx = 1$$

$$\int_0^{\infty} x^2e^{-ax^2}dx = \frac{1}{4a}\sqrt{\frac{\pi}{a}}$$

$$\int_0^{\infty} x^{2n}e^{-\frac{x^2}{a^2}}dx = \sqrt{\pi}\frac{1\cdot 3\cdot 5\cdots(2n-1)}{2^{n+1}}a^{2n-1} = \sqrt{\pi}\frac{(2n)!}{n!}\left(\frac{a}{2}\right)^{2n+1}$$

$$\int_0^{\infty} x^{2n+1}e^{-\frac{x^2}{a^2}}dx = \frac{n!}{2}a^{2n+2}$$

$$\int_{-\infty}^{\infty} e^{-x^2}dx = \sqrt{\pi}$$

To verify $\int_{-\infty}^{\infty} e^{-x^2}dx = \sqrt{\pi}$, consider $\left(\int_{-\infty}^{\infty} e^{-x^2}dx\right)^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\left(x^2+y^2\right)}dxdy$. Let $x = r\cos\theta$ and $y = r\sin\theta$. The Jacobian of this transformation of variables is

$$J(r,\theta) = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r$$

Thus,

$$\left(\int_{-\infty}^{\infty} e^{-x^2}dx\right)^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\left(x^2+y^2\right)}dxdy = \int_0^{\infty}\int_0^{2\pi} e^{-r^2}J(r,\theta)\,drd\theta$$

$$= \int_0^{\infty} e^{-r^2}rdr\int_0^{2\pi} d\theta$$

$$= -2\pi\left[\frac{e^{-r^2}}{2}\right]_0^{\infty} = \pi$$

Thus, $\int_{-\infty}^{\infty} e^{-x^2}dx = \sqrt{\pi}$.

Miscellaneous integrals

$$\int_{x=0}^{1} x^{\alpha-1}(1-\alpha)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

For definition of the gamma function see Section 12.3 **Binomial coefficients**

The binomial coefficient $\binom{n}{k} = \frac{n!}{(n-k)!k!}$ is the number of ways of choosing k items from n. The number of ways of choosing $d+1$ items from $n+1$ items equals the number of ways of choosing the $d+1$ items from the first n items plus the number of ways of choosing d of the items from the first n items with the other item being the last of the $n+1$ items.

$$\binom{n}{d} + \binom{n}{d+1} = \binom{n+1}{d+1}.$$

The observation that the number of ways of choosing k items from $2n$ equals the number of ways of choosing i items from the first n and choosing $k-i$ items from the second n summed over all i, $0 \leq i \leq k$ yields the identity

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

Setting $k = n$ in the above formula and observing that $\binom{n}{i} = \binom{n}{n-i}$ yields

$$\sum_{i=0}^{n} \binom{n}{i}^2 = \binom{2n}{n}.$$

More generally $\sum_{i=0}^{k} \binom{n}{i}\binom{m}{k-i} = \binom{n+m}{k}$ by a similar derivation.

12.3 Useful Inequalities

$1 + x \leq e^x$ for all real x.

> One often establishes an inequality such as $1 + x \leq e^x$ by showing that the difference of the two sides, namely $e^x - (1 + x)$, is always positive. This can be done by taking derivatives. The first and second derivatives are $e^x - 1$ and e^x. Since e^x is always positive, $e^x - 1$ is monotonic and $e^x - (1 + x)$ is convex. Since $e^x - 1$ is monotonic, it can be zero only once and is zero at $x = 0$. Thus, $e^x - (1 + x)$ takes on its minimum at $x = 0$ where it is zero establishing the inequality.

$(1 - x)^n \geq 1 - nx$ for $0 \leq x \leq 1$

$1 + x \leq e^x$ for all real x

$(1-x)^n \geq 1 - nx$ for $0 \leq x \leq 1$

$(x+y)^2 \leq 2x^2 + 2y^2$

Triangle Inequality $\quad |\mathbf{x} + \mathbf{y}| \leq |\mathbf{x}| + |\mathbf{y}|$.

Cauchy-Schwartz Inequality $\quad |\mathbf{x}||\mathbf{y}| \geq \mathbf{x}^T\mathbf{y}$

Young's Inequality For positive real numbers p and q where $\frac{1}{p} + \frac{1}{q} = 1$ and positive reals x and y,

$$xy \leq \frac{1}{p}x^p + \frac{1}{q}y^q.$$

Hölder's inequalityHölder's inequality For positive real numbers p and q with $\frac{1}{p} + \frac{1}{q} = 1$,

$$\sum_{i=1}^{n} |x_i y_i| \ \leq \ \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \left(\sum_{i=1}^{n} |y_i|^q \right)^{1/q}.$$

Jensen's inequality For a convex function f,

$$f\left(\sum_{i=1}^{n} \alpha_i x_i \right) \leq \sum_{i=1}^{n} \alpha_i f\left(x_i\right),$$

Let $g(x) = (1-x)^n - (1-nx)$. We establish $g(x) \geq 0$ for x in $[0,1]$ by taking the derivative.

$$g'(x) = -n(1-x)^{n-1} + n = n\big(1 - (1-x)^{n-1}\big) \geq 0$$

for $0 \leq x \leq 1$. Thus, g takes on its minimum for x in $[0,1]$ at $x = 0$ where $g(0) = 0$ proving the inequality.

$(x+y)^2 \leq 2x^2 + 2y^2$

The inequality follows from $(x+y)^2 + (x-y)^2 = 2x^2 + 2y^2$.

Lemma 12.1 *For any nonnegative reals $a_1, a_2, \ldots, a_n$ and any $\rho \in [0,1]$, $\left(\sum_{i=1}^n a_i\right)^\rho \leq \sum_{i=1}^n a_i^\rho$.*

Proof: We will see that we can reduce the proof of the lemma to the case when only one of the a_i is nonzero and the rest are zero. To this end, suppose a_1 and a_2 are both positive and without loss of generality, assume $a_1 \geq a_2$. Add an infinitesimal positive amount ϵ to a_1 and subtract the same amount from a_2. This does not alter the left hand side. We claim it does not increase the right hand side. To see this, note that

$$(a_1 + \epsilon)^\rho + (a_2 - \epsilon)^\rho - a_1^\rho - a_2^\rho = \rho(a_1^{\rho-1} - a_2^{\rho-1})\epsilon + O(\epsilon^2),$$

and since $\rho - 1 \leq 0$, we have $a_1^{\rho-1} - a_2^{\rho-1} \leq 0$, proving the claim. Now by repeating this process, we can make $a_2 = 0$ (at that time a_1 will equal the sum of the original a_1 and a_2). Now repeating on all pairs of a_i, we can make all but one of them zero and in the process, we have left the left hand side the same, but have not increased the right hand side. So it suffices to prove the inequality at the end which clearly holds. This method of proof is called the variational method. ∎

The Triangle Inequality

For any two vectors $\mathbf{x}$ and $\mathbf{y}$, $|\mathbf{x}+\mathbf{y}| \leq |\mathbf{x}| + |\mathbf{y}|$. Since $\mathbf{x} \cdot \mathbf{y} \leq |\mathbf{x}||\mathbf{y}|$,

$$|\mathbf{x}+\mathbf{y}|^2 = (\mathbf{x}+\mathbf{y})^T \cdot (\mathbf{x}+\mathbf{y}) = |\mathbf{x}|^2 + |\mathbf{y}|^2 + 2\mathbf{x}^T \cdot \mathbf{y} \leq |\mathbf{x}|^2 + |\mathbf{y}|^2 + 2|\mathbf{x}||\mathbf{y}| = (|\mathbf{x}| + |\mathbf{y}|)^2.$$

The inequality follows by taking square roots.

Stirling approximation

$$n! \cong \left(\frac{n}{e}\right)^n \sqrt{2\pi n} \qquad \binom{2n}{n} \cong \frac{1}{\sqrt{\pi n}} 2^{2n}$$

$$\sqrt{2\pi n}\frac{n^n}{e^n} < n! < \sqrt{2\pi n}\frac{n^n}{e^n}\left(1 + \frac{1}{12n-1}\right)$$

We prove the inequalities, except for constant factors. Namely, we prove that

$$1.4 \quad \left(\frac{n}{e}\right)^n \sqrt{n} \leq n! \leq e\left(\frac{n}{e}\right)^n \sqrt{n}.$$

Write $\ln(n!) = \ln 1 + \ln 2 + \cdots + \ln n$. This sum is approximately $\int_{x=1}^{n} \ln x \, dx$. The indefinite integral $\int \ln x \, dx = (x \ln x - x)$ gives an approximation, but without the $\sqrt{n}$ term. To get the $\sqrt{n}$, differentiate twice and note that $\ln x$ is a concave function. This means that for any positive x_0,

$$\frac{\ln x_0 + \ln(x_0+1)}{2} \leq \int_{x=x_0}^{x_0+1} \ln x \, dx,$$

since for $x \in [x_0, x_0+1]$, the curve $\ln x$ is always above the spline joining $(x_0, \ln x_0)$ and $(x_0+1, \ln(x_0+1))$. Thus,

$$\begin{aligned}
\ln(n!) &= \frac{\ln 1}{2} + \frac{\ln 1 + \ln 2}{2} + \frac{\ln 2 + \ln 3}{2} + \cdots + \frac{\ln(n-1) + \ln n}{2} + \frac{\ln n}{2} \\
&\leq \int_{x=1}^{n} \ln x \, dx + \frac{\ln n}{2} = [x \ln x - x]_1^n + \frac{\ln n}{2} \\
&= n \ln n - n + 1 + \frac{\ln n}{2}.
\end{aligned}$$

Thus, $n! \leq n^n e^{-n}\sqrt{n}e$. For the lower bound on $n!$, start with the fact that for any $x_0 \geq 1/2$ and any real ρ

$$\ln x_0 \geq \frac{1}{2}(\ln(x_0+\rho)+\ln(x_0-\rho)) \quad \text{implies} \quad \ln x_0 \geq \int_{x=x_0-0.5}^{x_0+.5} \ln x \; dx.$$

Thus,

$$\ln(n!) = \ln 2 + \ln 3 + \cdots + \ln n \geq \int_{x=1.5}^{n+.5} \ln x \; dx,$$

from which one can derive a lower bound with a calculation.

Stirling approximation for the binomial coefficient

$$\binom{n}{k} \leq \left(\frac{en}{k}\right)^k$$

Using the Stirling approximation for $k!$,

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} \leq \frac{n^k}{k!} \cong \left(\frac{en}{k}\right)^k.$$

The gamma function

For $a > 0$

$$\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$$

$$\Gamma\left(\tfrac{1}{2}\right) = \sqrt{\pi}, \quad \Gamma(1) = \Gamma(2) = 1, \text{ and for } n \geq 2, \quad \Gamma(n) = (n-1)\Gamma(n-1).$$

The last statement is proved by induction on n. It is easy to see that $\Gamma(1) = 1$. For $n \geq 2$, we use integration by parts.

$$\int f(x)\, g'(x)\, dx = f(x)\, g(x) - \int f'(x)\, g(x)\, dx$$

Write $\Gamma(n) = \int_{x=0}^{\infty} f(x)g'(x)\; dx$, where, $f(x) = x^{n-1}$ and $g'(x) = e^{-x}$. Thus,

$$\Gamma(n) = [f(x)g(x)]_{x=0}^{\infty} + \int_{x=0}^{\infty} (n-1)x^{n-2}e^{-x}\; dx = (n-1)\Gamma(n-1),$$

as claimed.

Cauchy-Schwartz Inequality

$$\left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n y_i^2\right) \geq \left(\sum_{i=1}^n x_i y_i\right)^2$$

In vector form, $|\mathbf{x}||\mathbf{y}| \geq \mathbf{x}^T\mathbf{y}$, the inequality states that the dot product of two vectors is at most the product of their lengths. The Cauchy-Schwartz inequality is a special case of Hölder's inequality with $p = q = 2$.

Young's inequality

For positive real numbers p and q where $\frac{1}{p} + \frac{1}{q} = 1$ and positive reals x and y,

$$\frac{1}{p}x^p + \frac{1}{q}y^q \geq xy.$$

The left hand side of Young's inequality, $\frac{1}{p}x^p + \frac{1}{q}y^q$, is a convex combination of x^p and y^q since $\frac{1}{p}$ and $\frac{1}{q}$ sum to 1. $\ln(x)$ is a concave function for $x > 0$ and so the ln of the convex combination of the two elements is greater than or equal to the convex combination of the ln of the two elements

$$\ln(\frac{1}{p}x^p + \frac{1}{q}y^p) \geq \frac{1}{p}\ln(x^p) + \frac{1}{q}\ln(y^q) = \ln(xy).$$

Since for $x \geq 0$, $\ln x$ is a monotone increasing function, $\frac{1}{p}x^p + \frac{1}{q}y^q \geq xy$..

Hölder's inequalityHölder's inequality

For positive real numbers p and q with $\frac{1}{p} + \frac{1}{q} = 1$,

$$\sum_{i=1}^{n} |x_i y_i| \ \leq \ \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \left(\sum_{i=1}^{n} |y_i|^q\right)^{1/q}.$$

Let $x_i' = x_i \ / \ (\sum_{i=1}^n |x_i|^p)^{1/p}$ and $y_i' = y_i \ / \ (\sum_{i=1}^n |y_i|^q)^{1/q}$. Replacing x_i by x_i' and y_i by y_i' does not change the inequality. Now $\sum_{i=1}^n |x_i'|^p = \sum_{i=1}^n |y_i'|^q = 1$, so it suffices to prove $\sum_{i=1}^n |x_i'y_i'| \leq 1$. Apply Young's inequality to get $|x_i'y_i'| \leq \frac{|x_i'|^p}{p} + \frac{|y_i'|^q}{q}$. Summing over i, the right hand side sums to $\frac{1}{p} + \frac{1}{q} = 1$ finishing the proof.

For $a_1, a_2, \ldots, a_n$ real and k a positive integer,

$$(a_1 + a_2 + \cdots + a_n)^k \leq n^{k-1}(|a_1|^k + |a_2|^k + \cdots + |a_n|^k).$$

Using Hölder's inequality with $p = k$ and $q = k/(k-1)$,

$$\begin{aligned} |a_1 + a_2 + \cdots + a_n| &\leq |a_1 \cdot 1| + |a_2 \cdot 1| + \cdots + |a_n \cdot 1| \\ &\leq \left(\sum_{i=1}^{n} |a_i|^k\right)^{1/k} (1 + 1 + \cdots + 1)^{(k-1)/k}, \end{aligned}$$

Figure 12.1: Approximating sums by integrals

from which the current inequality follows.

Arithmetic and geometric means

The arithmetic mean of a set of nonnegative reals is at least their geometric mean. For $a_1, a_2, \ldots, a_n > 0$,

$$\frac{1}{n}\sum_{i=1}^{n} a_i \geq \sqrt[n]{a_1 a_2 \cdots a_n}.$$

Assume that $a_1 \geq a_2 \geq \ldots \geq a_n$. We reduce the proof to the case when all the a_i are equal using the variational method; in this case the inequality holds with equality. Suppose $a_1 > a_2$. Let ε be a positive infinitesimal. Add ε to a_2 and subtract ε from a_1 to get closer to the case when they are equal. The left hand side $\frac{1}{n}\sum_{i=1}^{n} a_i$ does not change.

$$\begin{aligned}(a_1 - \varepsilon)(a_2 + \varepsilon)a_3 a_4 \cdots a_n &= a_1 a_2 \cdots a_n + \varepsilon(a_1 - a_2)a_3 a_4 \cdots a_n + O(\varepsilon^2) \\ &> a_1 a_2 \cdots a_n\end{aligned}$$

for small enough $\varepsilon > 0$. Thus, the change has increased $\sqrt[n]{a_1 a_2 \cdots a_n}$. So if the inequality holds after the change, it must hold before. By continuing this process, one can make all the a_i equal.

Approximating sums by integrals

For monotonic decreasing $f(x)$,

$$\int_{x=m}^{n+1} f(x)dx \leq \sum_{i=m}^{n} f(i) \leq \int_{x=m-1}^{n} f(x)dx.$$

See Fig. 12.1. Thus,

$$\int_{x=2}^{n+1} \frac{1}{x^2}dx \leq \sum_{i=2}^{n} \frac{1}{i^2} = \frac{1}{4} + \frac{1}{9} + \cdots + \frac{1}{n^2} \leq \int_{x=1}^{n} \frac{1}{x^2}dx$$

and hence $\frac{3}{2} - \frac{1}{n+1} \leq \sum_{i=1}^{n} \frac{1}{i^2} \leq 2 - \frac{1}{n}$.

Jensen's Inequality

For a convex function f,

$$f\left(\frac{1}{2}(x_1+x_2)\right) \leq \frac{1}{2}(f(x_1)+f(x_2)).$$

More generally for any convex function f,

$$f\left(\sum_{i=1}^{n}\alpha_i x_i\right) \leq \sum_{i=1}^{n}\alpha_i f(x_i),$$

where $0 \leq \alpha_i \leq 1$ and $\sum_{i=1}^{n}\alpha_i = 1$. From this, it follows that for any convex function f and random variable x,

$$E(f(x)) \geq f(E(x)).$$

We prove this for a discrete random variable x taking on values $a_1, a_2, \ldots$ with Prob($x = a_i$) $= \alpha_i$:

$$E(f(x)) = \sum_i \alpha_i f(a_i) \geq f\left(\sum_i \alpha_i a_i\right) = f(E(x)).$$

Figure 12.2: For a convex function f, $f\left(\frac{x_1+x_2}{2}\right) \leq \frac{1}{2}(f(x_1)+f(x_2))$.

Example: Let $f(x) = x^k$ for k an even positive integer. Then, $f''(x) = k(k-1)x^{k-2}$ which since $k-2$ is even is nonnegative for all x implying that f is convex. Thus,

$$E(x) \leq \sqrt[k]{E(x^k)},$$

since $t^{\frac{1}{k}}$ is a monotone function of t, $t > 0$. It is easy to see that this inequality does not necessarily hold when k is odd; indeed for odd k, x^k is not a convex function. ∎

Tails of Gaussian

For bounding the tails of Gaussian densities, the following inequality is useful. The proof uses a technique useful in many contexts. For $t > 0$,

$$\int_{x=t}^{\infty} e^{-x^2}\,dx \leq \frac{e^{-t^2}}{2t}.$$

In proof, first write: $\int_{x=t}^{\infty} e^{-x^2}\,dx \leq \int_{x=t}^{\infty} \frac{x}{t}e^{-x^2}\,dx$, using the fact that $x \geq t$ in the range of integration. The latter expression is integrable in closed form since $d(e^{-x^2}) = (-2x)e^{-x^2}$

yielding the claimed bound.

A similar technique yields an upper bound on

$$\int_{x=\beta}^{1} (1-x^2)^\alpha \, dx,$$

for $\beta \in [0,1]$ and $\alpha > 0$. Just use $(1-x^2)^\alpha \leq \frac{x}{\beta}(1-x^2)^\alpha$ over the range and integrate in closed form the last expression.

$$\begin{aligned}\int_{x=\beta}^{1} (1-x^2)^\alpha dx \leq \int_{x=\beta}^{1} \frac{x}{\beta}(1-x^2)^\alpha dx &= \frac{-1}{2\beta(\alpha+1)}(1-x^2)^{\alpha+1}\Big|_{x=\beta}^{1} \\ &= \frac{(1-\beta^2)^{\alpha+1}}{2\beta(\alpha+1)}\end{aligned}$$

12.4 Probability

Consider an experiment such as flipping a coin whose outcome is determined by chance. To talk about the outcome of a particular experiment, we introduce the notion of a *random variable* whose value is the outcome of the experiment. The set of possible outcomes is called the *sample space*. If the sample space is finite, we can assign a probability of occurrence to each outcome. In some situations where the sample space is infinite, we can assign a probability of occurrence. The probability $p\,(i) = \frac{6}{\pi^2}\frac{1}{i^2}$ for i an integer greater than or equal to one is such an example. The function assigning the probabilities is called a *probability distribution function*.

In many situations, a probability distribution function does not exist. For example, for the uniform probability on the interval [0,1], the probability of any specific value is zero. What we can do is define a *probability density function* $p(x)$ such that

$$\text{Prob}(a < x < b) = \int_a^b p(x)dx$$

If x is a continuous random variable for which a density function exists, then the *cumulative distribution function* $f\,(a)$ is defined by

$$f(a) = \int_{-\infty}^{a} p(x)dx$$

which gives the probability that $x \leq a$.

12.4.1 Sample Space, Events, Independence

There may be more than one relevant random variable in a situation. For example, if one tosses n coins, there are n random variables, $x_1, x_2, \ldots, x_n$, taking on values 0 and 1, a 1 for heads and a 0 for tails. The set of possible outcomes, the sample space, is $\{0,1\}^n$. An *event* is a subset of the sample space. The event of an odd number of heads, consists of all elements of $\{0,1\}^n$ with an odd number of 1's.

Let A and B be two events. The joint occurrence of the two events is denoted by $(A \wedge B)$. The *conditional probability* of event A given that event B has occurred is denoted by Prob$(A|B)$and is given by

$$\text{Prob}(A|B) = \frac{\text{Prob}(A \wedge B)}{\text{Prob}(B)}.$$

Events A and B are *independent* if the occurrence of one event has no influence on the probability of the other. That is, Prob$(A|B) =$ Prob(A) or equivalently, Prob$(A \wedge B) =$ Prob(A)Prob(B). Two random variables x and y are *independent* if for every possible set A of values for x and every possible set B of values for y, the events x in A and y in B are independent.

A collection of n random variables $x_1, x_2, \ldots, x_n$ is *mutually independent* if for all possible sets $A_1, A_2, \ldots, A_n$ of values of $x_1, x_2, \ldots, x_n$,

$$\text{Prob}(x_1 \in A_1, x_2 \in A_2, \ldots, x_n \in A_n) = \text{Prob}(x_1 \in A_1)\text{Prob}(x_2 \in A_2) \cdots \text{Prob}(x_n \in A_n).$$

If the random variables are discrete, it would suffice to say that for any real numbers $a_1, a_2, \ldots, a_n$

$$\text{Prob}(x_1 = a_1, x_2 = a_2, \ldots, x_n = a_n) = \text{Prob}(x_1 = a_1)\text{Prob}(x_2 = a_2) \cdots \text{Prob}(x_n = a_n).$$

Random variables $x_1, x_2, \ldots, x_n$ are pairwise independent if for any a_i and a_j, $i \neq j$, Prob$(x_i = a_i, x_j = a_j) =$ Prob$(x_i = a_i)$Prob$(x_j = a_j)$. Mutual independence is much stronger than requiring that the variables are pairwise independent. Consider the example of 2-universal hash functions discussed in Chapter **??**.

If (x, y) is a random vector and one normalizes it to a unit vector $\left(\frac{x}{\sqrt{x^2+y^2}}, \frac{y}{\sqrt{x^2+y^2}}\right)$ the coordinates are no longer independent since knowing the value of one coordinate uniquely determines the value of the other.

12.4.2 Linearity of Expectation

An important concept is that of the expectation of a random variable. The *expected value*, $E(x)$, of a random variable x is $E(x) = \sum_x xp(x)$ in the discrete case and $E(x) =$

$\int_{-\infty}^{\infty} xp(x)dx$ in the continuous case. The expectation of a sum of random variables is equal to the sum of their expectations. The linearity of expectation follows directly from the definition and does not require independence.

12.4.3 Union Bound

Let $A_1, A_2, \ldots, A_n$ be events. The actual probability of the union of events is given by Boole's formula.

$$\text{Prob}(A_1 \cup A_2 \cup \cdots A_n) = \sum_{i=1}^{n} \text{Prob}(A_i) - \sum_{ij} \text{Prob}(A_i \wedge A_j) + \sum_{ijk} \text{Prob}(A_i \wedge A_j \wedge A_k) - \cdots$$

Often we only need an upper bound on the probability of the union and use

$$\text{Prob}(A_1 \cup A_2 \cup \cdots A_n) \leq \sum_{i=1}^{n} \text{Prob}(A_i)$$

This upper bound is called the *union bound.*

12.4.4 Indicator Variables

A useful tool is that of an indicator variable that takes on value 0 or 1 to indicate whether some quantity is present or not. The indicator variable is useful in determining the expected size of a subset. Given a random subset of the integers $\{1, 2, \ldots, n\}$, the expected size of the subset is the expected value of $x_1 + x_2 + \cdots + x_n$ where x_i is the indicator variable that takes on value 1 if i is in the subset.

Example: Consider a random permutation of n integers. Define the indicator function $x_i = 1$ if the i^{th} integer in the permutation is i. The expected number of fixed points is given by

$$E\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} E(x_i) = n\frac{1}{n} = 1.$$

Note that the x_i are not independent. But, linearity of expectation still applies. ∎

Example: Consider the expected number of vertices of degree d in a random graph $G(n, p)$. The number of vertices of degree d is the sum of n indicator random variables, one for each vertex, with value one if the vertex has degree d. The expectation is the sum of the expectations of the n indicator random variables and this is just n times the expectation of one of them. Thus, the expected number of degree d vertices is $n\binom{n}{d}p^d(1-p)^{n-d}$. ∎

12.4.5 Variance

In addition to the expected value of a random variable, another important parameter is the variance. The *variance* of a random variable x, denoted var(x) or often $\sigma^2(x)$ is $E\left(x - E\left(x\right)\right)^2$ and measures how close to the expected value the random variable is likely to be. The *standard deviation* σ is the square root of the variance. The units of σ are the same as those of x.

By linearity of expectation

$$\sigma^2 = E\left(x - E\left(x\right)\right)^2 = E(x^2) - 2E(x)E(x) + E^2(x) = E\left(x^2\right) - E^2\left(x\right).$$

12.4.6 Variance of the Sum of Independent Random Variables

In general, the variance of the sum is not equal to the sum of the variances. However, if x and y are independent, then $E\left(xy\right) = E\left(x\right)E\left(y\right)$ and

$$var(x + y) = var\left(x\right) + var\left(y\right).$$

To see this

$$\begin{aligned} var(x + y) &= E\left((x + y)^2\right) - E^2(x + y) \\ &= E(x^2) + 2E(xy) + E(y^2) - E^2(x) - 2E(x)E(y) - E^2(y). \end{aligned}$$

From independence, $2E(xy) - 2E(x)E(y) = 0$ and

$$\begin{aligned} var(x + y) &= E(x^2) - E^2(x) + E(y^2) - E^2(y) \\ &= var(x) + var(y). \end{aligned}$$

More generally, if $x_1, x_2, \ldots, x_n$ are pairwise independent random variables, then

$$var(x_1 + x_2 + \cdots + x_n) = var(x_1) + var(x_2) + \cdots + var(x_n).$$

For the variance of the sum to be the sum of the variances only requires pairwise independence not full independence.

12.4.7 Median

One often calculates the average value of a random variable to get a feeling for the magnitude of the variable. This is reasonable when the probability distribution of the variable is Gaussian, or has a small variance. However, if there are outliers, then the average may be distorted by outliers. An alternative to calculating the expected value is to calculate the median, the value for which half of the probability is above and half is below.

12.4.8 The Central Limit Theorem

Let $s = x_1 + x_2 + \cdots + x_n$ be a sum of n independent random variables where each x_i has probability distribution

$$x_i = \begin{cases} 0 & \frac{1}{2} \\ 1 & \frac{1}{2} \end{cases} .$$

The expected value of each x_i is $1/2$ with variance

$$\sigma_i^2 = \left(\frac{1}{2} - 0\right)^2 \frac{1}{2} + \left(\frac{1}{2} - 1\right)^2 \frac{1}{2} = \frac{1}{4}.$$

The expected value of s is $n/2$ and since the variables are independent, the variance of the sum is the sum of the variances and hence is $n/4$. How concentrated s is around its mean depends on the standard deviation of s which is $\frac{\sqrt{n}}{2}$. For n equal 100 the expected value of s is 50 with a standard deviation of 5 which is 10% of the mean. For $n = 10,000$ the expected value of s is 5,000 with a standard deviation of 50 which is 1% of the mean. Note that as n increases, the standard deviation increases, but the ratio of the standard deviation to the mean goes to zero. More generally, if x_i are independent and identically distributed, each with standard deviation σ, then the standard deviation of $x_1 + x_2 + \cdots + x_n$ is $\sqrt{n}\sigma$. So, $\frac{x_1+x_2+\cdots+x_n}{\sqrt{n}}$ has standard deviation σ. The central limit theorem makes a stronger assertion that in fact $\frac{x_1+x_2+\cdots+x_n}{\sqrt{n}}$ has Gaussian distribution with standard deviation σ.

Theorem 12.2 *Suppose $x_1, x_2, \ldots, x_n$ is a sequence of identically distributed independent random variables, each with mean μ and variance σ^2. The distribution of the random variable*

$$\frac{1}{\sqrt{n}} (x_1 + x_2 + \cdots + x_n - n\mu)$$

converges to the distribution of the Gaussian with mean 0 and variance σ^2.

12.4.9 Probability Distributions

The Gaussian or normal distribution

The normal distribution is

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-m)^2}{\sigma^2}}$$

where m is the mean and σ^2 is the variance. The coefficient $\frac{1}{\sqrt{2\pi}\sigma}$ makes the integral of the distribution be one. If we measure distance in units of the standard deviation σ from the mean, then

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

Standard tables give values of the integral

$$\int_0^t \phi(x)dx$$

and from these values one can compute probability integrals for a normal distribution with mean m and variance σ^2.

General Gaussians

So far we have seen spherical Gaussian densities in $\mathbf{R}^d$. The word spherical indicates that the level curves of the density are spheres. If a random vector $\mathbf{y}$ in $\mathbf{R}^d$ has a spherical Gaussian density with zero mean, then y_i and y_j, $i \neq j$, are independent. However, in many situations the variables are correlated. To model these Gaussians, level curves that are ellipsoids rather than spheres are used.

For a random vector $\mathbf{x}$, the covariance of x_i and x_j is $E((x_i - \mu_i)(x_j - \mu_j))$. We list the covariances in a matrix called the *covariance matrix*, denoted Σ.[1] Since $\mathbf{x}$ and $\boldsymbol{\mu}$ are column vectors, $(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T$ is a $d \times d$ matrix. Expectation of a matrix or vector means componentwise expectation.

$$\Sigma = E\big((\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T\big).$$

The general Gaussian density with mean $\boldsymbol{\mu}$ and positive definite covariance matrix Σ is

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right).$$

To compute the covariance matrix of the Gaussian, substitute $\mathbf{y} = \Sigma^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$. Noting that a positive definite symmetric matrix has a square root:

$$\begin{aligned} E((\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T &= E(\Sigma^{1/2}\mathbf{y}\mathbf{y}^T\Sigma^{1/2}) \\ &= \Sigma^{1/2}\left(E(\mathbf{y}\mathbf{y}^T)\right)\Sigma^{1/2} = \Sigma. \end{aligned}$$

The density of $\mathbf{y}$ is the unit variance, zero mean Gaussian, thus $E(yy^T) = I$.

Bernoulli trials and the binomial distribution

A Bernoulli trial has two possible outcomes, called success or failure, with probabilities p and $1 - p$, respectively. If there are n independent Bernoulli trials, the probability of exactly k successes is given by the *binomial distribution*

$$B(n, p) = \binom{n}{k} p^k (1 - p)^{n-k}$$

[1] Σ is the standard notation for the covariance matrix. We will use it sparingly so as not to confuse with the summation sign.

The mean and variance of the binomial distribution $B(n,p)$ are np and $np(1-p)$, respectively. The mean of the binomial distribution is np, by linearity of expectations. The variance is $np(1-p)$ since the variance of a sum of independent random variables is the sum of their variances.

Let x_1 be the number of successes in n_1 trials and let x_2 be the number of successes in n_2 trials. The probability distribution of the sum of the successes, $x_1 + x_2$, is the same as the distribution of $x_1 + x_2$ successes in $n_1 + n_2$ trials. Thus, $B(n_1,p) + B(n_2,p) = B(n_1+n_2,p)$.
When p is a constant, the expected degree of vertices in $G(n,p)$ increases with n. For example, in $G\left(n,\frac{1}{2}\right)$, the expected degree of a vertex is $n/2$. In many real applications, we will be concerned with $G(n,p)$ where $p = d/n$, for d a constant; i.e., graphs whose expected degree is a constant d independent of n. Holding $d = np$ constant as n goes to infinity, the binomial distribution

$$\text{Prob}(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

approaches the Poisson distribution

$$\text{Prob}(k) = \frac{(np)^k}{k!} e^{-np} = \frac{d^k}{k!} e^{-d}.$$

move text beginning here to appendix
To see this, assume $k = o(n)$ and use the approximations $n - k \cong n$, $\binom{n}{k} \cong \frac{n^k}{k!}$, and $\left(1-\frac{1}{n}\right)^{n-k} \cong e^{-1}$ to approximate the binomial distribution by

$$\lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} = \frac{n^k}{k!} \left(\frac{d}{n}\right)^k (1-\frac{d}{n})^n = \frac{d^k}{k!} e^{-d}.$$

Note that for $p = \frac{d}{n}$, where d is a constant independent of n, the probability of the binomial distribution falls off rapidly for $k > d$, and is essentially zero for all but some finite number of values of k. This justifies the $k = o(n)$ assumption. Thus, the Poisson distribution is a good approximation.
end of material to move

Poisson distribution

The Poisson distribution describes the probability of k events happening in a unit of time when the average rate per unit of time is λ. Divide the unit of time into n segments. When n is large enough, each segment is sufficiently small so that the probability of two events happening in the same segment is negligible. The Poisson distribution gives the probability of k events happening in a unit of time and can be derived from the binomial distribution by taking the limit as $n \to \infty$.

Let $p = \frac{\lambda}{n}$. Then

$$\begin{aligned}\text{Prob}(k \text{ successes in a unit of time}) &= \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-k} \\ &= \lim_{n\to\infty} \frac{\lambda^k}{k!} e^{-\lambda}\end{aligned}$$

In the limit as n goes to infinity the binomial distribution $p(k) = \binom{n}{k} p^k (1-p)^{n-k}$ becomes the Poisson distribution $p(k) = e^{-\lambda}\frac{\lambda^k}{k!}$. The mean and the variance of the Poisson distribution have value λ. If x and y are both Poisson random variables from distributions with means λ_1 and λ_2 respectively, then $x + y$ is Poisson with mean $m_1 + m_2$. For large n and small p the binomial distribution can be approximated with the Poisson distribution.

The binomial distribution with mean np and variance $np(1-p)$ can be approximated by the normal distribution with mean np and variance $np(1-p)$. The central limit theorem tells us that there is such an approximation in the limit. The approximation is good if both np and $n(1-p)$ are greater than 10 provided k is not extreme. Thus,

$$\binom{n}{k}\left(\frac{1}{2}\right)^k\left(\frac{1}{2}\right)^{n-k} \cong \frac{1}{\sqrt{\pi n/2}} e^{-\frac{(n/2-k)^2}{\frac{1}{2}n}}.$$

This approximation is excellent provided k is $\Theta(n)$. The Poisson approximation

$$\binom{n}{k} p^k (1-p)^k \cong e^{-np}\frac{(np)^k}{k!}$$

is off for central values and tail values even for $p = 1/2$. The approximation

$$\left(\begin{array}{c} n \\ k \end{array}\right) p^k (1-p)^{n-k} \cong \frac{1}{\sqrt{\pi pn}} e^{-\frac{(pn-k)^2}{pn}}$$

is good for $p = 1/2$ but is off for other values of p.

Generation of random numbers according to a given probability distribution

Suppose one wanted to generate a random variable with probability density $p(x)$ where $p(x)$ is continuous. Let $P(x)$ be the cumulative distribution function for x and let u be a random variable with uniform probability density over the interval [0,1]. Then the random variable $x = P^{-1}(u)$ has probability density $p(x)$.

Example: For a Cauchy density function the cumulative distribution function is

$$P(x) = \int_{t=-\infty}^{x} \frac{1}{\pi}\frac{1}{1+t^2}dt = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(x).$$

Setting $u = P(x)$ and solving for x yields $x = \tan\left(\pi\left(u - \frac{1}{2}\right)\right)$. Thus, to generate a random number $x \geq 0$ using the Cauchy distribution, generate u, $0 \leq u \leq 1$, uniformly and calculate $x = \tan\left(\pi\left(u - \frac{1}{2}\right)\right)$. The value of x varies from $-\infty$ to ∞ with $x = 0$ for $u = 1/2$. ∎

12.4.10 Bayes Rule and Estimators

Bayes rule

Bayes rule relates the conditional probability of A given B to the conditional probability of B given A.

$$\text{Prob}(A|B) = \frac{\text{Prob}(B|A)\,\text{Prob}(A)}{\text{Prob}(B)}$$

Suppose one knows the probability of A and wants to know how this probability changes if we know that B has occurred. Prob(A) is called the prior probability. The conditional probability Prob($A|B$) is called the posterior probability because it is the probability of A after we know that B has occurred.

The example below illustrates that if a situation is rare, a highly accurate test will often give the wrong answer.

Example: Let A be the event that a product is defective and let B be the event that a test says a product is defective. Let Prob($B|A$) be the probability that the test says a product is defective assuming the product is defective and let Prob$\left(B|\bar{A}\right)$ be the probability that the test says a product is defective if it is not actually defective.

What is the probability Prob($A|B$) that the product is defective if the test say it is defective? Suppose Prob(A) = 0.001, Prob($B|A$) = 0.99, and Prob$\left(B|\bar{A}\right) = 0.02$. Then

$$\begin{aligned}\text{Prob}(B) &= \text{Prob}(B|A)\,\text{Prob}(A) + \text{Prob}\left(B|\bar{A}\right)\text{Prob}\left(\bar{A}\right)\\ &= 0.99 \times 0.001 + 0.02 \times 0.999\\ &= 0.02087\end{aligned}$$

and

$$\text{Prob}(A|B) = \frac{\text{Prob}(B|A)\,\text{Prob}(A)}{\text{Prob}(B)} \approx \frac{0.99 \times 0.001}{0.0210} = 0.0471$$

Even though the test fails to detect a defective product only 1% of the time when it is defective and claims that it is defective when it is not only 2% of the time, the test is correct only 4.7% of the time when it says a product is defective. This comes about because of the low frequencies of defective products. ∎

The words prior, a posteriori, and likelihood come from Bayes theorem.

$$\text{a posteriori} = \frac{\text{likelihood } \times \text{ prior}}{\text{normalizing constant}}$$

$$\text{Prob}\,(A|B) = \frac{\text{Prob}\,(B|A)\,\text{Prob}\,(A)}{\text{Prob}\,(B)}$$

The a posteriori probability is the conditional probability of A given B. The likelihood is the conditional probability $\text{Prob}(B|A)$.

Unbiased Estimators

Consider n samples $x_1, x_2, \ldots, x_n$ from a Gaussian distribution of mean μ and variance σ^2. For this distribution, $m = \frac{x_1+x_2+\cdots+x_n}{n}$ is an unbiased estimator of μ, which means that $E(m) = \mu$ and $\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$ is an unbiased estimator of σ^2. However, if μ is not known and is approximated by m, then $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2$ is an unbiased estimator of σ^2.

Maximum Likelihood Estimation MLE

Suppose the probability distribution of a random variable x depends on a parameter r. With slight abuse of notation, since r is a parameter rather than a random variable, we denote the probability distribution of x as $p\,(x|r)$. This is the likelihood of observing x if r was in fact the parameter value. The job of the maximum likelihood estimator, MLE, is to find the best r after observing values of the random variable x. The likelihood of r being the parameter value given that we have observed x is denoted $L(r|x)$. This is again not a probability since r is a parameter, not a random variable. However, if we were to apply Bayes' rule as if this was a conditional probability, we get

$$L(r|x) = \frac{\text{Prob}(x|r)\text{Prob}(r)}{\text{Prob}(x)}.$$

Now, assume $\text{Prob}(r)$ is the same for all r. The denominator $\text{Prob}(x)$ is the absolute probability of observing x and is independent of r. So to maximize $L(r|x)$, we just maximize $\text{Prob}(x|r)$. In some situations, one has a prior guess as to the distribution $\text{Prob}(r)$. This is then called the "prior" and in that case, we call $\text{Prob}(x|r)$ the posterior which we try to maximize.

Example: Consider flipping a coin 100 times. Suppose 62 heads and 38 tails occur. What is the most likely value of the probability of the coin to come down heads when the coin is flipped? In this case, it is $r = 0.62$. The probability that we get 62 heads if the unknown probability of heads in one trial is r is

$$\text{Prob}\,(62 \text{ heads}|r) = \binom{100}{62} r^{62}(1-r)^{38}.$$

This quantity is maximized when $r = 0.62$. To see this take the logarithm, which as a function of r is $\ln \binom{100}{62} + 62 \ln r + 38 \ln(1-r)$. The derivative with respect to r is zero at $r = 0.62$ and the second derivative is negative indicating a maximum. Thus, $r = 0.62$ is the maximum likelihood estimator of the probability of heads in a trial. ∎

12.4.11 Tail Bounds and Chernoff inequalities

Markov's inequality bounds the probability that a nonnegative random variable exceeds a value a.

$$p(x \geq a) \leq \frac{E(x)}{a}.$$

or

$$p\left(x \geq aE(x)\right) \leq \frac{1}{a}$$

If one also knows the variance, σ^2, then using Chebyshev's inequality one can bound the probability that a random variable differs from its expected value by more than a standard deviations.

$$p(|x - m| \geq a\sigma) \leq \frac{1}{a^2}$$

If a random variable s is the sum of n independent random variables $x_1, x_2, \ldots, x_n$ of finite variance, then better bounds are possible. For any $\delta > 0$,

$$\text{Prob}(s > (1+\delta)m) < \left[\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right]^m$$

and for $0 < \gamma \leq 1$,

$$\text{Prob}\left(s < (1-\gamma)m\right) < \left[\frac{e^{-\gamma}}{(1+\gamma)^{(1+\gamma)}}\right]^m < e^{-\frac{\gamma^2 m}{2}}$$

Chernoff inequalities

Chebyshev's inequality bounds the probability that a random variable will deviate from its mean by more than a given amount. Chebyshev's inequality holds for any probability distribution. For some distributions we can get much tighter bounds. For example, the probability that a Gaussian random variable deviates from its mean falls off exponentially with the distance from the mean. Here we shall be concerned with the situation where we have a random variable that is the sum of n independent random variables. This is another situation in which we can derive a tighter bound than that given by the Chebyshev inequality. We consider the case where the n independent variables are binomial but similar results can be shown for independent random variables from any distribution that has a finite variance.

Let $x_1, x_2, \ldots, x_n$ be independent random variables where

$$x_i = \begin{cases} 0 & \text{Prob } 1-p \\ 1 & \text{Prob } \quad p \end{cases}.$$

Consider the sum $s = \sum_{i=1}^{n} x_i$. Here the expected value of each x_i is p and by linearity of expectation, the expected value of the sum is $m{=}np$. Theorem **??** bounds the probability that the sum s exceeds $(1+\delta)\,m$.

Theorem 12.3 *For any $\delta > 0$, Prob $\left(s > (1+\delta)m\right) < \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{m}$*

Proof: For any $\lambda > 0$, the function $e^{\lambda x}$ is monotone. Thus,

$$\text{Prob}\left(s > (1+\delta)m\right) = \text{Prob}\left(e^{\lambda s} > e^{\lambda(1+\delta)m}\right).$$

$e^{\lambda x}$ is nonnegative for all x, so we can apply Markov's inequality to get

$$\text{Prob}\left(e^{\lambda s} > e^{\lambda(1+\delta)m}\right) \leq e^{-\lambda(1+\delta)m} E\left(e^{\lambda s}\right).$$

Since the x_i are independent,

$$\begin{aligned} E\left(e^{\lambda s}\right) = E\left(e^{\lambda \sum_{i=1}^{n} x_i}\right) = E\left(\prod_{i=1}^{n} e^{\lambda x_i}\right) = \prod_{i=1}^{n} E\left(e^{\lambda x_i}\right) \\ = \prod_{i=1}^{n} \left(e^{\lambda} p + 1 - p\right) = \prod_{i=1}^{n} \left(p(e^{\lambda} - 1) + 1\right). \end{aligned}$$

Using the inequality $1 + x < e^{x}$ with $x = p(e^{\lambda} - 1)$ yields

$$E\left(e^{\lambda s}\right) < \prod_{i=1}^{n} e^{p(e^{\lambda}-1)}.$$

Thus, for all $\lambda > 0$

$$\begin{aligned} \text{Prob}\left(s > (1+\delta)m\right) &\leq \text{Prob}\left(e^{\lambda s} > e^{\lambda(1+\delta)m}\right) \\ &\leq e^{-\lambda(1+\delta)m} E\left(e^{\lambda s}\right) \\ &\leq e^{-\lambda(1+\delta)m} \prod_{i=1}^{n} e^{p(e^{\lambda}-1)}. \end{aligned}$$

Setting $\lambda = \ln(1+\delta)$

$$\begin{aligned} \text{Prob}\left(s > (1+\delta)m\right) &\leq \left(e^{-\ln(1+\delta)}\right)^{(1+\delta)m} \prod_{i=1}^{n} e^{p(e^{\ln(1+\delta)}-1)} \\ &\leq \left(\frac{1}{1+\delta}\right)^{(1+\delta)m} \prod_{i=1}^{n} e^{p\delta} \\ &\leq \left(\frac{1}{(1+\delta)}\right)^{(1+\delta)m} e^{np\delta} \\ &\leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{m}. \end{aligned}$$

■

To simplify the bound of Theorem 12.3, observe that

$$(1+\delta)\ln(1+\delta) = \delta + \frac{\delta^2}{2} - \frac{\delta^3}{6} + \frac{\delta^4}{12} - \cdots .$$

Therefore

$$(1+\delta)^{(1+\delta)} = e^{\delta + \frac{\delta^2}{2} - \frac{\delta^3}{6} + \frac{\delta^4}{12} - \cdots}$$

and hence

$$\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}} = e^{-\frac{\delta^2}{2} + \frac{\delta^3}{6} - \cdots}.$$

Thus, the bound simplifies to

$$\text{Prob}\left(s < (1+\delta)\,m\right) \leq e^{-\frac{\delta^2}{2}m + \frac{\delta^3}{6}m - \cdots}.$$

For small δ the probability drops exponentially with δ^2.

When δ is large another simplification is possible. First

$$\text{Prob}\left(s > (1+\delta)\,m\right) \leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{m} \leq \left(\frac{e}{1+\delta}\right)^{(1+\delta)m}$$

If $\delta > 2e-1$, substituting $2e-1$ for δ in the denominator yields

$$\text{Prob}(s > (1+\delta)\,m) \leq 2^{-(1+\delta)m}.$$

Theorem 12.3 gives a bound on the probability of the sum being greater than the mean. We now bound the probability that the sum will be less than its mean.

Theorem 12.4 *Let* $0 < \gamma \leq 1$, *then* $\text{Pr}\,ob\left(s < (1-\gamma)m\right) < \left(\frac{e^{-\gamma}}{(1+\gamma)^{(1+\gamma)}}\right)^{m} < e^{-\frac{\gamma^2 m}{2}}$.

Proof: For any $\lambda > 0$

$$\text{Prob}\left(s < (1-\gamma)m\right) = \text{Prob}\left(-s > -(1-\gamma)m\right) = \text{Prob}\left(e^{-\lambda s} > e^{-\lambda(1-\gamma)m}\right).$$

Applying Markov's inequality

$$\text{Prob}\left(s < (1-\gamma)m\right) < \frac{E(e^{-\lambda x})}{e^{-\lambda(1-\gamma)m}} < \frac{\prod_{i=1}^{n} E(e^{-\lambda X_i})}{e^{-\lambda(1-\gamma)m}}.$$

Now

$$E(e^{-\lambda x_i}) = pe^{-\lambda} + 1 - p = 1 + p(e^{-\lambda} - 1) + 1.$$

Thus,

$$\text{Prob}(s < (1-\gamma)m) < \frac{\prod_{i=1}^{n}[1+p(e^{-\lambda}-1)]}{e^{-\lambda(1-\gamma)m}}.$$

Since $1+x < e^x$

$$\text{Prob}\left(s < (1-\gamma)m\right) < \frac{e^{np(e^{-\lambda}-1)}}{e^{-\lambda(1-\gamma)m}}.$$

Setting $\lambda = \ln\frac{1}{1-\gamma}$

$$\begin{aligned}\text{Prob}\left(s < (1-\gamma)m\right) &< \frac{e^{np(1-\gamma-1)}}{(1-\gamma)^{(1-\gamma)m}} \\ &< \left(\frac{e^{-\gamma}}{(1-\gamma)^{(1-\gamma)}}\right)^m.\end{aligned}$$

But for $0 < \gamma \leq 1$, $(1-\gamma)^{(1-\gamma)} > e^{-\gamma+\frac{\gamma^2}{2}}$. To see this note that

$$\begin{aligned}(1-\gamma)\ln(1-\gamma) &= (1-\gamma)\left(-\gamma - \frac{\gamma^2}{2} - \frac{\gamma^3}{3} - \cdots\right) \\ &= -\gamma - \frac{\gamma^2}{2} - \frac{\gamma^3}{3} - \cdots + \gamma^2 + \frac{\gamma^3}{2} + \frac{\gamma^4}{3} + \cdots \\ &= -\gamma + \left(\gamma^2 - \frac{\gamma^2}{2}\right) + \left(\frac{\gamma^3}{2} - \frac{\gamma^3}{3}\right) + \cdots \\ &= -\gamma + \frac{\gamma^2}{2} + \frac{\gamma^3}{6} + \cdots \\ &\geq -\gamma + \frac{\gamma^2}{2}.\end{aligned}$$

It then follows that

$$\text{Prob}\left(s < (1-\gamma)m\right) < \left(\frac{e^{-\gamma}}{(1-\gamma)^{(1-\gamma)}}\right)^m < e^{-\frac{m\gamma^2}{2}}.$$

■

12.5 Eigenvalues and Eigenvectors

12.5.1 Eigenvalues and Eigenvectors

Let A be an $n \times n$ real matrix. The scalar λ is called an eigenvalue of A if there exists a nonzero vector $\mathbf{x}$ satisfying the equation $A\mathbf{x} = \lambda\mathbf{x}$. The vector $\mathbf{x}$ is called the eigenvector of A associated with λ. The set of all eigenvectors associated with a given eigenvalue form a subspace as seen from the fact that if $A\mathbf{x} = \lambda\mathbf{x}$ and $A\mathbf{y} = \lambda\mathbf{y}$, then for any scalers c and d, $A(c\mathbf{x} + d\mathbf{y}) = \lambda(c\mathbf{x} + d\mathbf{y})$. The equation $A\mathbf{x} = \lambda\mathbf{x}$ has a nontrivial solution only if

$\det(A - \lambda I) = 0$. The equation $\det(A - \lambda I) = 0$ is called the *characteristic* equation and has n not necessarily distinct roots.

Matrices A and B are similar if there is an invertible matrix P such that $A = P^{-1}BP$.

Theorem 12.5 *If A and B are similar, then they have the same eigenvalues.*

Proof: Let A and B be similar matrices. Then there exists an invertible matrix P such that $A = P^{-1}BP$. For an eigenvector $\mathbf{x}$ of A with eigenvalue λ, $A\mathbf{x} = \lambda\mathbf{x}$, which implies $P^{-1}BP\mathbf{x} = \lambda\mathbf{x}$ or $B(P\mathbf{x}) = \lambda(P\mathbf{x})$. So, $P\mathbf{x}$ is an eigenvector of B with the same eigenvalue λ. Since the reverse also holds, the theorem follows. ∎

Even though two similar matrices, A and B, have the same eigenvalues, their eigenvectors are in general different.

The matrix A is *diagonalizable* if A is similar to a diagonal matrix.

Theorem 12.6 *A is diagonalizable if and only if A has n linearly independent eigenvectors.*

Proof:

(only if) Assume A is diagonalizable. Then there exists an invertible matrix P and a diagonal matrix D such that $D = P^{-1}AP$. Thus, $PD = AP$. Let the diagonal elements of D be $\lambda_1, \lambda_2, \ldots, \lambda_n$ and let $\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_n}$ be the columns of P. Then $AP = [A\mathbf{p_1}, A\mathbf{p_2}, \ldots, A\mathbf{p_n}]$ and $PD = [\lambda_1\mathbf{p_1}, \lambda_2\mathbf{p_2}, \ldots, \lambda_n\mathbf{p_n}]$. Hence $A\mathbf{p_i} = \lambda_i\mathbf{p_i}$. That is, the λ_i are the eigenvalues of A and the $\mathbf{p_i}$ are the corresponding eigenvectors. Since P is invertible, the $\mathbf{p_i}$ are linearly independent.

(if) Assume that A has n linearly independent eigenvectors $\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_n}$ with corresponding eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. Then $A\mathbf{p_i} = \lambda_i\mathbf{p_i}$ and reversing the above steps

$$AP = [A\mathbf{p_1}, A\mathbf{p_2}, \ldots, A\mathbf{p_n}] = [\lambda_1\mathbf{p_1}, \lambda_2\mathbf{p_2}, \ldots \lambda_n\mathbf{p_n}] = PD.$$

Thus, $AP = DP$. Since the $\mathbf{p_i}$ are linearly independent, P is invertible and hence $A = P^{-1}DP$. Thus, A is diagonalizable. ∎

It follows from the proof of the theorem that if A is diagonalizable and has eigenvalue λ with multiplicity k, then there are k linearly independent eigenvectors associated with λ.

A matrix P is *orthogonal* if it is invertible and $P^{-1} = P^T$. A matrix A is *orthogonally diagonalizable* if there exists an orthogonal matrix P such that $P^{-1}AP = D$ is diagonal. If A is orthogonally diagonalizable, then $A = PDP^T$ and $AP = PD$. Thus, the columns of P are the eigenvectors of A and the diagonal elements of D are the corresponding

eigenvalues.

If P is an orthogonal matrix, then P^TAP and A are both representations of the same linear transformation with respect to different bases. To see this, note that if $\mathbf{e_1}, \mathbf{e_2}, \ldots, \mathbf{e_n}$ is the standard basis, then a_{ij} is the component of $A\mathbf{e_j}$ along the direction $\mathbf{e_i}$, namely, $a_{ij} = \mathbf{e_i}^T A\mathbf{e_j}$. Thus, A defines a linear transformation by specifying the image under the transformation of each basis vector. Denote by $\mathbf{p_j}$ the j^{th} column of P. It is easy to see that $(P^TAP)_{ij}$ is the component of $A\mathbf{p_j}$ along the direction $\mathbf{p_i}$, namely, $(P^TAP)_{ij} = \mathbf{p_i}^T A\mathbf{p_j}$. Since P is orthogonal, the $\mathbf{p_j}$ form a basis of the space and so P^TAP represents the same linear transformation as A, but in the basis $p_1, p_2, \ldots, p_n$.

Another remark is in order. Check that

$$A = PDP^T = \sum_{i=1}^{n} d_{ii}\mathbf{p_i}\mathbf{p_i}^T.$$

Compare this with the singular value decomposition where

$$A = \sum_{i=1}^{n} \sigma_i \mathbf{u_i}\mathbf{v_i}^T,$$

the only difference being that $\mathbf{u_i}$ and $\mathbf{v_i}$ can be different and indeed if A is not square, they will certainly be.

12.5.2 Symmetric Matrices

For an arbitrary matrix, some of the eigenvalues may be complex. However, for a symmetric matrix with real entries, all eigenvalues are real. The number of eigenvalues of a symmetric matrix, counting multiplicities, equals the dimension of the matrix. The set of eigenvectors associated with a given eigenvalue form a vector space. For a non-symmetric matrix, the dimension of this space may be less than the multiplicity of the eigenvalue. Thus, a nonsymmetric matrix may not be diagonalizable. However, for a symmetric matrix the eigenvectors associated with a given eigenvalue form a vector space of dimension equal to the multiplicity of the eigenvalue. Thus, all symmetric matrices are diagonalizable. The above facts for symmetric matrices are summarized in the following theorem.

Theorem 12.7 (Real Spectral Theorem) *Let A be a real symmetric matrix. Then*

1. *The eigenvalues, $\lambda_1, \lambda_2, \ldots, \lambda_n$, are real, as are the components of the corresponding eigenvectors, $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_n}$.*
2. **(Spectral Decomposition)** *A is orthogonally diagonalizable and indeed*

$$A = VDV^T = \sum_{i=1}^{n} \lambda_i \mathbf{v_i}\mathbf{v_i}^T,$$

where V is the matrix with columns $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_n}$, $|\mathbf{v_i}| = 1$ and D is a diagonal matrix with entries $\lambda_1, \lambda_2, \ldots, \lambda_n$.

Proof: $A\mathbf{v_i} = \lambda_i \mathbf{v_i}$ and $\mathbf{v_i}^c A\mathbf{v_i} = \lambda_i \mathbf{v_i}^c \mathbf{v_i}$. Here the c superscript means conjugate transpose. Then

$$\lambda_i = \mathbf{v_i}^c A\mathbf{v_i} = (\mathbf{v_i}^c A\mathbf{v_i})^{cc} = (\mathbf{v_i}^c A^c \mathbf{v_i})^c = (\mathbf{v_i}^c A\mathbf{v_i})^c = \lambda_i^c$$

and hence λ_i is real.

Since λ_i is real, a nontrivial solution to $(A - \lambda_i I)\,\mathbf{x} = 0$ has real components.

Let P be a real symmetric matrix such that $P\mathbf{v_1} = \mathbf{e_1}$ where $\mathbf{e_1} = (1, 0, 0, \ldots, 0)^T$ and $P^{-1} = P^T$. We will construct such a P shortly. Since $A\mathbf{v_1} = \lambda_1 \mathbf{v_1}$,

$$PAP^T\mathbf{e_1} = PAv_1 = \lambda Pv_1 = \lambda_1 \mathbf{e_1}.$$

The condition $PAP^T\mathbf{e_1} = \lambda_1 \mathbf{e_1}$ plus symmetry implies that $PAP^T = \begin{pmatrix} \lambda_1 & 0 \\ 0 & A' \end{pmatrix}$ where A' is $n-1$ by $n-1$ and symmetric. By induction, A' is orthogonally diagonalizable. Let Q be the orthogonal matrix with $QA'Q^T = D'$, a diagonal matrix. Q is $(n-1) \times (n-1)$. Augment Q to an $n \times n$ matrix by putting 1 in the $(1,1)$ position and 0 elsewhere in the first row and column. Call the resulting matrix R. R is orthogonal too.

$$R \begin{pmatrix} \lambda_1 & 0 \\ 0 & A' \end{pmatrix} R^T = \begin{pmatrix} \lambda_1 & 0 \\ 0 & D' \end{pmatrix} \quad \Longrightarrow \quad RPAP^T R^T = \begin{pmatrix} \lambda_1 & 0 \\ 0 & D' \end{pmatrix}.$$

Since the product of two orthogonal matrices is orthogonal, this finishes the proof of (2) except it remains to construct P. For this, take an orthonormal basis of space containing $\mathbf{v_1}$. Suppose the basis is $\{\mathbf{v_1}, \mathbf{w_2}, \mathbf{w_3}, \ldots\}$ and V is the matrix with these basis vectors as its columns. Then $P = V^T$ will do. ∎

Theorem 12.8 (The fundamental theorem of symmetric matrices) *A real matrix A is orthogonally diagonalizable if and only if A is symmetric.*

Proof: (if) Assume A is orthogonally diagonalizable. Then there exists P such that $D = P^{-1}AP$. Since $P^{-1} = P^T$, we get

$$A = PDP^{-1} = PDP^T$$

which implies

$$A^T = (PDP^T)^T = PDP^T = A$$

and hence A is symmetric.
(only if) Already roved. ∎

Note that a nonsymmetric matrix may not be diagonalizable, it may have eigenvalues that are not real, and the number of linearly independent eigenvectors corresponding to an eigenvalue may be less than its multiplicity. For example, the matrix

$$\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

has eigenvalues 2, $\frac{1}{2} + i\frac{\sqrt{3}}{2}$, and $\frac{1}{2} - i\frac{\sqrt{3}}{2}$. The matrix $\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$ has characteristic equation $(1-\lambda)^2 = 0$ and thus has eigenvalue 1 with multiplicity 2 but has only one linearly independent eigenvector associated with the eigenvalue 1, namely $\mathbf{x} = c\begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad c \neq 0$. Neither of these situations is possible for a symmetric matrix.

12.5.3 Relationship between SVD and Eigen Decomposition

The singular value decomposition exists for any $n \times d$ matrix whereas the eigenvalue decomposition exists only for certain square matrices. For symmetric matrices the decompositions are essentially the same.

The singular values of a matrix are always positive since they are the sum of squares of the projection of a row of a matrix onto a singular vector. Given a symmetric matrix, the eigenvalues can be positive or negative. If A is a symmetric matrix with eigenvalue decomposition $A = V_E D_E V_E^T$ and singular value decomposition $A = U_S D_S V_S^T$, what is the relationship between D_E and D_S, and between V_E and V_S, and between U_S and V_E? Observe that if A can be expressed as QDQ^T where Q is orthonormal and D is diagonal, then $AQ = QD$. That is, each column of Q is an eigenvector and the elements of D are the eigenvalues. Thus, if the eigenvalues of A are distinct, then Q is unique up to a permutation of columns. If an eigenvalue has multiplicity k, then the space spanned the k columns is unique. In the following we will use the term essentially unique to capture this situation. Now $AA^T = U_S D_S^2 U_S^T$ and $A^T A = V_S D_S^2 V_S^T$. By an argument similar to the one above, U_S and V_S are essentially unique and are the eigenvectors or negatives of the eigenvectors of A and A^T. The eigenvalues of AA^T or $A^T A$ are the squares of the eigenvalues of A. If A is not positive semi definite and has negative eigenvalues, then in the singular value decomposition $A = U_S D_S V_S$, some of the left singular vectors are the negatives of the eigenvectors. Let S be a diagonal matrix with $\pm 1's$ on the diagonal depending on whether the corresponding eigenvalue is positive or negative. Then $A = (U_S S)(S D_S) V_S$ where $U_S S = V_E$ and $S D_S = D_E$.

12.5.4 Extremal Properties of Eigenvalues

In this section we derive a min max characterization of eigenvalues that implies that the largest eigenvalue of a symmetric matrix A has a value equal to the maximum of

$\mathbf{x}^T A\mathbf{x}$ over all vectors $\mathbf{x}$ of unit length. That is, the largest eigenvalue of A equals the 2-norm of A. If A is a real symmetric matrix there exists an orthogonal matrix P that diagonalizes A. Thus

$$P^T AP = D$$

where D is a diagonal matrix with the eigenvalues of A, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, on its diagonal. Rather than working with A, it is easier to work with the diagonal matrix D. This will be an important technique that will simplify many proofs.

Consider maximizing $\mathbf{x}^T A\mathbf{x}$ subject to the conditions

1. $\sum_{i=1}^{n} x_i^2 = 1$
2. $\mathbf{r}_i^T \mathbf{x} = 0, \quad 1 \leq i \leq s$

where the r_i are any set of nonzero vectors. We ask over all possible sets $\{r_i | 1 \leq i \leq s\}$ of s vectors, what is the minimum value assumed by this maximum.

Theorem 12.9 (Min max theorem) *For a symmetric matrix A, $\min_{\mathbf{r}_1,\ldots,\mathbf{r}_s} \max_{\substack{\mathbf{x} \\ \mathbf{r}_i \perp \mathbf{x}}} (\mathbf{x}^t A\mathbf{x}) = \lambda_{s+1}$ where the minimum is over all sets $\{r_1, r_2, \ldots, r_s\}$ of s nonzero vectors and the maximum is over all unit vectors x orthogonal to the s nonzero vectors.*

Proof: A is orthogonally diagonalizable. Let P satisfy $P^T P = I$ and $P^T AP = D$, D diagonal. Let $\mathbf{y} = P^T \mathbf{x}$. Then $\mathbf{x} = P\mathbf{y}$ and

$$\mathbf{x}^T A\mathbf{x} = \mathbf{y}^T P^T AP\mathbf{y} = \mathbf{y}^T D\mathbf{y} = \sum_{i=1}^{n} \lambda_i y_i^2$$

Since there is a one-to-one correspondence between unit vectors $\mathbf{x}$ and $\mathbf{y}$, maximizing $\mathbf{x}^T A\mathbf{x}$ subject to $\sum x_i^2 = 1$ is equivalent to maximizing $\sum_{i=1}^{n} \lambda_i y_i^2$ subject to $\sum y_i^2 = 1$. Since $\lambda_1 \geq \lambda_i$, $2 \leq i \leq n$, $\mathbf{y} = (1, 0, \ldots, 0)$ maximizes $\sum_{i=1}^{n} \lambda_i y_i^2$ at λ_1. Then $\mathbf{x} = P\mathbf{y}$ is the first column of P and is the first eigenvector of A. Similarly λ_n is the minimum value of $\mathbf{x}^T A\mathbf{x}$ subject to the same conditions.

Now consider maximizing $\mathbf{x}^T A\mathbf{x}$ subject to the conditions

1. $\sum x_i^2 = 1$
2. $\mathbf{r}_i^T \mathbf{x} = 0$

where the $\mathbf{r}_i$ are any set of nonzero vectors. We ask over all possible choices of s vectors what is the minimum value assumed by this maximum.

$$\min_{\mathbf{r}_1,\ldots,\mathbf{r}_s} \max_{\substack{\mathbf{x} \\ \mathbf{r}_i^T \mathbf{x} = 0}} \mathbf{x}^T A\mathbf{x}$$

As above, we may work with $\mathbf{y}$. The conditions are

1. $\sum y_i^2 = 1$
2. $\mathbf{q}_i^T \mathbf{y} = 0$ where, $\mathbf{q}_i^T = \mathbf{r}_i^T P$

Consider any choice for the vectors $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_s$. This gives a corresponding set of $\mathbf{q}_i$. The $\mathbf{y}_i$ therefore satisfy s linear homogeneous equations. If we add $y_{s+2} = y_{s+3} = \cdots y_n = 0$ we have $n-1$ homogeneous equations in n unknowns $y_1, \ldots, y_n$. There is at least one solution that can be normalized so that $\sum y_i^2 = 1$. With this choice of $\mathbf{y}$

$$\mathbf{y}^T D \mathbf{y} = \sum \lambda_i y_i^2 \geq \lambda_{s+1}$$

since coefficients greater than or equal to $s+1$ are zero. Thus, for any choice of $\mathbf{r_i}$ there will be a $\mathbf{y}$ such that

$$\max_{\substack{\mathbf{y} \\ \mathbf{r}_i^T \mathbf{y} = 0}} (\mathbf{y}^T P^T A P \mathbf{y}) \geq \lambda_{s+1}$$

and hence

$$\min_{\mathbf{r_1}, \mathbf{r_2}, \ldots, \mathbf{r_s}} \max_{\substack{\mathbf{y} \\ \mathbf{r}_i^T \mathbf{y} = 0}} (\mathbf{y}^T P^T A P \mathbf{y}) \geq \lambda_{s+1}.$$

However, there is a set of s constraints for which the minimum is less than or equal to λ_{s+1}. Fix the relations to be $y_i = 0, \quad 1 \leq i \leq s$. There are s equations in n unknowns and for any $\mathbf{y}$ subject to these relations

$$\mathbf{y}^T D \mathbf{y} = \sum_{s+1}^{n} \lambda_i y_i^2 \leq \lambda_{s+1}.$$

Combining the two inequalities, $\min \max \mathbf{y}^T D \mathbf{y} = \lambda_{s+1}$. ∎

The above theorem tells us that the maximum of $\mathbf{x}^T A \mathbf{x}$ subject to the constraint that $|\mathbf{x}|^2 = 1$ is λ_1. Consider the problem of maximizing $\mathbf{x}^T A \mathbf{x}$ subject to the additional restriction that $\mathbf{x}$ is orthogonal to the first eigenvector. This is equivalent to maximizing $\mathbf{y}^t P^t A P \mathbf{y}$ subject to $\mathbf{y}$ being orthogonal to (1,0,...,0), i.e. the first component of $\mathbf{y}$ being 0. This maximum is clearly λ_2 and occurs for $\mathbf{y} = (0, 1, 0, \ldots, 0)$. The corresponding $\mathbf{x}$ is the second column of P or the second eigenvector of A.

Similarly the maximum of $\mathbf{x}^T A \mathbf{x}$ for $\mathbf{p_1}^T \mathbf{x} = \mathbf{p_2}^T \mathbf{x} = \cdots \mathbf{p_s}^T \mathbf{x} = 0$ is λ_{s+1} and is obtained for $\mathbf{x} = \mathbf{p_{s+1}}$.

12.5.5 Eigenvalues of the Sum of Two Symmetric Matrices

The min max theorem is useful in proving many other results. The following theorem shows how adding a matrix B to a matrix A changes the eigenvalues of A. The theorem is useful for determining the effect of a small perturbation on the eigenvalues of A.

Theorem 12.10 *Let A and B be $n \times n$ symmetric matrices. Let C=A+B. Let α_i, β_i, and γ_i denote the eigenvalues of A, B, and C respectively, where $\alpha_1 \geq \alpha_2 \geq \ldots \alpha_n$ and similarly for β_i, γ_i. Then $\alpha_s + \beta_1 \geq \gamma_s \geq \alpha_s + \beta_n$.*

Proof: By the min max theorem we have

$$\alpha_s = \min_{\mathbf{r}_1,\ldots,\mathbf{r}_{s-1}} \max_{\substack{\mathbf{x} \\ \mathbf{r}_i \perp \mathbf{x}}} (\mathbf{x}^T A \mathbf{x}).$$

Suppose $\mathbf{r_1}, \mathbf{r_2}, \ldots, \mathbf{r_{s-1}}$ attain the minimum in the expression. Then using the min max theorem on C,

$$\begin{aligned} \gamma_s &\leq \max_{\mathbf{x} \perp \mathbf{r_1}, \mathbf{r_2}, \ldots \mathbf{r_{s-1}}} \left(\mathbf{x}^T (A+B) \mathbf{x}\right) \\ &\leq \max_{\mathbf{x} \perp \mathbf{r_1}, \mathbf{r_2}, \ldots \mathbf{r_{s-1}}} (\mathbf{x}^T A \mathbf{x}) + \max_{\mathbf{x} \perp \mathbf{r_1}, \mathbf{r_2}, \ldots \mathbf{r_{s-1}}} (\mathbf{x}^T B \mathbf{x}) \\ &\leq \alpha_s + \max_{\mathbf{x}} (\mathbf{x}^T B \mathbf{x}) \leq \alpha_s + \beta_1. \end{aligned}$$

Therefore, $\gamma_s \leq \alpha_s + \beta_1$.

An application of the result to $A = C + (-B)$, gives $\alpha_s \leq \gamma_s - \beta_n$. The eigenvalues of -$B$ are minus the eigenvalues of B and thus $-\beta_n$ is the largest eigenvalue. Hence $\gamma_s \geq \alpha_s + \beta_n$ and combining inequalities yields $\alpha_s + \beta_1 \geq \gamma_s \geq \alpha_s + \beta_n$. ∎

Lemma 12.11 *Let A and B be $n \times n$ symmetric matrices. Let C=A+B. Let α_i, β_i, and γ_i denote the eigenvalues of A, B, and C respectively, where $\alpha_1 \geq \alpha_2 \geq \ldots \alpha_n$ and similarly for β_i, γ_i. Then $\gamma_{r+s-1} \leq \alpha_r + \beta_s$.*

Proof: There is a set of $r-1$ relations such that over all $\mathbf{x}$ satisfying the $r-1$ relationships

$$\max(\mathbf{x}^T A \mathbf{x}) = \alpha_r.$$

And a set of $s-1$ relations such that over all $\mathbf{x}$ satisfying the $s-1$ relationships

$$\max(\mathbf{x}^T B \mathbf{x}) = \beta_s.$$

Consider $\mathbf{x}$ satisfying all these $r+s-2$ relations. For any such $\mathbf{x}$

$$\mathbf{x}^T C \mathbf{x} = \mathbf{x}^T A \mathbf{x} + \mathbf{x}^T B \mathbf{x} x \leq \alpha_r + \beta_s$$

and hence over all the $\mathbf{x}$

$$\max(\mathbf{x}^T C \mathbf{x}) \leq \alpha_s + \beta_r$$

Taking the minimum over all sets of $r+s-2$ relations

$$\gamma_{r+s-1} = \min \max(\mathbf{x}^T C \mathbf{x}) \leq \alpha_r + \beta_s$$

∎

12.5.6 Norms

A set of vectors $\{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ is *orthogonal* if $\mathbf{x_i}^T\mathbf{x_j} = 0$ for $i \neq j$ and is *orthonormal* if in addition $|\mathbf{x_i}| = 1$ for all i. A matrix A is *orthonormal* if $A^TA = I$. If A is a square orthonormal matrix, then rows as well as columns are orthogonal. In other words, if A is square orthonormal, then A^T is also. In the case of matrices over the complexes, the concept of an orthonormal matrix is replaced by that of a unitary matrix. A^* is the conjugate transpose of A if $a^*_{ij} = \bar{a}_{ji}$ where a^*_{ij} is the ij^{th} entry of A^* and $\bar{a}^*_{ij}$ is the complex conjugate of the ij^{th} element of A. A matrix A over the field of complex numbers is **unitary** if $AA^* = I$.

Norms

A **norm** on $\mathbf{R}^n$ is a function $f : \mathbf{R}^n \to \mathbf{R}$ satisfying the following three axioms:

1. $f(\mathbf{x}) \geq 0$,
2. $f(\mathbf{x} + \mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$, and
3. $f(\alpha\mathbf{x}) = |\alpha| f(\mathbf{x})$.

A norm on a vector space provides a distance function where

$$\text{distance}(\mathbf{x}, \mathbf{y}) = norm(\mathbf{x} - \mathbf{y}).$$

An important class of norms for vectors is the p-norms defined for $p > 0$ by

$$|\mathbf{x}|_p = (|\mathbf{x_1}|^p + \cdots + |\mathbf{x_n}|^p)^{\frac{1}{p}}.$$

Important special cases are

$$\begin{aligned}
&|\mathbf{x}|_0 \text{ the number of non zero entries} \\
&|\mathbf{x}|_1 = |x_1| + \cdots + |x_n| \\
&|\mathbf{x}|_2 = \sqrt{|x_1|^2 + \cdots + |x_n|^2} \\
&|\mathbf{x}|_\infty = \max |x_i|.
\end{aligned}$$

Lemma 12.12 *For any* $1 \leq p < q$, $|\mathbf{x}|_q \leq |\mathbf{x}|_p$.

Proof:

$$|\mathbf{x}|_q^q = \sum_i |x_i|^q.$$

Let $a_i = |x_i|^q$ and $\rho = p/q$. Using Jensen's inequality (see Section 12.3) that for any nonnegative reals $a_1, a_2, \ldots, a_n$ and any $\rho \in (0, 1)$, we have $(\sum_{i=1}^n a_i)^\rho \leq \sum_{i=1}^n a_i^\rho$, the lemma is proved. ∎

There are two important matrix norms, the matrix p-norm

$$||A||_p = \max_{|\mathbf{x}|=1} \|A\mathbf{x}\|_p$$

and the Frobenius norm

$$||A||_F = \sqrt{\sum_{ij} a_{ij}^2}.$$

Let $\mathbf{a_i}$ be the i^{th} column of A. Then $\|A\|_F^2 = \sum_i \mathbf{a_i}^T \mathbf{a_i} = tr\left(A^T A\right)$. A similar argument on the rows yields $\|A\|_F^2 = tr\left(AA^T\right)$. Thus, $\|A\|_F^2 = tr\left(A^T A\right) = tr\left(AA^T\right)$.
If A is symmetric and rank k

$$||A||_2^2 \leq ||A||_F^2 \leq k\,||A||_2^2.$$

12.5.7 Important Norms and Their Properties

Lemma 12.13 $||AB||_2 \leq ||A||_2\,||B||_2$

Proof: $||AB||_2 = \max_{|\mathbf{x}|=1} |AB\mathbf{x}|$. Let $\mathbf{y}$ be the value of $\mathbf{x}$ that achieves the maximum and let $\mathbf{z} = B\mathbf{y}$. Then

$$||AB||_2 = |AB\mathbf{y}| = |A\mathbf{z}| = \left|A\frac{\mathbf{z}}{|\mathbf{z}|}\right| |\mathbf{z}|$$

But $\left|A\frac{\mathbf{z}}{|\mathbf{z}|}\right| \leq \max_{|\mathbf{x}|=1} |A\mathbf{x}| = ||A||_2$ and $|\mathbf{z}| \leq \max_{|\mathbf{x}|=1} |B\mathbf{x}| = ||B||_2$. Thus $||AB||_2 \leq ||A||_2\,||B||_2$. ∎

Let Q be an orthonormal matrix.

Lemma 12.14 *For all* $\mathbf{x}$, $|Q\mathbf{x}| = |\mathbf{x}|$.

Proof: $|Q\mathbf{x}|_2^2 = \mathbf{x}^T Q^T Q\mathbf{x} = \mathbf{x}^T\mathbf{x} = |\mathbf{x}|_2^2$. ∎

Lemma 12.15 $||QA||_2 = ||A||_2$

Proof: For all $\mathbf{x}$, $|Q\mathbf{x}| = |\mathbf{x}|$. Replacing $\mathbf{x}$ by $A\mathbf{x}$, $|QA\mathbf{x}| = |A\mathbf{x}|$ and thus $\max_{|\mathbf{x}|=1} |QA\mathbf{x}| = \max_{|\mathbf{x}|=1} |A\mathbf{x}|$ ∎

Lemma 12.16 $||AB||_F^2 \leq ||A||_F^2\,||B||_F^2$

Proof: Let $\mathbf{a_i}$ be the i^{th} column of A and let $\mathbf{b_j}$ be the j^{th} column of B. By the Cauchy-Schwartz inequality $\left\|\mathbf{a_i}^T\mathbf{b_j}\right\| \leq \|\mathbf{a_i}\|\,\|\mathbf{b_j}\|$. Thus $||AB||_F^2 = \sum_i \sum_j \left|\mathbf{a_i}^T\mathbf{b_j}\right|^2 \leq \sum_i \sum_j \|\mathbf{a_i}\|^2\,\|\mathbf{b_j}\|^2 = \sum_i \|\mathbf{a_i}\|^2 \sum_j \|\mathbf{b_j}\|^2 = ||A||_F^2\,||B||_F^2$ ∎

Lemma 12.17 $||QA||_F = ||A||_F$

Proof: $||QA||_F^2 = \mathrm{Tr}(A^T Q^T Q A) = \mathrm{Tr}(A^T A) = ||A||_F^2$. ∎

Lemma 12.18 *For real, symmetric matrix A with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots$, $\|A\|_2^2 = \max(\lambda_1^2, \lambda_n^2)$ and $\|A\|_F^2 = \lambda_1^2 + \lambda_2^2 + \cdots + \lambda_n^2$*

Proof: Suppose the spectral decomposition of A is PDP^T, where P is an orthogonal matrix and D is diagonal. We saw that $||P^T A||_2 = ||A||_2$. Applying this again, $||P^T AP||_2 = ||A||_2$. But, $P^T AP = D$ and clearly for a diagonal matrix D, $||D||_2$ is the largest absolute value diagonal entry from which the first equation follows. The proof of the second is analogous. ∎

If A is real and symmetric and of rank k then $||A||_2^2 \leq ||A||_F^2 \leq k\,||A||_2^2$

Theorem 12.19 $||A||_2^2 \leq ||A||_F^2 \leq k\,||A||_2^2$

Proof: It is obvious for diagonal matrices that $||D||_2^2 \leq ||D||_F^2 \leq k\,||D||_2^2$. Let $D = Q^t AQ$ where Q is orthonormal. The result follows immediately since for Q orthonormal, $||QA||_2 = ||A||_2$ and $||QA||_F = ||A||_F$. ∎

Real and symmetric are necessary for some of these theorems. This condition was needed to express $\Sigma = Q^T AQ$. For example, in Theorem 12.19 suppose A is the $n \times n$ matrix

$$A = \begin{pmatrix} 1 & 1 & \\ 1 & 1 & \\ \vdots & \vdots & 0 \\ 1 & 1 & \end{pmatrix}.$$

$||A||_2 = 2$ and $||A||_F = \sqrt{2n}$. But A is rank 2 and $||A||_F > 2\,||A||_2$ for $n > 8$.

Lemma 12.20 *Let A be a symmetric matrix. Then $\|A\|_2 = \max_{|\mathbf{x}|=1} \left|\mathbf{x}^T A\mathbf{x}\right|$.*

Proof: By definition, the 2-norm of A is $\|A\|_2 = \max_{|\mathbf{x}|=1} |A\mathbf{x}|$. Thus,

$$\|A\|_2 = \max_{|\mathbf{x}|=1} |A\mathbf{x}| = \max_{|\mathbf{x}|=1} \sqrt{\mathbf{x}^T A^T A\mathbf{x}} = \sqrt{\lambda_1^2} = \lambda_1 = \max_{|\mathbf{x}|=1} \left|\mathbf{x}^T A\mathbf{x}\right|$$

∎

The two norm of a matrix A is greater than or equal to the 2-norm of any of its columns. Let $\mathbf{a_u}$ be a column of A.

Lemma 12.21 $|\mathbf{a_u}| \leq \|A\|_2$

Proof: Let $\mathbf{e_u}$ be the unit vector with a 1 in position u and all other entries zero. Note $\lambda = \max_{|x|=1} |Ax|$. Let $\mathbf{x} = \mathbf{e_u}$ where $\mathbf{a_u}$ is row u. Then $|\mathbf{a_u}| = |A\mathbf{e_u}| \leq \max_{|\mathbf{x}|=1} |A\mathbf{x}| = \lambda$ ∎

12.5.8 Linear Algebra

Lemma 12.22 *Let A be an $n \times n$ symmetric matrix. Then* $\det(A) = \lambda_1 \lambda_2 \cdots \lambda_n$.

Proof: The $\det(A - \lambda I)$ is a polynomial in λ of degree n. The coefficient of λ^n will be ± 1 depending on whether n is odd or even. Let the roots of this polynomial be $\lambda_1, \lambda_2, \ldots, \lambda_n$. Then $\det(A - \lambda I) = (-1)^n \prod_{i=1}^{n} (\lambda - \lambda_i)$. Thus

$$\det(A) = \det(A - \lambda I)|_{\lambda=0} = (-1)^n \prod_{i=1}^{n} (\lambda - \lambda_i) \Bigg|_{\lambda=0} = \lambda_1 \lambda_2 \cdots \lambda_n$$

∎

The trace of a matrix is defined to be the sum of its diagonal elements. That is, $\mathrm{tr}(A) = a_{11} + a_{22} + \cdots + a_{nn}$.

Lemma 12.23 *$tr(A) = \lambda_1 + \lambda_2 + \cdots + \lambda_n$.*

Proof: Consider the coefficient of λ^{n-1} in $\det(A - \lambda I) = (-1)^n \prod_{i=1}^{n} (\lambda - \lambda_i)$. Write

$$A - \lambda I = \begin{pmatrix} a_{11} - \lambda & a_{12} & \cdots \\ a_{21} & a_{22} - \lambda & \cdots \\ \vdots & \vdots & \vdots \end{pmatrix}.$$

Calculate $\det(A - \lambda I)$ by expanding along the first row. Each term in the expansion involves a determinant of size $n - 1$ which is a polynomial in λ of deg $n - 2$ except for the principal minor which is of deg $n - 1$. Thus the term of deg $n - 1$ comes from

$$(a_{11} - \lambda)(a_{22} - \lambda) \cdots (a_{nn} - \lambda)$$

and has coefficient $(-1)^{n-1}(a_{11} + a_{22} + \cdots + a_{nn})$. Now

$$\begin{aligned} (-1)^n \prod_{i=1}^{n} (\lambda - \lambda_i) &= (-1)^n (\lambda - \lambda_1)(\lambda - \lambda_2) \cdots (\lambda - \lambda_n) \\ &= (-1)^n \left(\lambda^n - (\lambda_1 + \lambda_2 + \cdots + \lambda_n)\lambda^{n-1} + \cdots \right) \end{aligned}$$

Therefore equating coefficients $\lambda_1 + \lambda_2 + \cdots + \lambda_n = a_{11} + a_{22} + \cdots + a_{nn} = tr(A)$

Note that $(tr(A))^2 \neq tr(A^2)$. For example $A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ has trace 3, $A^2 = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$ has trace 5 $\neq$9. However $tr(A^2) = \lambda_1^2 + \lambda_2^2 + \cdots + \lambda_n^2$. To see this, observe that $A^2 = (V^T D V)^2 = V^T D^2 V$. Thus, the eigenvalues of A^2 are the squares of the eigenvalues for A. ∎

Alternative proof that $tr(A) = \lambda_1 + \lambda_2 + \cdots + \lambda_n$. Suppose the spectral decomposition of A is $A = PDP^T$. We have

$$\mathrm{tr}\,(A) = tr\left(PDP^T\right) = \mathrm{tr}\left(DP^TP\right) = \mathrm{tr}\,(D) = \lambda_1 + \lambda_2 + \cdots + \lambda_n.$$

Lemma 12.24 *If A is $n \times m$ and B is a $m \times n$ matrix, then tr(AB)=tr(BA).*

$$\mathrm{tr}(AB) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}b_{ji} = \sum_{j=1}^{n}\sum_{i=1}^{n} b_{ji}a_{ij} = \mathrm{tr}\,(BA)$$

Pseudo inverse

Let A be an $n \times m$ rank r matrix and let $A = U\Sigma V^T$ be the singular value decomposition of A. Let $\Sigma' = diag\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_r}, 0, \ldots, 0\right)$ where $\sigma_1, \ldots, \sigma_r$ are the nonzero singular values of A. Then $A' = V\Sigma'U^T$ is the pseudo inverse of A. It is the unique X that minimizes $\|AX - I\|_F$.

Second eigenvector

Suppose the eigenvalues of a matrix are $\lambda_1 \geq \lambda_2 \geq \cdots$. The second eigenvalue, λ_2, plays an important role for matrices representing graphs. It may be the case that $|\lambda_n| > |\lambda_2|$.

Why is the second eigenvalue so important? Consider partitioning the vertices of a regular degree d graph $G = (V, E)$ into two blocks of equal size so as to minimize the number of edges between the two blocks. Assign value +1 to the vertices in one block and -1 to the vertices in the other block. Let $\mathbf{x}$ be the vector whose components are the ± 1 values assigned to the vertices. If two vertices, i and j, are in the same block, then x_i and x_j are both +1 or both –1 and $(x_i - x_j)^2 = 0$. If vertices i and j are in different blocks then $(x_i - x_j)^2 = 4$. Thus, partitioning the vertices into two blocks so as to minimize the edges between vertices in different blocks is equivalent to finding a vector $\mathbf{x}$ with coordinates ± 1 of which half of its coordinates are +1 and half of which are –1 that minimizes

$$E_{cut} = \frac{1}{4}\sum_{(i,j)\in E} (x_i - x_j)^2$$

Let A be the adjacency matrix of G. Then

$$\begin{aligned}
&\mathbf{x}^T A\mathbf{x} = \sum_{ij} a_{ij}x_ix_j = 2\sum_{edges} x_ix_j \\
&= 2 \times \begin{pmatrix} \text{number of edges} \\ \text{within components} \end{pmatrix} - 2 \times \begin{pmatrix} \text{number of edges} \\ \text{between components} \end{pmatrix} \\
&= 2 \times \begin{pmatrix} \text{total number} \\ \text{of edges} \end{pmatrix} - 4 \times \begin{pmatrix} \text{number of edges} \\ \text{between components} \end{pmatrix}
\end{aligned}$$

Maximizing $\mathbf{x}^T A\mathbf{x}$ over all $\mathbf{x}$ whose coordinates are ± 1 and half of whose coordinates are $+1$ is equivalent to minimizing the number of edges between components.

Since finding such an $\mathbf{x}$ is computational difficult, replace the integer condition on the components of $\mathbf{x}$ and the condition that half of the components are positive and half of the components are negative with the conditions $\sum_{i=1}^{n} x_i^2 = 1$ and $\sum_{i=1}^{n} x_i = 0$. Then finding the optimal $\mathbf{x}$ gives us the second eigenvalue since it is easy to see that the first eigenvector Is along $\mathbf{1}$

$$\lambda_2 = \max_{\mathbf{x} \perp \mathbf{v_1}} \frac{\mathbf{x}^T A\mathbf{x}}{\sum x_i^2}$$

Actually we should use $\sum_{i=1}^{n} x_i^2 = n$ not $\sum_{i=1}^{n} x_i^2 = 1$. Thus $n\lambda_2$ must be greater than $2 \times \begin{pmatrix} \text{total number} \\ \text{of edges} \end{pmatrix} - 4 \times \begin{pmatrix} \text{number of edges} \\ \text{between components} \end{pmatrix}$ since the maximum is taken over a larger set of $\mathbf{x}$. The fact that λ_2 gives us a bound on the minimum number of cross edges is what makes it so important.

12.5.9 Distance between subspaces

Suppose S_1 and S_2 are two subspaces. Choose a basis of S_1 and arrange the basis vectors as the columns of a matrix X_1; similarly choose a basis of S_2 and arrange the basis vectors as the columns of a matrix X_2. Note that S_1 and S_2 can have different dimensions. Define the square of the distance between two subspaces by

$$dist^2(S_1, S_2) = dist^2(X_1, X_2) = ||X_1 - X_2 X_2^T X_1||_F^2$$

Since $X_1 - X_2 X_2^T X_1$ and $X_2 X_2^T X_1$ are orthogonal

$$\|X_1\|_F^2 = \left\|X_1 - X_2 X_2^T X_1\right\|_F^2 + \left\|X_2 X_2^T X_1\right\|_F^2$$

and hence

$$dist^2\left(X_1, X_2\right) = \|X_1\|_F^2 - \left\|X_2 X_2^T X_1\right\|_F^2 .$$

Intuitively, the distance between X_1 and X_2 is the Frobenius norm of the component of X_1 not in the space spanned by the columns of X_2.

If X_1 and X_2 are 1-dimensional unit length vectors, $dist^2\left(X_1, X_2\right)$ is the sin squared of the angle between the spaces.

Example: Consider two subspaces in four dimensions

$$X_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{1}{\sqrt{3}} \end{pmatrix} \qquad X_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$$

Here

$$dist^2\left(X_1, X_2\right) = \left\| \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{1}{\sqrt{3}} \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{1}{\sqrt{3}} \end{pmatrix} \right\|_F^2$$

$$= \left\| \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{1}{\sqrt{3}} \end{pmatrix} \right\|_F^2 = \frac{7}{6}$$

In essence, we projected each column vector of X_1 onto X_2 and computed the Frobenius norm of X_1 minus the projection. The Frobenius norm of each column is the sin squared of the angle between the original column of X_1 and the space spanned by the columns of X_2. ∎

12.6 Generating Functions

A sequence $a_0, a_1, \ldots$, can be represented by a generating function $g(x) = \sum_{i=0}^{\infty} a_i x^i$. The advantage of the generating function is that it captures the entire sequence in a closed form that can be manipulated as an entity. For example, if $g(x)$ is the generating function for the sequence $a_0, a_1, \ldots$, then $x\frac{d}{dx}g(x)$ is the generating function for the sequence $0,\ a_1,\ 2a_2,\ 3a_3, \ldots$ and $x^2 g''(x) + xg'(x)$ is the generating function for the sequence for $0,\ a_1,\ 4a_2,\ 9a_3, \ldots$

Example: The generating function for the sequence $1, 1, \ldots$ is $\sum_{i=0}^{\infty} x^i = \frac{1}{1-x}$. The generating function for the sequence $0, 1, 2, 3, \ldots$ is

$$\sum_{i=0}^{\infty} i x^i = \sum_{i=0}^{\infty} x\frac{d}{dx}x^i = x\frac{d}{dx}\sum_{i=0}^{\infty} x^i = x\frac{d}{dx}\frac{1}{1-x} = \frac{x}{(1-x)^2}.$$

∎

Example: If A can be selected 0 or 1 times and B can be selected 0, 1, or 2 times and C can be selected 0, 1, 2, or 3 times, in how many ways can five objects be selected. Consider the generating function for the number of ways to select objects. The generating function for the number of ways of selecting objects, selecting only A's is $1+x$, only B's is $1+x+x^2$, and only C's is $1 + x + x^2 + x^3$. The generating function when selecting A's, B's, and C's is the product.

$$(1 + x)(1 + x + x^2)(1 + x + x^2 + x^3) = 1 + 3x + 5x^2 + 6x^3 + 5x^4 + 3x^5 + x^6$$

The coefficient of x^5 is 3 and hence we can select five objects in three ways: ABBCC, ABCCC, or BBCCC. ∎

The generating functions for the sum of random variables

Let $f(x) = \sum_{i=0}^{\infty} p_i x^i$ be the generating function for an integer valued random variable where p_i is the probability that the random variable takes on value i. Let $g(x) = \sum_{i=0}^{\infty} q_i x^i$ be the generating function of an independent integer valued random variable where q_i is the probability that the random variable takes on the value i. The sum of these two random variables has the generating function $f(x)g(x)$. This is because the coefficient of x^i in the product $f(x)g(x)$ is $\sum_{k=0}^{i} p_k q_{k-i}$ and this is also the probability that the sum of the random variables is i. Repeating this, the generating function of a sum of independent nonnegative integer valued random variables is the product of their generating functions.

12.6.1 Generating Functions for Sequences Defined by Recurrence Relationships

Consider the Fibonacci sequence

$$0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, \ldots$$

defined by the recurrence relationship

$$f_0 = 0 \qquad f_1 = 1 \qquad f_i = f_{i-1} + f_{i-2} \quad i \geq 2$$

Multiply each side of the recurrence by x^i and sum from i equals two to infinity.

$$\begin{aligned} \sum_{i=2}^{\infty} f_i x^i &= \sum_{i=2}^{\infty} f_{i-1} x^i + \sum_{i=2}^{\infty} f_{i-2} x^i \\ f_2 x^2 + f_3 x^3 + \cdots &= f_1 x^2 + f_2 x^3 + \cdots + f_0 x^2 + f_1 x^3 + \cdots \\ &= x\left(f_1 x + f_2 x^2 + \cdots\right) + x^2\left(f_0 + f_1 x + \cdots\right) \end{aligned} \tag{12.1}$$

Let

$$f(x) = \sum_{i=0}^{\infty} f_i x^i. \tag{12.2}$$

Substituting (12.2) into (12.1) yields

$$\begin{aligned} f(x) - f_0 - f_1 x &= x\left(f(x) - f_0\right) + x^2 f(x) \\ f(x) - x &= x f(x) + x^2 f(x) \\ f(x)(1 - x - x^2) &= x \end{aligned}$$

Thus, $f(x) = \frac{x}{1-x-x^2}$ is the generating function for the Fibonacci sequence.

Note that generating functions are formal manipulations and do not necessarily converge outside some region of convergence. Consider the generating function $f(x) = \sum_{i=0}^{\infty} f_i x^i = \frac{x}{1-x-x^2}$ for the Fibonacci sequence. Using $\sum_{i=0}^{\infty} f_i x^i$,

$$f(1) = f_0 + f_1 + f_2 + \cdots = \infty$$

and using $f(x) = \frac{x}{1-x-x^2}$

$$f(1) = \frac{1}{1-1-1} = -1.$$

Asymptotic behavior

To determine the asymptotic behavior of the Fibonacci sequence write

$$f(x) = \frac{x}{1-x-x^2} = \frac{\frac{\sqrt{5}}{5}}{1-\phi_1 x} + \frac{-\frac{\sqrt{5}}{5}}{1-\phi_2 x}$$

where $\phi_1 = \frac{1+\sqrt{5}}{2}$ and $\phi_1 = \frac{1-\sqrt{5}}{2}$ are the reciprocals of the two roots of the quadratic $1-x-x^2 = 0$.

Then

$$f(x) = \frac{\sqrt{5}}{5}\Big(1 + \phi_1 x + (\phi_1 x)^2 + \cdots - \big(1 + \phi_2 x + (\phi_2 x)^2 + \cdots\big)\Big).$$

Thus,

$$f_n = \frac{\sqrt{5}}{5}\left(\phi_1^n - \phi_2^n\right).$$

Since $\phi_2 < 1$ and $\phi_1 > 1$, for large n, $f_n \cong \frac{\sqrt{5}}{5}\phi_1^n$. In fact, since $f_n = \frac{\sqrt{5}}{5}\left(\phi_1^n - \phi_2^n\right)$ is an integer and $\phi_2 < 1$, it must be the case that $f_n = \left\lfloor f_n + \frac{\sqrt{5}}{2}\phi_2^n \right\rfloor$. Hence $f_n = \left\lfloor \frac{\sqrt{5}}{5}\phi_1^n \right\rfloor$ for all n.

Means and standard deviations of sequences

Generating functions are useful for calculating the mean and standard deviation of a sequence. Let z be an integral valued random variable where p_i is the probability that z equals i. The expected value of z is given by $m = \sum_{i=0}^{\infty} i p_i$. Let $p(x) = \sum_{i=0}^{\infty} p_i x^i$ be the generating function for the sequence $p_1, p_2, \ldots$. The generating function for the sequence $p_1, 2p_2, 3p_3, \ldots$ is

$$x\frac{d}{dx}p(x) = \sum_{i=0}^{\infty} i p_i x^i.$$

Thus, the expected value of the random variable z is $m = xp'(x)|_{x=1} = p'(1)$. If p was not a probability function, its average value would be $\frac{p'(1)}{p(1)}$ since we would need to normalize the area under p to one.

The second moment of z, is $E(z^2) - E^2(z)$ and can be obtained as follows.

$$\begin{aligned} x^2 \frac{d}{dx} p(x)\Big|_{x=1} &= \sum_{i=0}^{\infty} i(i-1)x^i p(x)\Big|_{x=1} \\ &= \sum_{i=0}^{\infty} i^2 x^i p(x)\Big|_{x=1} - \sum_{i=0}^{\infty} i x^i p(x)\Big|_{x=1} \\ &= E(z^2) - E(z). \end{aligned}$$

Thus, $\sigma^2 = E(z^2) - E^2(z) = E(z^2) - E(z) + E(z) - E^2(z) = p^{"}(1) + p'(1) - \big(p'(1)\big)^2$.

12.6.2 The Exponential Generating Function and the Moment Generating Function

Besides the ordinary generating function there are a number of other types of generating functions. One of these is the exponential generating function. Given a sequence $a_0, a_1, \ldots$, the associated *exponential generating function* is $g(x) = \sum_{i=0}^{\infty} a_i \frac{x^i}{i!}$.

Moment generating functions

The k^{th} moment of a random variable x around the point b is given by $E((x-b)^k)$. Usually the word moment is used to denote the moment around the value 0 or around the mean. In the following, we use moment to mean the moment about the origin.

The *moment generating function* of a random variable x is defined by

$$\Psi(t) = E(e^{tx}) = \int_{-\infty}^{\infty} e^{tx} p(x) dx$$

Replacing e^{tx} by its power series expansion $1 + tx + \frac{(tx)^2}{2!} \cdots$ gives

$$\Psi(t) = \int_{-\infty}^{\infty} \left(1 + tx + \frac{(tx)^2}{2!} + \cdots\right) p(x) dx$$

Thus, the k^{th} moment of x about the origin is $k!$ times the coefficient of t^k in the power series expansion of the moment generating function. Hence, the moment generating function is the exponential generating function for the sequence of moments about the origin.

The moment generating function transforms the probability distribution $p(x)$ into a function $\Psi(t)$ of t. Note $\Psi(0) = 1$ and is the area or integral of $p(x)$. The moment generating function is closely related to the *characteristic function* which is obtained by replacing e^{tx} by e^{itx} in the above integral where $i = \sqrt{-1}$ and is related to the *Fourier*

transform which is obtained by replacing e^{tx} by e^{-itx}.

$\Psi(t)$ is closely related to the Fourier transform and its properties are essentially the same. In particular, $p(x)$ can be uniquely recovered by an inverse transform from $\Psi(t)$. More specifically, if all the moments m_i are finite and the sum $\sum_{i=0}^{\infty} \frac{m_i}{i!} t^i$ converges absolutely in a region around the origin, then $p(x)$ is uniquely determined.

The Gaussian probability distribution with zero mean and unit variance is given by $p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. Its moments are given by

$$\begin{aligned} u_n &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^n e^{-\frac{x^2}{2}} dx \\ &= \begin{cases} \frac{n!}{2^{\frac{n}{2}} \left(\frac{n}{2}\right)!} & \text{n even} \\ 0 & \text{n odd} \end{cases} \end{aligned}$$

To derive the above, use integration by parts to get $u_n = (n-1) u_{n-2}$ and combine this with $u_0 = 1$ and $u_1 = 0$. The steps are as follows. Let $u = e^{-\frac{x^2}{2}}$ and $v = x^{n-1}$. Then $u' = -xe^{-\frac{x^2}{2}}$ and $v' = (n-1) x^{n-2}$. Now $uv = \int u'v + \int uv'$ or

$$e^{-\frac{x^2}{2}} x^{n-1} = \int x^n e^{-\frac{x^2}{2}} dx + \int (n-1) x^{n-2} e^{-\frac{x^2}{2}} dx.$$

From which

$$\begin{aligned} \int x^n e^{-\frac{x^2}{2}} dx &= (n-1) \int x^{n-2} e^{-\frac{x^2}{2}} dx - e^{-\frac{x^2}{2}} x^{n-1} \\ \int_{-\infty}^{\infty} x^n e^{-\frac{x^2}{2}} dx &= (n-1) \int_{-\infty}^{\infty} x^{n-2} e^{-\frac{x^2}{2}} dx \end{aligned}$$

Thus, $u_n = (n-1) u_{n-2}$.

The moment generating function is given by

$$g(s) = \sum_{n=0}^{\infty} \frac{u_n s^n}{n!} = \sum_{\substack{n=0 \\ n \text{ even}}}^{\infty} \frac{n!}{2^{\frac{n}{2}} \frac{n}{2}!} \frac{s^n}{n!} = \sum_{i=0}^{\infty} \frac{s^{2i}}{2^i i!} = \sum_{i=0}^{\infty} \frac{1}{i!} \left(\frac{s^2}{2}\right)^i = e^{\frac{s^2}{2}}.$$

For the general Gaussian, the moment generating function is

$$g(s) = e^{su + \left(\frac{\sigma^2}{2}\right) s^2}$$

Thus, given two independent Gaussians with mean u_1 and u_2 and variances σ_1^2 and σ_2^2, the product of their moment generating functions is

$$e^{s(u_1+u_2) + \left(\sigma_1^2 + \sigma_2^2\right) s^2},$$

the moment generating function for a Gaussian with mean $u_1 + u_2$ and variance $\sigma_1^2 + \sigma_2^2$. Thus, the convolution of two Gaussians is a Gaussian and the sum of two random variables that are both Gaussian is a Gaussian random variable.

12.7 Miscellaneous

12.7.1 Lagrange multipliers

Lagrange multipliers are used to convert a constrained optimization problem into an unconstrained optimization. Suppose we wished to maximize a function $f(\mathbf{x})$ subject to a constraint $g(\mathbf{x}) = c$. The value of $f(\mathbf{x})$ along the constraint $g(\mathbf{x}) = c$ might increase for a while and then start to decrease. At the point where $f(\mathbf{x})$ stops increasing and starts to decrease, the contour line for $f(\mathbf{x})$ is tangent to the curve of the constraint $g(\mathbf{x}) = c$. Stated another way the gradient of $f(\mathbf{x})$ and the gradient of $g(\mathbf{x})$ are parallel.

By introducing a new variable λ we can express the condition by $\nabla_{\mathbf{x}} f = \lambda \nabla_{\mathbf{x}} g$ and $g = c$. These two conditions hold if and only if

$$\nabla_{\mathbf{x}\lambda} \left(f(\mathbf{x}) + \lambda \left(g(\mathbf{x}) - c \right) \right) = 0$$

The partial with respect to λ establishes that $g(\mathbf{x}) = c$. We have converted the constrained optimization problem in x to an unconstrained problem with variables $\mathbf{x}$ and λ.

12.7.2 Finite Fields

For a prime p and integer n there is a unique finite field with p^n elements. In Section **??** we used the field GF(2^n), which consists of polynomials of degree less than n with coefficients over the field GF(2). In GF(2^8)

$$(x^7 + x^5 + x) + (x^6 + x^5 + x^4) = x^7 + x^6 = x^4 = x$$

Multiplication is modulo an irreducible polynomial. Thus

$$\begin{aligned}(x^7 + x^5 + x)(x^6 + x^5 + x^4) &= x^{13} + x^{12} + x^{11} + x^{11} + x^{10} + x^9 + x^7 + x^6 + x^5 \\ &= x^{13} + x^{12} + x^{10} + x^9 + x^7 + x^6 + x^5 \\ &= x^6 + x^4 + x^3 + x^2 \qquad \text{mod } x^8 + x^4 + x^3 + x + 1\end{aligned}$$

Division of $x^{13} + x^{12} + x^{10} + x^9 + x^7 + x^6 + x^5$ by $x^6 + x^4 + x^3 + x^2$ is illustrated below.

$$\begin{array}{lccccccccccccc}
 & x^{13} & +x^{12} & +x^{10} & +x^9 & & +x^7 & +x^6 & +x^5 & & & & \\
-x^5(x^8 + x^4 + x^3 + x^2 + 1) = & x^{13} & & & +x^9 & +x^8 & & +x^6 & +x^5 & & & & \\
\hline
 & & x^{12} & +x^{10} & & +x^8 & +x^7 & & & & & & \\
-x^4(x^8 + x^4 + x^3 + x^2 + 1) = & & x^{12} & & & +x^8 & +x^7 & & +x^5 & +x^4 & & & \\
\hline
 & & & x^{10} & & & & & +x^5 & x^4 & & & \\
-x^2(x^8 + x^4 + x^3 + x^2 + 1) = & & & x^{10} & & & & x^6 & +x^5 & & x^3 & x^2 & \\
\hline
 & & & & & & & x^6 & & +x^4 & +x^3 & +x^2 &
\end{array}$$

12.7.3 Hash Functions

Universal Hash Families

ADD PARAGRAPH ON MOTIVATION integrate material with Chapter

Let $M = \{1, 2, \ldots, m\}$ and $N = \{1, 2, \ldots, n\}$ where $m \geq n$. A family of hash functions $H = \{h | h : M \to N\}$ is said to be 2-universal if for all x and y, $x \neq y$, and for h chosen uniformly at random from H,

$$Prob\left[h\left(x\right) = h\left(y\right)\right] \leq \frac{1}{n}$$

Note that if H is the set of all possible mappings from M to N, then H is 2-universal. In fact $Prob\left[h\left(x\right) = h\left(y\right)\right] = \frac{1}{n}$. The difficulty in letting H consist of all possible functions is that a random h from H has no short representation. What we want is a small set H where each $h \in H$ has a short representation and is easy to compute.

Note that for a 2-universal H, for any two elements x and y, $h(x)$ and $h(y)$ behave as independent random variables. For a random f and any set X the set $\{f\left(x\right) | x \in X\}$ is a set of independent random variables.

12.7.4 Application of Mean Value Theorem

The mean value theorem states that if $f(x)$ is continuous and differentiable on the interval $[a, b]$, then there exists c, $a \leq c \leq b$ such that $f'(c) = \frac{f(b)-f(a)}{b-a}$. That is, at some point between a and b the derivative of f equals the slope of the line from $f(a)$ to $f(b)$. See Figure 12.7.4.

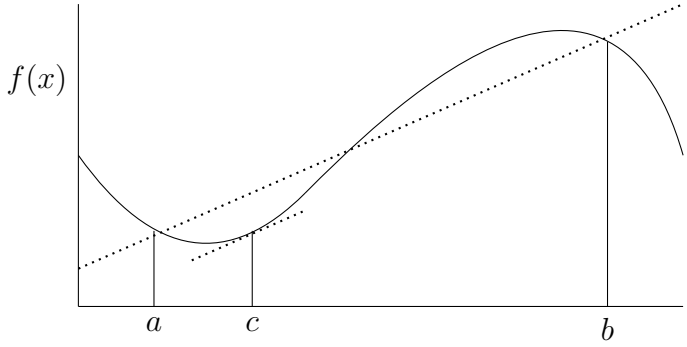

Figure 12.3: Illustration of the mean value theorem.

One application of the mean value theorem is with the Taylor expansion of a function. The Taylor expansion about the origin of $f(x)$ is

$$f(x) = f(0) + f'(0)x + \frac{1}{2!}f''(0)x^2 + \frac{1}{3!}f'''(0)x^3 + \cdots \tag{12.3}$$

By the mean value theorem there exists c, $0 \leq c \leq x$, such that $f'(c) = \frac{f(x)-f(0)}{x}$ or $f(x) - f(0) = xf'(c)$. Thus

$$xf'(c) = f'(0)x + \frac{1}{2!}f''(0)x^2 + \frac{1}{3!}f'''(0)x^3 + \cdots$$

and

$$f(x) = f(0) + xf'(c).$$

One could apply the mean value theorem to $f'(x)$ in

$$f'(x) = f'(0) + f''(0)x + \frac{1}{2!}f'''(0)x^2 + \cdots$$

Then there exists d, $0 \leq d \leq x$ such that

$$xf''(d) = f''(0)x + \frac{1}{2!}f'''(0)x^2 + \cdots$$

Integrating

$$\frac{1}{2}x^2f''(d) = \frac{1}{2!}f''(0)x + \frac{1}{3!}f'''(0)x^3 + \cdots$$

Substituting into Eq(12.3)

$$f(x) = f(0) + f'(0)x + \frac{1}{2}x^2f''(d).$$

12.7.5 Sperner's Lemma

Consider a triangulation of a 2-dimensional simplex. Let the vertices of the simplex be colored R, B, and G. If the vertices on each edge of the simplex are colored only with the two colors at the endpoints then the triangulation must have a triangle whose vertices are three different colors. In fact, it must have an odd number of such vertices. A generalization of the lemma to higher dimensions also holds.

Create a graph whose vertices correspond to the triangles of the triangulation plus an additional vertex corresponding to the outside region. Connect two vertices of the graph by an edge if the triangles corresponding to the two vertices share a common edge that is color R and B. The edge of the original simplex must have an odd number of such triangular edges. Thus, the outside vertex of the graph must be of odd degree. The graph must have an even number of odd degree vertices. Each odd vertex is of degree 0, 1, or 2. The vertices of odd degree, i.e. degree one, correspond to triangles which have all three colors.

12.7.6 Prüfer

Here we prove that the number of labeled trees with n vertices is n^{n-2}. By a labeled tree we mean a tree with n vertices and n distinct labels, each label assigned to one vertex.

Theorem 12.25 *The number of labeled trees with n vertices is* n^{n-2}.

Proof: (Prüfer sequence) There is a one-to-one correspondence between labeled trees and sequences of length $n-2$ of integers between 1 and n. An integer may repeat in the sequence. The number of such sequences is clearly n^{n-2}. Although each vertex of the tree has a unique integer label the corresponding sequence has repeating labels. The reason for this is that the labels in the sequence refer to interior vertices of the tree and the number of times the integer corresponding to an interior vertex occurs in the sequence is related to the degree of the vertex. Integers corresponding to leaves do not appear in the sequence.

To see the one-to-one correspondence, first convert a tree to a sequence by deleting the lowest numbered leaf. If the lowest numbered leaf is i and its parent is j, append j to the tail of the sequence. Repeating the process until only two vertices remain yields the sequence. Clearly a labeled tree gives rise to only one sequence.

It remains to show how to construct a unique tree from a sequence. The proof is by induction on n. For $n = 1$ or 2 the induction hypothesis is trivially true. Assume the induction hypothesis true for $n-1$. Certain numbers from 1 to n do not appear in the sequence and these numbers correspond to vertices that are leaves. Let i be the lowest number not appearing in the sequence and let j be the first integer in the sequence. Then i corresponds to a leaf connected to vertex j. Delete the integer j from the sequence. By the induction hypothesis there is a unique labeled tree with integer labels $1, \ldots, i-1, i+1, \ldots, n$. Add the leaf i by connecting the leaf to vertex j. We need to argue that no other sequence can give rise to the same tree. Suppose some other sequence did. Then the i^{th} integer in the sequence must be j. By the induction hypothesis the sequence with j removed is unique.

Algorithm

Create leaf list - the list of labels not appearing in the Prüfer sequence. n is the length of the Prüfer list plus two.
while Prüfer sequence is non empty do
begin
p =first integer in Prüfer sequence
e =smallest label in leaf list
Add edge (p, e)
Delete e from leaf list
Delete p from Prüfer sequence
If p no longer appears in Prüfer sequence add p to leaf list
end
There are two vertices e and f on leaf list, add edge (e, f)

12.8 Exercises

Exercise 12.1 *What is the difference between saying $f(n)$ is $O(n^3)$ and $f(n)$ is $o(n^3)$?*

Exercise 12.2 *If $f(n) \sim g(n)$ what can we say about $f(n) + g(n)$ and $f(n) - g(n)$?*

Exercise 12.3 *What is the difference between $\sim$ and Θ?*

Exercise 12.4 *If $f(n)$ is $O(g(n))$ does this imply that $g(n)$ is $\Omega(f(n))$?*

Exercise 12.5 *What is $\lim_{k\to\infty} \left(\frac{k-1}{k-2}\right)^{k-2}$.*

Exercise 12.6 *Select a, b, and c uniformly at random from $[0, 1]$. The probability that $b < a$ is $1/2$. The probability that $c<a$ is $1/2$. However, the probability that both b and c are less than a is $\frac{1}{3}$ not $1/4$. Why is this? Note that the six possible permutations abc, acb, bac, cab, bca, and cba, are all equally likely. Assume that a, b, and c are drawn from the interval (0,1]. Given that $b < a$, what is the probability that $c < a$?*

Exercise 12.7 *Let $A_1, A_2, \ldots, A_n$ be events. Prove that $Prob(A_1 \cup A_2 \cup \cdots A_n) \leq \sum_{i=1}^{n} Prob(A_i)$*

Exercise 12.8 *Give an example of three random variables that are pairwise independent but not fully independent.*

Exercise 12.9 *Give examples of nonnegative valued random variables with median >> mean. Can we have median << mean?*

Exercise 12.10 *Consider n samples $x_1, x_2, \ldots, x_n$ from a Gaussian distribution of mean μ and variance σ. For this distribution $m = \frac{x_1+x_2+\cdots+x_n}{n}$ is an unbiased estimator of μ. If μ is known then $\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$ is an unbiased estimator of σ^2. Prove that if we approximate μ by m, then $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2$ is an unbiased estimator of σ^2.*

Exercise 12.11 *Given the distribution $\frac{1}{\sqrt{2\pi}3} e^{-\frac{1}{2}\left(\frac{x}{3}\right)^2}$ what is the probability that $x > 1$?*

Exercise 12.12 *$e^{-\frac{x^2}{2}}$ has value 1 at $x = 0$ and drops off very fast as x increases. Suppose we wished to approximate $e^{-\frac{x^2}{2}}$ by a function $f(x)$ where*

$$f(x) = \begin{cases} 1 & |x| \leq a \\ 0 & |x| > a \end{cases}.$$

What value of a should we use? What is the integral of the error between $f(x)$ and $e^{-\frac{x^2}{2}}$?

Exercise 12.13 *Given two sets of red and black balls with the number of red and black balls in each set shown in the table below.*

	red	*black*
Set 1	*40*	*60*
Set 2	*50*	*50*

Randomly draw a ball from one of the sets. Suppose that it turns out to be red. What is the probability that it was drawn from Set 1?

Exercise 12.14 *Why cannot one prove an analogous type of theorem that states* $p(x \leq a) \leq \frac{E(x)}{a}$ *?*

Exercise 12.15 *Compare the Markov and Chebyshev bounds for the following probability distributions*

1. $p(x) = \begin{cases} 1 & x = 1 \\ 0 & otherwise \end{cases}$
2. $p(x) = \begin{cases} 1/2 & 0 \leq x \leq 2 \\ 0 & otherwise \end{cases}$

Exercise 12.16 *Let* s *be the sum of* n *independent random variables* $x_1, x_2, \ldots, x_n$ *where for each* i

$$x_i = \begin{cases} 0 & Prob \quad p \\ 1 & Prob \quad 1-p \end{cases}$$

1. *How large must* δ *be if we wish to have* $Prob\left(s < (1-\delta)\, m\right) < \varepsilon$*?*

2. *If we wish to have* $Prob\left(s > (1+\delta)\, m\right) < \varepsilon$*?*

Exercise 12.17 *What is the expected number of flips of a coin until a head is reached? Assume p is probability of a head on an individual flip. What is value if p=1/2?*

Exercise 12.18 *Given the joint probability*

P(A,B)	A=0	A=1
B=0	1/16	1/8
B=1	1/4	9/16

1. *What is the marginal probability of* A*? of* B*?*

2. *What is the conditional probability of* B *given* A*?*

Exercise 12.19 *Consider independent random variables x_1, x_2, and x_3, each equal to zero with probability $\frac{1}{2}$. Let $S = x_1 + x_2 + x_3$ and let F be event that $S \in \{1, 2\}$. Conditioning on F, the variables x_1, x_2, and x_3 are still each zero with probability $\frac{1}{2}$. Are they still independent?*

Exercise 12.20 *Consider rolling two dice A and B. What is the probability that the sum S will add to nine? What is the probability that the sum will be 9 if the roll of A is 3?*

Exercise 12.21 *Write the generating function for the number of ways of producing chains using only pennies, nickels, and dines. In how many ways can you produce 23 cents?*

Exercise 12.22 *A dice has six faces, each face of the dice having one of the numbers 1 though 6. The result of a role of the dice is the integer on the top face. Consider two roles of the dice. In how many ways can an integer be the sum of two roles of the dice.*

Exercise 12.23 *If a(x) is the generating function for the sequence $a_0, a_1, a_2, \ldots$, for what sequence is a(x)(1-x) the generating function.*

Exercise 12.24 *How many ways can one draw n $a's$ and $b's$ with an even number of $a's$.*

Exercise 12.25 *Find the generating function for the recurrence $a_i = 2a_{i-1} + i$ where $a_0 = 1$.*

Exercise 12.26 *Find a closed form for the generating function for the infinite sequence of prefect squares 1, 4, 9, 16, 25, ...*

Exercise 12.27 *Given that $\frac{1}{1-x}$ is the generating function for the sequence $1, 1, \ldots$, for what sequence is $\frac{1}{1-2x}$ the generating function?*

Exercise 12.28 *Find a closed form for the exponential generating function for the infinite sequence of prefect squares 1, 4, 9, 16, 25, ...*

Exercise 12.29 *Prove that the L_2 norm of $(a_1, a_2, \ldots, a_n)$ is less than or equal to the L_1 norm of $(a_1, a_2, \ldots, a_n)$.*

Exercise 12.30 *Prove that there exists a y, $0 \leq y \leq x$, such that $f(x) = f(0) + f'(y)x$.*

Exercise 12.31 *Show that the eigenvectors of a matrix A are not a continuous function of changes to the matrix.*

Exercise 12.32 *What are the eigenvalues of the two graphs shown below? What does this say about using eigenvalues to determine if two graphs are isomorphic.*

Exercise 12.33 *Let A be the adjacency matrix of an undirected graph G. Prove that eigenvalue λ_1 of A is at least the average degree of G.*

Exercise 12.34 *Show that if A is a symmetric matrix and λ_1 and λ_2 are distinct eigenvalues then their corresponding eigenvectors x_1 and x_2 are orthogonal.*
Hint:

Exercise 12.35 *Show that a matrix is rank k if and only if it has k nonzero eigenvalues and eigenvalue 0 of rank n-k.*

Exercise 12.36 *Prove that maximizing $\frac{x^T Ax}{x^T x}$ is equivalent to maximizing $x^T Ax$ subject to the condition that x be of unit length.*

Exercise 12.37 *Let A be a symmetric matrix with smallest eigenvalue $\lambda_{\min}$. Give a bound on the largest element of A^{-1}.*

Exercise 12.38 *Let A be the adjacency matrix of an n vertex clique with no self loops. Thus, each row of A is all ones except for the diagonal entry which is zero. What is the spectrum of A.*

Exercise 12.39 *Let A be the adjacency matrix of an undirect graph G. Prove that the eigenvalue λ_1 of A is at least the average degree of G.*

Exercise 12.40 *We are given the probability distribution for two random vectors x and y and we wish to stretch space to maximize the expected distance between them. Thus, we will multiply each coordinate by some quantity a_i. We restrict $\sum_{i=1}^{d} a_i^2 = d$. Thus, if we increase some coordinate by $a_i > 1$, some other coordinate must shrink. Given random vectors $x = (x_1, x_2, \ldots, x_d)$ and $y = (y_1, y_2, \ldots, y_d)$ how should we select a_i to maximize $E\left(|x-y|^2\right)$? The a_i stretch different coordinates. Assume*

$$y_i = \begin{cases} 0 & \frac{1}{2} \\ 1 & \frac{1}{2} \end{cases}$$

and that x_i has some arbitrary distribution.

$$E\left(|x-y|^2\right) = E\sum_{i=1}^{d}\left[a_i^2\left(x_i - y_i\right)^2\right] = \sum_{i=1}^{d} a_i^2 E\left(x_i^2 - 2x_i y_i + y_i^2\right)$$
$$= \sum_{i=1}^{d} a_i^2 E\left(x_i^2 - x_i + \tfrac{1}{2}\right)$$

Since $E\left(x_i^2\right) = E\left(x_i\right)$ we get . Thus, weighting the coordinates has no effect assuming $\sum_{i=1}^{d} a_i^2 = 1$. Why is this? Since $E\left(y_i\right) = \frac{1}{2}$.

$E\left(|x-y|^2\right)$ is independent of the value of x_i hence its distribution.

What if $y_i = \begin{cases} 0 & \frac{3}{4} \\ 1 & \frac{1}{4} \end{cases}$ *and* $E(y_i) = \frac{1}{4}$. *Then*

$$E\left(|x-y|^2\right) = \sum_{i=1}^{d} a_i^2 E\left(x_i^2 - 2x_i y_i + y_i^2\right) = \sum_{i=1}^{d} a_i^2 E\left(x_i - \tfrac{1}{2}x_i + \tfrac{1}{4}\right)$$
$$= \sum_{i=1}^{d} a_i^2 \left(\tfrac{1}{2}E(x_i) + \tfrac{1}{4}\right).$$

To maximize put all weight on the coordinate of x with highest probability of one. What if we used 1-norm instead of the two norm?

$$E(|x-y|) = E\sum_{i=1}^{d} a_i |x_i - y_i| = \sum_{i=1}^{d} a_i E|x_i - y_i| = \sum_{i=1}^{d} a_i b_i$$

where $b_i = E(x_i - y_i)$. *If* $\sum_{i=1}^{d} a_i^2 = 1$, *then to maximize let* $a_i = \frac{b_i}{b}$. *Taking the dot product of a and b is maximized when both are in the same direction.*

Exercise 12.41 *Maximize x+y subject to the constraint that* $x^2 + y^2 = 1$.

Exercise 12.42 *Draw a tree with 10 vertices and label each vertex with a unique integer from 1 to 10. Construct the Prfer sequence for the tree. Given the Prfer sequence recreate the tree.*

Exercise 12.43 *Construct the tree corresponding to the following Prfer sequences*

1. *113663*
2. *552833226*

References

[ABC+08] Reid Andersen, Christian Borgs, Jennifer T. Chayes, John E. Hopcroft, Vahab S. Mirrokni, and Shang-Hua Teng. Local computation of pagerank contributions. *Internet Mathematics*, 5(1):23–45, 2008.

[AF] David Aldous and James Fill. *Reversible Markov Chains and Random Walks on Graphs*. http://www.stat.berkeley.edu/ aldous/RWG/book.html.

[AK] Sanjeev Arora and Ravindran Kannan. Learning mixtures of separated nonspherical gaussians. *Annals of Applied Probability*, 15(1A):6992.

[Alo86] Noga Alon. Eigenvalues and expanders. *Combinatorica*, 6:83–96, 1986.

[AM05] Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *COLT*, pages 458–469, 2005.

[AN72] Krishna Athreya and P. E. Ney. *Branching Processes*, volume 107. Springer, Berlin, 1972.

[AP03] Dimitris Achlioptas and Yuval Peres. The threshold for random k-sat is 2^{k} (ln 2 - o(k)). In *STOC*, pages 223–231, 2003.

[Aro11] Multiplicative weights method: a meta-algorithm and its applications. *Theory of Computing journal - to appear*, 2011.

[AS08] Noga Alon and Joel H. Spencer. *The probabilistic method*. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons Inc., Hoboken, NJ, third edition, 2008. With an appendix on the life and work of Paul Erdős.

[BA] Albert-Lszl Barabsi and Rka Albert. Emergence of scaling in random networks. *Science*, 286(5439).

[BEHW] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the Association for Computing Machinary*.

[BGG97] C Sidney Burrus, Ramesh A Gopinath, and Haitao Guo. Introduction to wavelets and wavelet transforms: a primer. 1997.

[Ble12] David M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, 2012.

[Blo62] H.D. Block. The perceptron: A model for brain functioning. *Reviews of Modern Physics*, 34:123–135, 1962. Reprinted in *Neurocomputing*, Anderson and Rosenfeld.

[BMPW98] Sergey Brin, Rajeev Motwani, Lawrence Page, and Terry Winograd. What can you do with a web in your pocket? *Data Engineering Bulletin*, 21:37–47, 1998.

[Bol01] Béla Bollobás. *Random Graphs*. Cambridge University Press, 2001.

[BT87] Béla Bollobás and Andrew Thomason. Threshold functions. *Combinatorica*, 7(1):35–38, 1987.

[CF86] Ming-Te Chao and John V. Franco. Probabilistic analysis of two heuristics for the 3-satisfiability problem. *SIAM J. Comput.*, 15(4):1106–1118, 1986.

[CGTS99] Moses Charikar, Sudipto Guha, Éva Tardos, and David B. Shmoys. A constant-factor approximation algorithm for the k-median problem (extended abstract). In *Proceedings of the thirty-first annual ACM symposium on Theory of computing*, STOC '99, pages 1–10, New York, NY, USA, 1999. ACM.

[CHK+] Duncan S. Callaway, John E. Hopcroft, Jon M. Kleinberg, M. E. J. Newman, and Steven H. Strogatz. Are randomly grown graphs really random?

[Chv92] *33rd Annual Symposium on Foundations of Computer Science, 24-27 October 1992, Pittsburgh, Pennsylvania, USA*. IEEE, 1992.

[CLMW11] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3):11, 2011.

[DFK91] Martin Dyer, Alan Frieze, and Ravindran Kannan. A random polynomial time algorithm for approximating the volume of convex bodies. *Journal of the Association for Computing Machinary*, 1991.

[DFK+99] Petros Drineas, Alan M. Frieze, Ravi Kannan, Santosh Vempala, and V. Vinay. Clustering in large graphs and matrices. In *SODA*, pages 291–299, 1999.

[DG99] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of the johnson-lindenstrauss lemma. 99(006), 1999.

[DS84] Peter G. Doyle and J. Laurie Snell. *Random walks and electric networks*, volume 22 of *Carus Mathematical Monographs*. Mathematical Association of America, Washington, DC, 1984.

[DS07] Sanjoy Dasgupta and Leonard J. Schulman. A probabilistic analysis of em for mixtures of separated, spherical gaussians. *Journal of Machine Learning Research*, 8:203–226, 2007.

[ER60] Paul Erdös and Alfred Rényi. On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61, 1960.

[Fel68] William Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, January 1968.

[FK99] Alan M. Frieze and Ravindan Kannan. Quick approximation to matrices and applications. *Combinatorica*, 19(2):175–220, 1999.

[Fri99] Friedgut. Sharp thresholds of graph properties and the k-sat problem. *Journal of the American Math. Soc.*, 12, no 4:1017–1054, 1999.

[FS96] Alan M. Frieze and Stephen Suen. Analysis of two simple heuristics on a random instance of k-sat. *J. Algorithms*, 20(2):312–355, 1996.

[GKP94] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete mathematics - a foundation for computer science (2. ed.)*. Addison-Wesley, 1994.

[GvL96] Gene H. Golub and Charles F. van Loan. *Matrix computations (3. ed.)*. Johns Hopkins University Press, 1996.

[HBB10] Matthew D. Hoffman, David M. Blei, and Francis R. Bach. Online learning for latent dirichlet allocation. In *NIPS*, pages 856–864, 2010.

[Jer98] Mark Jerrum. Mathematical foundations of the markov chain monte carlo method. In Dorit Hochbaum, editor, *Approximation Algorithms for NP-hard Problems*, 1998.

[JKLP93] Svante Janson, Donald E. Knuth, Tomasz Luczak, and Boris Pittel. The birth of the giant component. *Random Struct. Algorithms*, 4(3):233–359, 1993.

[JLR00] Svante Janson, Tomasz Ĺuczak, and Andrzej Ruciński. *Random Graphs*. John Wiley and Sons, Inc, 2000.

[Kan09] Ravindran Kannan. A new probability inequality using typical moments and concentration results. In *FOCS*, pages 211–220, 2009.

[Kar90] Richard M. Karp. The transitive closure of a random digraph. *Random Structures and Algorithms*, 1(1):73–94, 1990.

[Kle99] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *JOURNAL OF THE ACM*, 46(5):604–632, 1999.

[Kle00] Jon M. Kleinberg. The small-world phenomenon: an algorithm perspective. In *STOC*, pages 163–170, 2000.

[Kle02] Jon M. Kleinberg. An impossibility theorem for clustering. In *NIPS*, pages 446–453, 2002.

[KV95] Michael Kearns and Umesh Vazirani. *An introduction to Computational Learning Theory*. MIT Press, 1995.

[KV09] Ravi Kannan and Santosh Vempala. Spectral algorithms. *Foundations and Trends in Theoretical Computer Science*, 4(3-4):157–288, 2009.

[Liu01] Jun Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, 2001.

[Mat10] Jiří Matoušek. *Geometric discrepancy*, volume 18 of *Algorithms and Combinatorics*. Springer-Verlag, Berlin, 2010. An illustrated guide, Revised paperback reprint of the 1999 original.

[Mit97] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

[MP69] M. Minsky and S. Papert. *Perceptrons: An Introduction to Computational Geometry*. The MIT Press, 1969.

[MR95a] Michael Molloy and Bruce A. Reed. A critical point for random graphs with a given degree sequence. *Random Struct. Algorithms*, 6(2/3):161–180, 1995.

[MR95b] Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[MR99] Rajeev Motwani and Prabhakar Raghavan. Randomized algorithms. In *Algorithms and theory of computation handbook*, pages 15–1–15–23. CRC, Boca Raton, FL, 1999.

[MU05] Michael Mitzenmacher and Eli Upfal. *Probability and computing - randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.

[MV10] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *FOCS*, pages 93–102, 2010.

[Nov62] A.B.J. Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata, Vol. XII*, pages 615–622, 1962.

[Pal85] Edgar M. Palmer. *Graphical evolution*. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons Ltd., Chichester, 1985. An introduction to the theory of random graphs, A Wiley-Interscience Publication.

[Par98] Beresford N. Parlett. *The symmetric eigenvalue problem*, volume 20 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998. Corrected reprint of the 1980 original.

[per10] *Markov Chains and Mixing Times*. American Mathematical Society, 2010.

[Sch90] Rob Schapire. Strength of weak learnability. *Machine Learning*, 5:197–227, 1990.

[SJ] Alistair Sinclair and Mark Jerrum. Approximate counting, uniform generation and rapidly mixing markov chains. *Information and Computation*.

[Sly10] Allan Sly. Computational transition at the uniqueness threshold. In *FOCS*, pages 287–296, 2010.

[SS01] Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

[SWY75] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18:613–620, November 1975.

[Val84] Leslie G. Valiant. A theory of the learnable. In *STOC*, pages 436–445, 1984.

[Val13] L. Valiant. *Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World*. Basic Books, 2013.

[VC71] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

[Vem04] Santosh Vempala. *The Random Projection Method*. DIMACS, 2004.

[VW02] Santosh Vempala and Grant Wang. A spectral algorithm for learning mixtures of distributions. *Journal of Computer and System Sciences*, pages 113–123, 2002.

[Wil06] H.S. Wilf. *Generatingfunctionology*. Ak Peters Series. A K Peters, 2006.

[WS98a] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393 (6684), 1998.

[WS98b] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393, 1998.

[WW96] E. T. Whittaker and G. N. Watson. *A course of modern analysis*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 1996. An introduction to the general theory of infinite processes and of analytic functions; with an account of the principal transcendental functions, Reprint of the fourth (1927) edition.