Note: This document was originally compiled by Jessica Su, with minor modifications by Jayadev Bhaskaran, Vasco Portilheiro, and Manan Shah.

1 Proof techniques

Here we will learn to prove universal mathematical statements, like "the square of any odd number is odd". It's easy enough to show that this is true in specific cases – for example, $3^2 = 9$, which is an odd number, and $5^2 = 25$, which is another odd number. However, to prove the statement, we must show that it works for *all* odd numbers, which is hard because you can't try every single one of them.

Note that if we want to *disprove* a universal statement, we only need to find one counterexample. For instance, if we want to disprove the statement "the square of any odd number is even", it suffices to provide a specific example of an odd number whose square is not even. (For instance, $3^2 = 9$, which is not an even number.)

Rule of thumb:

- To **prove** a universal statement, you must show it works in all cases.
- To **disprove** a universal statement, it suffices to find one counterexample.

(For "existence" statements, this is reversed. For example, if your statement is "there exists at least one odd number whose square is odd, then proving the statement just requires saying $3^2 = 9$, while disproving the statement would require showing that none of the odd numbers have squares that are odd.)

1.0.1 Proving something is true for all members of a group

If we want to prove something is true for all odd numbers (for example, that the square of any odd number is odd), we can pick an arbitrary odd number x, and try to prove the statement for that number. In the proof, we cannot assume anything about x other than that it's an odd number. (So we can't just set x to be a specific number, like 3, because then our proof might rely on special properties of the number 3 that don't generalize to all odd numbers).

Example: Prove that the square of any odd number is odd.

Proof: Let x be an arbitrary odd number. By definition, an odd number is an integer that can be written in the form $2k + 1$, for some integer k. This means we can write $x = 2k + 1$, where k is some integer. So $x^2 = (2k + 1)^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1$. Since k is an integer, $2k^2 + 2k$ is also an integer, so we can write $x^2 = 2\ell + 1$, where $\ell = 2k^2 + 2k$ is an integer. Therefore, x^2 is odd.

Since this logic works for *any* odd number x, we have shown that the square of any odd number is odd.

1.1 Special techniques

In addition to the "pick an arbitrary element" trick, here are several other techniques commonly seen in proofs.

1.1.1 Proof by contrapositive

Consider the statement "If it is raining today, then I do not go to class."

This is logically equivalent to the statement "If I go to class, then it is not raining today."

So if we want to prove the first statement, it suffices to prove the second statement (which is called the **contrapositive**).

Note that it is **not** equivalent to the statement "If I do not go to class, then it is raining today" (this is called the fallacy of the converse).

Example: Let x be an integer. Prove that x^2 is an odd number if and only if x is an odd number.

Proof: The "if and only if" in this statement requires us to prove both directions of the implication. First, we must prove that if x is an odd number, then x^2 is an odd number. Then we should prove that if x^2 is an odd number, then x is an odd number.

We have already proven the first statement, so now we just need to prove the second statement. The second statement is logically equivalent to its contrapositive, so it suffices to prove that "if x is an even number, then x^2 is even."

Suppose x is an even number. This means we can write $x = 2k$ for some integer k. This means $x^2 = 4k^2 = 2(2k^2)$. Since k is an integer, $2k^2$ is also an integer, so we can write $x^2 = 2\ell$ for the integer $\ell = 2k^2$. By definition, this means x^2 is an even number.

1.1.2 Proof by contradiction

In proof by contradiction, you assume your statement is not true, and then derive a contradiction. This is really a special case of proof by contrapositive (where your "if" is all of mathematics, and your "then" is the statement you are trying to prove).

Example: Prove that $\sqrt{2}$ is irrational.

Proof: Suppose that $\sqrt{2}$ was rational. By definition, this means that $\sqrt{2}$ can be written as m/n for some integers m and n. Since $\sqrt{2} = m/n$, it follows that $2 = m^2/n^2$, so $m^2 = 2n^2$. Now any square number x^2 must have an even number of prime factors, since any prime factor found in the first x must also appear in the second x. Therefore, m^2 must have an even number of prime factors. However, since n^2 must also have an even number of prime factors, and 2 is a prime number, $2n^2$ must have an odd number of prime factors. This is a contradiction, since we claimed that $m^2 = 2n^2$, and no number can have both an even number of prime factors and an odd number of prime factors. Therefore, our initial assumption was wrong, and $\sqrt{2}$ must be irrational.

1.1.3 Proof by cases

Sometimes it's hard to prove the whole theorem at once, so you split the proof into several cases, and prove the theorem separately for each case.

Example: Let n be an integer. Show that if n is not divisible by 3, then $n^2 = 3k + 1$ for some integer k.

Proof: If n is not divisible by 3, then either $n = 3m+1$ (for some integer m) or $n = 3m+2$ (for some integer m.

Case 1: Suppose $n = 3m + 1$. Then $n^2 = (3m + 1)^2 = 9m^2 + 6m + 1 = 3(3m^2 + 2m) + 1$. Since $3m^2 + 2m$ is an integer, it follows that we can write $n^2 = 3k + 1$ for $k = 3m^2 + 2m$.

Case 2: Suppose $n = 3m+2$. Then $n^2 = (3m+2)^2 = 9m^2+12m+4 = 9m^2+12m+3+1 = 3(3m^2 + 4m + 1) + 1$. So we can write $n^2 = 3k + 1$ for $k = 3m^2 + 4m + 1$.

Since we have proven the statement for both cases, and since Case 1 and Case 2 reflect all possible possibilities, the theorem is true.

1.2 Proof by induction

We can use induction when we want to show a statement is true for all positive integers n. (Note that this is not the only situation in which we can use induction, and that induction is not (usually) the only way to prove a statement for all positive integers.)

To use induction, we prove two things:

- **Base case:** The statement is true in the case where $n = 1$.
- **Inductive step:** If the statement is true for $n = k$, then the statement is also true for $n = k + 1$.

This actually produces an infinite chain of implications:

- The statement is true for $n = 1$
- If the statement is true for $n = 1$, then it is also true for $n = 2$
- If the statement is true for $n = 2$, then it is also true for $n = 3$
- If the statement is true for $n = 3$, then it is also true for $n = 4$
- $\ldots$

Together, these implications prove the statement for all positive integer values of n. (It does not prove the statement for non-integer values of n, or values of n less than 1.)

Example: Prove that $1 + 2 + \cdots + n = n(n + 1)/2$ for all integers $n \geq 1$.

Proof: We proceed by induction.

Base case: If $n = 1$, then the statement becomes $1 = 1(1 + 1)/2$, which is true.

Inductive step: Suppose the statement is true for $n = k$. This means $1 + 2 + \cdots + k = k(k+1)/2$. We want to show the statement is true for $n = k+1$, i.e. $1+2+\cdots+k+(k+1) = (k+1)(k+2)/2$.

By the induction hypothesis (i.e. because the statement is true for $n = k$), we have $1 + 2 + \cdots + k + (k+1) = k(k+1)/2 + (k+1)$. This equals $(k+1)(k/2+1)$, which is equal to $(k+1)(k+2)/2$. This proves the inductive step.

Therefore, the statement is true for all integers $n \geq 1$.

1.2.1 Strong induction

Strong induction is a useful variant of induction. Here, the inductive step is changed to

- **Base case:** The statement is true when $n = 1$.
- **Inductive step:** If the statement is true for all values of $1 \leq n < k$, then the statement is also true for $n = k$.

This also produces an infinite chain of implications:

- The statement is true for $n = 1$
- If the statement is true for $n = 1$, then it is true for $n = 2$
- If the statement is true for both $n = 1$ and $n = 2$, then it is true for $n = 3$
- If the statement is true for $n = 1$, $n = 2$, and $n = 3$, then it is true for $n = 4$
- ...

Strong induction works on the same principle as weak induction, but is generally easier to prove theorems with.

Example: Prove that every integer n greater than or equal to 2 can be factored into prime numbers.

Proof: We proceed by (strong) induction.

Base case: If $n = 2$, then n is a prime number, and its factorization is itself.

Inductive step: Suppose k is some integer larger than 2, and assume the statement is true for all numbers $n < k$. Then there are two cases:

Case 1: k is prime. Then its prime factorization is just k.

Case 2: k is composite. This means it can be decomposed into a product xy, where x and y are both greater than 1 and less than k. Since x and y are both less than k, both x and y can be factored into prime numbers (by the inductive hypothesis). That is, $x = p_1 \ldots p_s$ and $y = q_1 \ldots q_t$ where $p_1, \ldots, p_s$ and $q_1, \ldots, q_t$ are prime numbers.

Thus, k can be written as $(p_1 \ldots p_s) \cdot (q_1 \ldots q_t)$, which is a factorization into prime numbers.

This proves the statement.

2 Important fact from calculus

The definition of the exponential function says that

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

In particular, this means that $\lim_{n\to\infty}(1 + \frac{1}{n})^n = e$ and $\lim_{n\to\infty}(1 - \frac{1}{n})^n = \frac{1}{e}$.

3 Linear algebra

In this section we will discuss vectors and matrices. We denote the (i, j)th entry of a matrix A as A_{ij}, and the ith entry of a vector as v_i.

3.1 Vectors and vector operations

A vector is a one dimensional matrix, and it can be written as a column vector:

$$\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

or a row vector:

$$\begin{bmatrix} v_1 & v_2 & \dots & v_n \end{bmatrix}$$

3.1.1 Norm

The ℓ_2 norm, or length, of a vector $(v_1, \dots, v_n)$ is just $\sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}$. The norm of a vector v is usually written as $||v||$.

3.1.2 Triangle inequality

For two vectors u and v, we have

$$||u + v|| \leq ||u|| + ||v||$$

and

$$||u - v|| \geq \Big| ||u|| - ||v|| \Big|$$

3.1.3 Dot product

The dot product of two equal-length vectors $(u_1, \dots, u_n)$ and $(v_1, \dots, v_n)$ is $u \cdot v = u_1v_1 + u_2v_2 + \cdots + u_nv_n$.

Two vectors are orthogonal if their dot product is zero. The angle between the vectors is acute if the dot product is greater than 0, and is obtuse otherwise.

If θ is the angle swept out between two vectors u and v, the dot product between them is $u \cdot v = ||u||||v|| \cos\theta$. This leads to the notion of "cosine similarity" between vectors, a similarity measure between -1 and 1: $\cos\theta = \frac{u \cdot v}{||u||||v||}$.

Finally, an identity which comes in useful is $u \cdot u = ||u||^2$.

3.2 Matrix operations

3.2.1 Matrix addition

Matrix addition is defined for matrices of the same dimension. Matrices are added componentwise:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} + \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 1+5 & 2+6 \\ 3+7 & 4+8 \end{bmatrix} = \begin{bmatrix} 6 & 8 \\ 10 & 12 \end{bmatrix}$$

3.2.2 Matrix multiplication

Matrices can be multiplied like so:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 1\cdot5+2\cdot7 & 1\cdot6+2\cdot8 \\ 3\cdot5+4\cdot7 & 3\cdot6+4\cdot8 \end{bmatrix} = \begin{bmatrix} 19 & 22 \\ 43 & 50 \end{bmatrix}$$

You can also multiply non-square matrices, but the dimensions have to match (i.e. the number of columns of the first matrix has to equal the number of rows of the second matrix).

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} = \begin{bmatrix} 1\cdot1+2\cdot4 & 1\cdot2+2\cdot5 & 1\cdot3+2\cdot6 \\ 3\cdot1+4\cdot4 & 3\cdot2+4\cdot5 & 3\cdot3+4\cdot6 \\ 5\cdot1+6\cdot4 & 5\cdot2+6\cdot5 & 5\cdot3+6\cdot6 \end{bmatrix} = \begin{bmatrix} 9 & 12 & 15 \\ 19 & 26 & 33 \\ 29 & 40 & 51 \end{bmatrix}$$

In general, if matrix A is multiplied by matrix B, we have $(AB)_{ij} = \sum_k A_{ik}B_{kj}$ for all entries (i, j) of the matrix product.

Matrix multiplication is associative, i.e. $(AB)C = A(BC)$. It is also distributive, i.e. $A(B + C) = AB + AC$. However, it is **not** commutative. That is, AB does not have to equal BA.

Note that if you multiply a 1-by-n matrix with an n-by-1 matrix, that is the same as taking the dot product of the corresponding vectors.

3.2.3 Matrix transpose

The transpose operation switches a matrix's rows with its columns, so

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}^T = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}$$

In other words, we define A^T by $(A^T)_{ij} = A_{ji}$.

Properties:

- $(A^T)^T = A$
- $(AB)^T = B^T A^T$
- $(A + B)^T = A^T + B^T$

3.2.4 Identity matrix

The identity matrix I_n is an n-by-n matrix with all 1's on the diagonal, and 0's everywhere else. It is usually abbreviated I, when it is clear what the dimensions of the matrix are.

It has the property that when you multiply it by any other matrix, you get that matrix. In other words, if A is an m-by-n matrix, then $AI_n = I_m A = A$.

3.2.5 Matrix inverse

The inverse of a matrix A is the matrix that you can multiply A by to get the identity matrix. Not all matrices have an inverse. (The ones that have an inverse are called *invertible.*)

In other words, A^{-1} is the matrix where $AA^{-1} = A^{-1}A = I$ (if it exists). Note that A must be square to have an inverse.

Properties:

- $(A^{-1})^{-1} = A$
- $(AB)^{-1} = B^{-1}A^{-1}$
- $(A^{-1})^T = (A^T)^{-1}$

3.3 Types of matrices

3.3.1 Diagonal matrix

A diagonal matrix is a matrix that has 0's everywhere except the diagonal. A diagonal matrix can be written $D = diag(d_1, d_2, \ldots, d_n)$, which corresponds to the matrix

$$\begin{bmatrix} d_1 & 0 & \ldots & 0 \\ 0 & d_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & d_n \end{bmatrix}$$

You may verify that

$$D^k = \begin{bmatrix} d_1^k & 0 & \ldots & 0 \\ 0 & d_2^k & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & d_n^k \end{bmatrix}$$

3.3.2 Triangular matrix

A lower triangular matrix is a matrix that has all its nonzero elements on or below the diagonal. An upper triangular matrix is a matrix that has all its nonzero elements on or above the diagonal.

3.3.3 Symmetric matrix

A is symmetric if $A = A^T$, i.e. $A_{ij} = A_{ji}$ for all entries (i, j) in A. Note that a matrix must be square in order to be symmetric.

3.3.4 Orthogonal matrix

A matrix U is orthogonal if $UU^T = U^TU = I$. (That is, the inverse of an orthogonal matrix is its transpose.)

Orthogonal matrices have the property that every row is orthogonal to every other row. That is, the dot product of any row vector with any other row vector is 0. In addition, every row is a unit vector, i.e. it has norm 1. (Try verifying this for yourself!)

Similarly, every column is a unit vector, and every column is orthogonal to every other column. (You can verify this by noting that if U is orthogonal, then U^T is also orthogonal.)

3.4 Linear independence and span

A linear combination of the vectors $v_1, \ldots, v_n$ is an expression of the form $a_1v_1 + a_2v_2 + \cdots + a_nv_n$, where $a_1, \ldots, a_n$ are real numbers. Note that some of the a_i's may be zero.

The span of a set of vectors is the set of all possible linear combinations of that set of vectors.

The vectors $v_1, \ldots, v_n$ are linearly independent if you cannot find coefficients $a_1, \ldots, a_n$ where $a_1v_1 + \cdots + a_nv_n = 0$ (except for the trivial solution $a_1 = a_2 = \cdots = 0$). Intuitively, this means you cannot write any of the vectors in terms of any linear combination of the other vectors. (A set of vectors is linearly dependent if it is not linearly independent.)

If we let $V = [v_1, \ldots, v_n]$ be the matrix whose columns are the vectors v_i, then the statement that the vectors are linearly independent is equivalent to say that the only solution to $Vx = 0$ is $x = 0$ (where 0 is the vector of all zeros). A matrix with this property is called *full-rank.*

3.5 Eigenvalues and eigenvectors

Sometimes, multiplying a matrix by a vector just stretches that vector. If that happens, the vector is called an eigenvector of the matrix, and the “stretching factor” is called the eigenvalue.

Definition: Given a square matrix A, λ is an eigenvalue of A with the corresponding eigenvector x if $Ax = \lambda x$. (Note that in this definition, x is a vector, and λ is a number.)

(By convention, the zero vector cannot be an eigenvector of any matrix.)

Example: If

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

then the vector $\begin{bmatrix} 3 \\ -3 \end{bmatrix}$ is an eigenvector with eigenvalue 1, because

$$Ax = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ -3 \end{bmatrix} = \begin{bmatrix} 3 \\ -3 \end{bmatrix} = 1 \cdot \begin{bmatrix} 3 \\ -3 \end{bmatrix}$$

3.5.1 Solving for eigenvalues and eigenvectors

We exploit the fact that $Ax = \lambda x$ if and only if $(A - \lambda I)x = 0$. (Note that λI is the diagonal matrix where all the diagonal entries are λ, and all other entries are zero.)

This equation has a nonzero solution for x if and only if the determinant of $A - \lambda I$ equals 0. (We won't prove this here, but you can google for "invertible matrix theorem".) Therefore, you can find the eigenvalues of the matrix A by solving the equation $det(A - \lambda I) = 0$ for λ. Once you have done that, you can find the corresponding eigenvector for each eigenvalue λ by solving the system of equations $(A - \lambda I)x = 0$ for x.

Example: If

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

then

$$A - \lambda I = \begin{bmatrix} 2 - \lambda & 1 \\ 1 & 2 - \lambda \end{bmatrix}$$

and

$$det(A - \lambda I) = (2 - \lambda)^2 - 1 = \lambda^2 - 4\lambda + 3$$

Setting this equal to 0, we find that $\lambda = 1$ and $\lambda = 3$ are possible eigenvalues.

To find the eigenvectors for $\lambda = 1$, we plug λ into the equation $(A - \lambda I)x = 0$. This gives us

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Any vector where $x_2 = -x_1$ is a solution to this equation, and in particular, $\begin{bmatrix} 3 \\ -3 \end{bmatrix}$ is one solution.

To find the eigenvectors for $\lambda = 3$, we again plug λ into the equation, and this time we get

$$\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Any vector where $x_2 = x_1$ is a solution to this equation.

(**Note:** The above method is never used to calculate eigenvalues and eigenvectors for large matrices in practice, iterative methods are used instead.)

3.5.2 Properties of eigenvalues and eigenvectors

- Usually eigenvectors are normalized to unit length
- If A is symmetric, then all its eigenvalues are real
- The eigenvalues of any triangular matrix are its diagonal entries
- The trace of a matrix (i.e. the sum of the elements on its diagonal) is equal to the sum of its eigenvalues
- The determinant of a matrix is equal to the product of its eigenvalues

3.6 Matrix eigendecomposition

Theorem: Suppose A is an n-by-n matrix with n linearly independent eigenvectors. Then A can be written as $A = PDP^{-1}$, where P is the matrix whose columns are the eigenvectors of A, and D is the diagonal matrix whose entries are the corresponding eigenvalues.

In addition, $A^2 = (PDP^{-1})(PDP^{-1}) = PD^2P^{-1}$, and $A^n = PD^nP^{-1}$. (This is interesting because it's much easier to raise a diagonal matrix to a power than to exponentiate an ordinary matrix.)

Note that if A is symmetric, then it has eigenvectors that are linearly independent. If we let P be the matrix whose columns are the normalized eigenvectors, then $P^{-1} = P^T$, since P is orthogonal, and $A = PDP^T$.

4 Probability

4.1 Fundamentals

The sample space Ω represents the set of all possible things that can happen. For example, if you are rolling a die, your sample space is $\{1, 2, 3, 4, 5, 6\}$.

An event is a subset of the sample space. For example, the event "I roll a number less than 4" can be represented by the subset $\{1, 2, 3\}$. The event "I roll a 6" can be represented by the subset $\{6\}$.

A probability function is a mapping from events to real numbers between 0 and 1. It must have the following properties:

- $P(\Omega) = 1$
- $P(A \cup B) = P(A) + P(B)$ for **disjoint** events A and B (i.e. when $A \cap B = \emptyset$)

Example: For the dice example, we can define the probability function by saying $P(\{i\}) = 1/6$ for $i = 1, \ldots, 6$. (That is, we say that each number has an equal probability of being rolled.) All events in the probability space can be represented as unions of these six disjoint events.

Using this definition, we can compute the probability of more complicated events, like

$$P(\text{we roll an odd number}) = 1/6 + 1/6 + 1/6 = 1/2.$$

(Note that we can add probabilities here because the events $\{1\}$, $\{3\}$, and $\{5\}$ are disjoint.)

4.2 Principle of Inclusion-Exclusion

We begin by considering two sets A and B. When A and B are not disjoint, we have

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Proof: You can derive this theorem from the probability axioms. $A \cup B$ can be split into three disjoint events: $A \setminus B$, $A \cap B$, and $B \setminus A$. Furthermore, A can be split into $A \setminus B$ and $A \cap B$, and B can be split into $B \setminus A$ and $A \cap B$. So

$$\begin{aligned} P(A \cup B) &= P(A \setminus B) + P(A \cap B) + P(B \setminus A) \\ &= P(A \setminus B) + P(A \cap B) + P(B \setminus A) + P(A \cap B) - P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

Example: Suppose k is chosen uniformly at random from the integers $1, 2, \ldots, 100$. (This means the probability of getting each integer is 1/100.) Find the probability that k is divisible by 2 or 5.

By the Principle of Inclusion-Exclusion, $P(k \text{ is divisible by 2 or 5}) = P(k \text{ is divisible by 2}) + P(k \text{ is divisible by 5}) - P(k \text{ is divisible by both 2 and 5})$. There are 50 numbers divisible by 2, 20 numbers divisible by 5, and 10 numbers divisible by 10 (i.e., divisible by both 2 and 5). Therefore, the probability is $50/100 + 20/100 - 10/100 = 60/100 = 0.6$.

We now generalize to the case of n sets $X_1 \ldots X_n \subseteq Y$. In this case, the inclusion-exclusion principle can be stated as

$$\left|\bigcup_{i=1}^{k} X_i\right| = \sum_{I \subseteq \{1,\ldots,k\}, I \neq \emptyset} (-1)^{|I|-1} \left|\bigcap_{i \in I} X_i\right|$$

We can prove this by induction: think of $X_1 \cup \cdots \cup X_k$ as $(X_1 \cup \cdots \cup X_{k-1}) \cup X_k$ and use the case of two sets:

$$|(X_1 \cup \cdots \cup X_{k-1}) \cup X_k| = |X_1 \cup \cdots \cup X_{k-1}| + |X_k| - |(X_1 \cup \cdots \cup X_{k-1}) \cap X_k|$$

and recall the inductive fact that we have formulas to count $|X_1 \cup \cdots \cup X_{k-1}|$ and $|(X_1 \cap X_k) \cup \cdots \cup (X_{k-1} \cap X_k)|$. Try this at home!

4.3 Union bound

For any collection of n events $A_1, \ldots, A_n$, we have

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i)$$

Proof: We can prove this by induction (for finite n).

Base case: Suppose $n = 1$. Then the statement becomes $P(A_1) \leq P(A_1)$, which is true.

Inductive step: Suppose the statement is true for $n = k$. We must prove that the statement is true for $n = k + 1$. We have

$$\bigcup_{i=1}^{k+1} A_i = \left(\bigcup_{i=1}^{k} A_i \right) \cup A_{k+1}$$

and by the Principle of Inclusion-Exclusion,

$$P\left(\bigcup_{i=1}^{k+1} A_i \right) \leq P\left(\bigcup_{i=1}^{k} A_i \right) + P(A_{k+1})$$

By the induction hypothesis, the first term is less than or equal to $\sum_{i=1}^{k} P(A_i)$. So

$$P\left(\bigcup_{i=1}^{k+1} A_i \right) \leq \sum_{i=1}^{k+1} P(A_i)$$

proving the theorem.

Example: Suppose you have a 1 in 100000 chance of getting into a car accident every time you drive to work. If you drive to work every day of the year, how likely are you to get in a car accident on your way to work?

Answer: The union bound will not tell you exactly how likely you are to get in a car accident. However, it will tell you that the probability is upper bounded by 365/100000.

4.4 Conditional Probability and Bayes' Rule

Suppose you are administering the GRE, and you discover that 2.5% of students get a perfect score on the math section.[1] By itself, this is not a very useful statistic, because the scores on the math section vary substantially by major. You dig a little deeper and find that 7.5% of physical sciences students get a perfect score, 6.3% of engineering students get a perfect score, and most other majors do substantially worse.[2]

In the language of conditional probability, we would say that the probability of getting a perfect score, given that you are a engineering major, is 6.3%:

$$P(\text{perfect score} \mid \text{engineering major}) = 0.063$$

If we want to actually compute this probability, we would take the number of engineering majors that receive a perfect score, and divide it by the total number of engineering majors. This is equivalent to computing the formula

[1] See https://www.ets.org/s/gre/pdf/gre_guide_table4.pdf for a breakdown by specific majors.

[2] For some reason, computer science is counted as part of the physical sciences, and not as engineering.

$$P(\text{perfect score} \mid \text{engineering major}) = \frac{P(\text{perfect score} \cap \text{engineering major})}{P(\text{engineering major})}$$

(In general, we can replace "perfect score" and "engineering major" with any two events, and we get the formal definition of conditional probability.)

Example: Suppose you toss a fair coin three times. What is the probability that all three tosses come up heads, given that the first toss came up heads?

Answer: This probability is

$$\frac{P(\text{all three tosses come up heads and the first toss came up heads})}{P(\text{the first toss came up heads})} = \frac{1/8}{1/2} = \frac{1}{4}$$

4.4.1 Independence

Two events are independent if the fact that one event happened does not affect the probability that the other event happens. In other words

$$P(A|B) = P(A)$$

This also implies that

$$P(A \cap B) = P(A)P(B)$$

Example: We implicitly used the independence assumption in the previous calculation, when we were computing the probability that all three coin tosses come up heads. This probability is 1/8 because the probability that each toss comes up heads is 1/2, and the three events are independent of each other.

4.4.2 Bayes' Rule

We can apply the definition of conditional probability to get

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

In addition, we can say

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B \cap A) + P(B \cap \text{not } A)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\text{not } A)P(\text{not } A)}$$

Example: Suppose 1% of women who enter your clinic have breast cancer, and a woman with breast cancer has a 90% chance of getting a positive result, while a woman without breast cancer has a 10% chance of getting a false positive result. What is the probability of a woman having breast cancer, given that she just had a positive test?

Answer: By Bayes' Rule,

$$\begin{aligned}
P(\text{cancer} \mid \text{positive}) &= \frac{P(\text{positive} \mid \text{cancer})P(\text{cancer})}{P(\text{positive})} \\
&= \frac{P(\text{positive} \mid \text{cancer})P(\text{cancer})}{P(\text{positive} \mid \text{cancer})P(\text{cancer}) + P(\text{positive} \mid \text{not cancer})P(\text{not cancer})} \\
&= \frac{0.9 \cdot 0.01}{0.9 \cdot 0.01 + 0.1 \cdot 0.99} \\
&= 8.3\%
\end{aligned}$$

4.5 Random variables

A random variable X is a variable that can take on different values depending on the outcome of some probabilistic process. It can be defined as a function $X : \Omega \to \mathbb{R}$ that yields a different real number depending on which point in the sample space you choose.

Example: Suppose we are tossing three coins. Let X be the number of coins that come up heads. Then $P(X = 0) = 1/8$.

4.5.1 PDFs and CDFs

A random variable can take on either a discrete range of values or a continuous range of values. If it takes a discrete range of values, the function that assigns a probability to each possible value is called the probability mass function.

Example: Let X be the number shown on a fair six-sided die. Then the probability mass function for X is $P(X = i) = 1/6$.

If the random variable takes a continuous range of values, the equivalent of the probability mass function is called the probability density function. The tricky thing about probability density functions is that often, the probability of getting a specific number (say $X = 3.258$) is zero. So we can only talk about the probability of getting a number that lies within a certain range.

We define $f(x)$ to be the probability density function of a continuous random variable X if $P(a \leq X \leq b) = \int_a^b f(x)dx$. Here the probability is just the area under the curve of the PDF.

The PDF must have the following properties:

- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x)dx = 1$
- $\int_{x \in A} f(x)dx = P(X \in A)$

The cumulative distribution function (or CDF) of a real valued random variable X expresses the probability that the random variable is less than or equal to the argument. It is given by $F(x) = P(X \leq x)$.

The CDF can be expressed as the integral of the PDF, in that

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

The CDF must have the following properties:

- $F(x)$ must be between 0 and 1
- $F(x)$ must be nondecreasing
- $F(x)$ must be right-continuous
- $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$

4.6 Expectation and variance

4.6.1 Expectation

The expected value (or mean) of a random variable can be interpreted as a weighted average. For a discrete random variable, we have

$$E[X] = \sum_{x} x \cdot P(X = x)$$

For a continuous random variable,

$$E[X] = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

where $f(x)$ is the probability density function.

Example: Suppose your happiness is a 10 when it's sunny outside, and a 2 when it's raining outside. It's sunny 80% of the time and raining 20% of the time. What is the expected value of your happiness?

Answer: $10 \cdot 0.8 + 2 \cdot 0.2 = 8.4$.

4.6.2 Linearity of expectation

If X and Y are two random variables, and a is a constant, then

$$E[X + Y] = E[X] + E[Y]$$

and

$$E[aX] = aE[X]$$

This is true even if X and Y are not independent.

4.6.3 Variance

The variance of a random variable is a measure of how far away the values are, on average, from the mean. It is defined as

$$Var(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

For a random variable X and a constant a, we have $Var(X + a) = Var(X)$ and $Var(aX) = a^2 Var(X)$.

We **do not** have $Var(X + Y) = Var(X) + Var(Y)$ unless X and Y are uncorrelated (which means they have covariance 0). In particular, independent random variables are always uncorrelated, although the reverse doesn't hold.

4.7 Special random variables

4.7.1 Bernoulli random variables

A Bernoulli random variable with parameter p can be interpreted as a coin flip that comes up heads with probability p, and tails with probability $1 - p$.

If X is a Bernoulli random variable, i.e. $X \sim \mathit{Bernoulli}(p)$, then $P(X = 1) = p$ and $P(X = 0) = 1 - p$.

We also have

$$E[X] = 1 \cdot p + 0 \cdot (1 - p) = p$$

and

$$Var(X) = E[X^2] - (E[X])^2 = p - p^2 = p(1 - p)$$

4.7.2 Geometric random variables

Suppose you keep flipping a coin until you get heads. A geometric random variable with parameter p measures how many times you have to flip the coin if each time it has a probability p of coming up heads. It is defined by the distribution

$$P(X = k) = p(1 - p)^{k-1}$$

Furthermore, $E[X] = 1/p$ and $Var(X) = (1 - p)/p^2$.

4.7.3 Uniform random variables

A uniform random variable is a continuous random variable, where you sample a point uniformly at random from a given interval. If $X \sim Uniform(a, b)$, then the probability density function is given by

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

We have $E[X] = (a + b)/2$, and $Var(X) = (b - a)^2/12$.

4.7.4 Normal random variables

A normal random variable is a point sampled from the normal distribution, which has all sorts of interesting statistical properties. If $X \sim Normal(\mu, \sigma^2)$, then the probability density function is given by

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

Also, $E[X] = \mu$ and $Var(X) = \sigma^2$.

4.8 Indicator random variables

An indicator random variable is a variable that is 1 if an event occurs, and 0 otherwise:

$$I_A = \begin{cases} 1 & \text{if event } A \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

The expectation of an indicator random variable is just the probability of the event occurring:

$$\begin{aligned} E[I_A] &= 1 \cdot P(I_A = 1) + 0 \cdot P(I_A = 0) \\ &= P(I_A = 1) \\ &= P(A) \end{aligned}$$

Indicator random variables are very useful for computing expectations of complicated random variables, especially when combined with the property that the expectation of a sum of random variables is the sum of the expectations.

Example: Suppose we are flipping n coins, and each comes up heads with probability p. What is the expected number of coins that come up heads?

Answer: Let X_i be the indicator random variable that is 1 if the ith coin comes up heads, and 0 otherwise. Then $E[\sum_{i=1}^n X_i] = \sum_{i=1}^n E[X_i] = \sum_{i=1}^n p = np$.

4.9 Inequalities

4.9.1 Markov's inequality

For any random variable X that takes only non-negative values, we have

$$P(X \geq a) \leq \frac{E[X]}{a}$$

for $a > 0$.

You can derive this as follows. Let $I_{X \geq a}$ be the indicator random variable that is 1 if $X \geq a$, and 0 otherwise. Then $aI_{X \geq a} \leq X$ (convince yourself of this!). Taking expectations on both sides, we get $aE[I_{X \geq a}] \leq E[X]$, so $P(X \geq a) \leq E[X]/a$.

4.9.2 Chebyshev's inequality

If we apply Markov's inequality to the random variable $(X - E[X])^2$, we get

$$P((X - E[X])^2) \geq a^2) \leq \frac{E[(X - E[X])^2]}{a^2}$$

or

$$P(|X - E[X]|) \geq a) \leq \frac{Var(X)}{a^2}$$

This gives a bound on how far a random variable can be from its mean.

4.9.3 Chernoff bound

Suppose $X_1, \ldots, X_n$ are independent Bernoulli random variables, where $P(X_i = 1) = p_i$. Denoting $\mu = E[\sum_{i=1}^n X_i] = \sum_{i=1}^n p_i$, we get

$$P\left(\sum_{i=1}^n X_i \geq (1+\delta)\mu\right) \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu$$

for any δ.

5 Material adapted from

Greg Baker, "Introduction to Proofs"

https://www.cs.sfu.ca/~ggbaker/zju/math/proof.html

CS 103 Winter 2016, "Guide to Proofs"

http://stanford.io/2dexnf9

Peng Hui How, "Proof? A Supplementary Note For CS161"

http://web.stanford.edu/class/archive/cs/cs161/cs161.1168/HowToWriteCorrectnessProof.pdf

Nihit Desai, Sameep Bagadia, David Hallac, Peter Lofgren, Yu "Wayne" Wu, Borja Pelato, "Quick Tour of Linear Algebra and Graph Theory"

http://snap.stanford.edu/class/cs224w-2014/recitation/linear_algebra/LA_Slides.pdf
http://snap.stanford.edu/class/cs224w-2015/recitation/linear_algebra.pdf

"Quick tour to Basic Probability Theory"

http://snap.stanford.edu/class/cs224w-2015/recitation/prob_tutorial.pdf

"Bayes' Formula"

http://www.math.cornell.edu/~mec/2008-2009/TianyiZheng/Bayes.html