

Appendix A

Probability Theory

©1999 by Dan Ashlock

This appendix reviews some terms and mathematical notions from probability theory used in this book that may not have appeared in your program of study or which you may have forgotten. Ubiquitous in the theory of artificial life is the notion of a *Markov chain*, a set of repeated trials that are not independent. On the way to the elementary parts of the theory of Markov chains, we will review a good deal of basic probability theory.

A.1 Basic Probability Theory.

A *distribution* D is a triple (Q, E, P) consisting of: a set of *points* Q , a collection of *events* E that are subsets of Q , and a function $P : E \rightarrow [0, 1]$ that assigns probabilities to events. How would we represent the familiar example of flipping a fair coin in this notation?

Example A.1 Flipping a fair coin *When D represents flipping a fair coin, we have point set $Q = \{heads, tails\}$, events $E = \{\{\}, \{heads\}, \{tails\}, \{heads, tails\}\}$, and probability assignment*

$$\begin{aligned}P(\{\}) &= 0 \\P(\{heads\}) &= 0.5 \\P(\{tails\}) &= 0.5 \\P(\{heads, tails\}) &= 1\end{aligned}$$

Probabilities are real numbers in the unit interval. There is one additional requirement to make a triple (Q, E, P) a distribution. As long as the set Q is finite or countably infinite, we demand that

$$\sum_{q \in Q} P(\{q\}) = 1. \tag{A.1}$$

In the event that Q is uncountable, we demand that

$$\int_{q \in Q} P(\{q\}) = 1. \quad (\text{A.2})$$

Typically, we confuse singleton sets with their sole member so that we define $P(q) := P(\{q\})$ for each $q \in Q$. You may wonder why we have points *and* events. Since events are built out of points, their presence seems redundant. There are two reasons. First, events consisting of many points in the distribution are often the actual objects of interest. Second, in the case in which Q is an uncountable set the probability of singleton point events is zero. This forces us to deal with multi-point events to get anything done.

Example A.2 The uniform distribution on $[0,1]$ A uniform distribution is one in which all points are equally likely. Notice the distribution in Example A.1 was uniform on two points. On an uncountable set, we achieve a uniform distribution by insisting that events the same size be assigned the same probability by P . Two events A and B are the same size if

$$\int_{a \in A} dx = \int_{b \in B} dx$$

A little work will show that, for the uniform distribution on $[0,1]$, we may take

$$P(x) = 1.$$

We compute the probability of an event by computing the integral of $P(x)$ on that event. Notice we have been vague about specifying what E is in this example. Events that are built from intervals by the operations of intersection, union, and complementation are safe. For a better treatment, a course in measure theory is required.

A *trial* is the result of sampling a point from a distribution, flipping a coin, for example. A way of looking at the probability of an event is that it is the chance that a point in the event will be chosen in a trial. A *set of repeated trials* is a collection of trials taken one after the other from the same distribution or a sequence of distributions. We can place a *product distribution* on repeated trials by letting the points in the product distribution be the possible sets of outcomes of a repeated trial and then inducing the events and their associated probabilities in the natural manner.

Example A.3 A product distribution Suppose we flip 3 coins. We then have an example of 3 repeated trials sampled from the distribution given in Example A.1. The set of 3 trials form a single trial in a (3-fold) product distribution. The points of this distribution are:

$$\{\{H, H, H\}, \{H, H, T\}, \{H, T, H\}, \{H, T, T\}, \{T, H, H\}, \{T, H, T\}, \{T, T, H\}, \{T, T, T\}\}.$$

The set of events consists of all 256 subsets of the set of points. Each single-point event has probability $1/8$, and the probability of an event is the number of points in it divided by 8.

Two events A and B are said to be *independent* if

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

An example of two independent events is as follows. If we flip two coins and put a product distribution on the 4 possible outcomes, then the events “the first coin comes up heads” and “the second coin comes up tails” are independent. If you want to know the probability of several independent events all happening, then you multiply their probabilities. The probability of getting 3 heads on 3 flips of a fair coin, for example, is $\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$. (Each of 3 independent flips has probability $\frac{1}{2}$ of producing a head. Multiply them to get the probability of 3 heads in a row.)

If two events are not independent then they are said to be *dependent*. Suppose, for example, we have a pot containing 5 black and 5 white balls, and we have two trials in which we draw balls out of the pot at random. If we do not replace the first ball before drawing the second, then the probability of drawing a black or white ball is dependent on what we drew the first time. In both trials, the events are $\{black\}$ or $\{white\}$, but the distribution of the second draw is changed by the first draw. The events “first ball is white” and “second ball is white” are *dependent* in the product distribution of the two trials.

If two events are such that either one happening completely precludes the other happening, then the events are said to be *disjoint*. Mathematically, A and B are disjoint if

$$P(A \cup B) = P(A) + P(B).$$

If you want to know the probability of one of several disjoint events happening, then you simply sum their probabilities. Each of the faces of a fair 6-sided die has probability $1/6$ of being rolled, and all 6 events are disjoint. The probability of rolling a prime number on a 6-sided die is $P(2) + P(3) + P(5) = 1/6 + 1/6 + 1/6 = 1/2$. (Try asking a friend to call “prime” or “non-prime” on a die instead of “heads” or “tails” on a coin. A humorous argument often ensues, especially in the presence of those who believe 1 to be prime.)

If a distribution is on a set of numbers, then a distribution has an *expected value*. One computes the expected value of a distribution on a set of numbers by summing the product of the numbers with their respective probabilities. Take, for example, the numbers 1 through 6, as generated by a fair die. The probability of each is $1/6$, and, so, the expected value of the roll of a single die is $\frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = 3.5$. The notion of expected value is a mathematical generalization of the more familiar notion of *average*. Formally, if $D = (Q, E, P)$ is a distribution for which $Q \subseteq \mathbb{R}$, then the expected value $E(D)$ is given by

$$E(D) = \sum_{q \in Q} q \cdot P(q) \tag{A.3}$$

Many introductory probability classes deal largely with sets of independent repeated trials or sets of disjoint events, because they are far easier to work with mathematically. The

modus operandi of evolution is to have strongly *dependent* trials. Rather than maintaining the same distribution by replacing balls in the pot between trials, we throw away most of the balls we draw and produce new balls by combining old ones in odd fashions. This means that dependent probability models are the norm in artificial life. The independent models are also useful; they can, for example, be used to understand the composition of the initial population in an evolutionary algorithm.

A.1.1 Choosing Things and Binomial Probability

The symbol $\binom{n}{k}$, pronounced “n choose k” is defined to be the number of different sets of k objects that can be chosen from a set of n objects. There is a simple formula for the choice numbers:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}. \quad (\text{A.4})$$

When choosing k objects out of n there are n choices for the first object, $n - 1$ choices for the second, and so on until there are

$$n \cdot (n - 1) \cdots (n - k + 1) = \frac{n!}{(n - k)!} \quad (\text{A.5})$$

ways to choose the set. These choices, however, have an implicit order, and, so, when choosing k objects, there are $k!$ distinct orders in which we could choose the same set. Dividing by $k!$ yields the desired formula. Since choosing and failing to choose objects are dual to one another, we obtain the useful identity

$$\binom{n}{k} = \binom{n}{n - k}, \quad (\text{A.6})$$

which also clearly follows from algebraic manipulation of the formula A.4. The choice numbers are also called the *binomial coefficients*, because of their starring role in the Binomial Theorem.

Theorem A.1 (*Binomial Theorem*)

$$(x + y)^n = \sum_{k=0}^n \binom{n}{k} x^k y^{n-k}.$$

A *Bernoulli trial* is a trial from a distribution $D = (Q, E, P)$ for which $|Q| = 2$. These two events happen with probability p and $1 - p$. One of the events is typically called a *success*, and the other is called a *failure*. The probability of success is p . The *Binomial Probability Model* is used to compute the probability of seeing some number of successes in an independent set of repeated Bernoulli trials.

Theorem A.2 (*Binomial Probability Model*) *If we are doing a set of n independent Bernoulli trials with probability p of success, then the probability of obtaining exactly k successes is*

$$\binom{n}{k} p^k (1-p)^{n-k}.$$

The Binomial Probability Model looks like a piece sliced out of the Binomial Theorem with p and $(1-p)$ taking the place of x and y . This is the result of identical counting arguments producing the Binomial Probability Model and the terms of the Binomial Theorem. If we are to have k successes, then we also have $n-k$ failures. Since the events are independent, we multiply the probabilities. Thus, any given sequence of successes and failures with k successes has probability $p^k(1-p)^{(n-k)}$. Since the successes form a k -subset of the trials, there are $\binom{n}{k}$ such sequences. We multiply the probability of a single sequence with k successes by the number of such sequences to obtain the probability of getting k successes - the Binomial Probability Model.

Example A.4 *Suppose that we have a population of 60 strings of length 20 that were produced by choosing characters “0” or “1” with a uniform distribution. What is the largest number of 1s we would expect to see in a member of the population? Answer this question by finding the number of 1s such that the expected number of creatures with that many 1s is (i) at least 1 and (ii) as small as possible.*

Answer:

The expected number of creatures with k 1s is just the population size times the result of plugging $p = 1/2$, $n = 20$ into the Binomial Probability Model. For 60 creatures, the expected number of creatures with 14 1s is 2.217. The expected number of creatures with 15 1s is 0.8871. So, 14 is a reasonable value for the largest number of 1s you would expect to see in such a population.

A quick way of generating binomial coefficients is to use *Pascal’s Triangle*, the first 11 rows of which are shown in Figure A.1. It is left to you to deduce how the triangle was generated and how to find a given binomial coefficient in the triangle.

A.1.2 Choosing Things to Count

In this section, we will use cards as our probability paradigm. We will use the machinery developed to learn something about single tournament selection. Some familiarity with poker is assumed, consult Hoyle or a friend if you are unfamiliar with this game.

Example A.5 *What is the number of 5-card poker hands that can be dealt?*

Answer:

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
1 6 15 20 15 6 1
1 7 21 35 35 21 7 1
1 8 28 56 70 56 28 8 1
1 9 36 84 126 126 84 36 9 1
1 10 45 120 210 252 210 120 45 10 1

Figure A.1: Pascal's Triangle from $n=0$ to $n=10$

Compute the number of ways to choose 5 out of 52 cards, that is:

$$\binom{52}{5} = \frac{52!}{5! \cdot 47!} = 2,598,960.$$

To get the probability of a given type of poker hand, you simply divide the number of ways to get the hand by the number of total hands. The next three examples illustrate this.

Example A.6 *What is the probability of getting three of a kind?*

Answer:

First let's solve the problem: "how many different poker hands are there that count as three of a kind?" Three of a kind is a hand that contains 3 cards with the same face value and 2 other cards with 2 other distinct face values. To get 3 cards the same, we choose the face value, choose 3 of the 4 cards with that face value, and then choose 2 of the other 49 cards, i.e., there are

$$\binom{13}{1} \cdot \binom{4}{3} \cdot \binom{49}{2} = 61,152$$

poker hands that contain 3 cards with the same face value.

We are not done yet! This counting includes hands with 4 cards the same ("four of a kind") and with 3 cards with one face value and the other 2 with another face value (a "full house"). Both of these are better than three of a kind and do not count as three of a kind.

To get the correct count, we must therefore count the number of ways to get four of a kind and a full house and subtract these from the total. Four of a kind is quite easy: simply choose a face value, choose all 4 cards of that face value, and then choose one of the 48 other cards. There are

$$\binom{13}{1} \cdot \binom{4}{4} \cdot \binom{48}{1} = 624$$

ways to get four of a kind.

A full house is a little harder: choose 1 of the 13 face values to be the “three the same,” choose 3 of those 4 cards, then choose 1 of the 12 remaining face values to be the “two the same,” and then choose 2 of the 4 cards with that face value. In short, there are

$$\binom{13}{1} \cdot \binom{4}{3} \cdot \binom{12}{1} \cdot \binom{4}{2} = 3744$$

different ways to get a full house.

Putting this all together, there are

$$61,152 - 624 - 3744 = 56,784$$

ways to get three of a kind.

To get the probability of getting three of a kind, we divide by the total number of poker hands.

$$P(\text{three-of-a-kind}) = \frac{56,784}{2,598,960} \approx 0.02185.$$

Example A.7 How many ways are there to get two of a kind?

Answer:

Again, we start by counting the number of hands that are two of a kind: 2 cards with the same face value and the other 3 with distinct face values. Since a large number of different types of good poker hands contain 2 cards with the same face value, it would be risky to follow the count-and-subtract technique used in Example A.6. We will, therefore, compute directly.

First, we select 1 of the 13 face values for our “two the same” and then choose 2 of those 4 values. This leaves 12 face values from which we must select 3 distinct face values to fill out the hand. Once we know these 3 face values, it follows that we must choose 1 of the 4 cards within each of these face values. This gives us

$$\binom{13}{1} \cdot \binom{4}{2} \cdot \binom{12}{3} \cdot \binom{4}{1}^3 = 1,098,240$$

ways to get two of a kind.

Dividing by the total number of poker hands, we get

$$P(\text{two-of-a-kind}) = \frac{1,098,240}{2,598,960} \approx 0.42256903.$$

One odd fact about poker hands is that the more valuable ones are easier to count. This is because they are not themselves included in still more valuable hands above them. The *flush*, a hand in which all 5 cards have the same suit, is quite easy to count, especially since a royal flush or a straight flush are, via linguistic technicality, still flushes.

Example A.8 *What is the probability of getting a flush?*

Answer:

First count the number of flush hands. We must choose 1 of 4 suits and then pick which 5 of the 13 cards in that suit we want. Thus, there are

$$\binom{4}{1} \cdot \binom{13}{5} = 5148$$

different ways to get a flush, yielding

$$P(\text{flush}) = \frac{5148}{2,598,960} \approx 0.001981.$$

Now, with the mental machinery all charged up to count things using *choose*, we can explore an issue concerning single tournament selection with tournament size 4. What is the expected number of children a creature participating in single tournament selection will have in each generation? First, let us agree that when two parents have two children, each incorporating some fraction of each parent's gene, this counts as one child. This means that, in single tournament selection, the expected number of children of a parent is one times the probability that parent will be placed by the random selection in a tournament in which it is one of the two most fit. Clearly, this probability can be computed from a creature's rank in the population in a given generation. (We will assume that, when there are ties in fitness, they do not lead to ties in rank, but, rather, rank is selected among equally fit creatures uniformly at random.)

Theorem A.3 *The expected number of children a creature with rank k out of a population of n creatures using single tournament selection as the model of evolution is:*

$$\frac{\binom{n-k}{3} + \binom{n-k}{2} \binom{k-1}{1}}{\binom{n-1}{3}}.$$

Proof:

There are two disjoint events that together make up the event in which we are interested, a creature being one of the 2 most fit creatures in its group of 4. Either it can be the top creature, or it can be the 2nd in its group of 4. The number of choices of other creatures that leave the creature in question at the top is simply the number of creatures less fit than it choose 3, $\binom{n-k}{3}$. If it is the second creature, then we choose 2 creatures from those less fit, $\binom{n-k}{2}$, and 1 from those more fit, $\binom{k-1}{1}$. Since these events are disjoint, they add. Finally, divide by the number of possible ways to choose 3 creatures to obtain a probability. Finally, notice that, in tournament selection, this probability is equal to the expected number of children \square

To give a feel for how the expected number of children is distributed, we show the probabilities for a population of size 24 in Example A.9. It is interesting to note that the probability of death is exactly one minus the probability of having children in this model of evolution when the tournament size is 4. As an exercise, you could compute the probability based on rank of becoming a parent or of dying for tournament sizes other than 4.

Example A.9 *Probability of tournament selection*

<i>Rank</i>	<i>Expected Children</i>	<i>Rank</i>	<i>Expected Children</i>
1	1	13	0.4658
2	1	14	0.3981
3	0.9881	15	0.3320
4	0.9656	16	0.2688
5	0.9334	17	0.2095
6	0.8927	18	0.1553
7	0.8447	19	0.1073
8	0.7905	20	0.0666
9	0.7312	21	0.0344
10	0.6680	22	0.0119
11	0.6019	23	0
12	0.5342	24	0

A.1.3 Binomial and Normal Confidence Intervals

In many of the experiments in this book, we record the time until success, in generations or mating events, for a large number of populations. When there are variations in the evolutionary algorithms used to produce those times, we can ask which variation of the algorithm worked better. Let us imagine we are studying the difference between single point and probabilistic mutation in a string evolver of the sort used in Chapter 2. Figure A.2 gives a graph of the fraction of populations that contain a copy of the reference string as a function of the number of generations. The graphs show that single point mutation outperforms probabilistic mutation at whatever rate it was performed. The question remains, “is the difference significant?”

Answering the question of significance can be done precisely. What you do is compute the probability that two experiments could be as different as they are by chance. To do this we construct confidence intervals. A *confidence interval with a given p value for a quantity q* is a range of values $q_l \leq q \leq q_h$ such that the probability that the true value of q is between q_l and q_h is p . A general treatment of confidence intervals is given in any mathematical statistics book. We will treat three different sorts of confidence intervals - binomial, normal, and normal approximation to the binomial. We now define some of the elementary terminology of statistics.

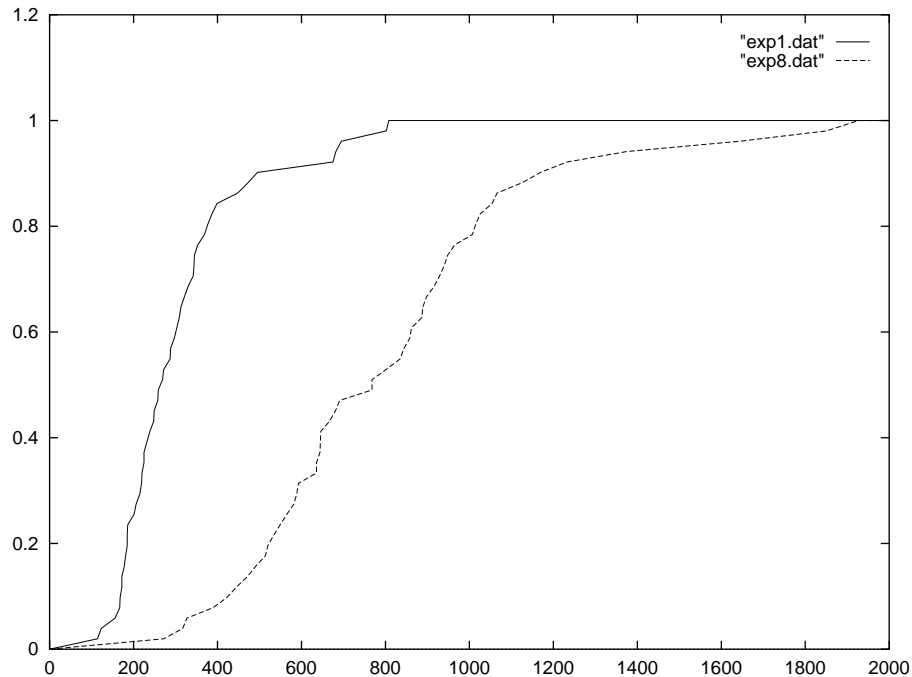


Figure A.2: Fraction of populations with a correct answer as a function of number of generations (Exp1.dat holds the data for a string evolver using single point mutation. Exp8.dat holds the data for a string evolver using probabilistic mutation.)

Definition A.1 A random variable X with distribution $D = (Q, E, P)$ is a surrogate for choosing a point from Q with probability as specified by P .

A random variable X associated with flipping a coin has the distribution given in Example A.1. It has two possible outcomes: “heads” and “tails”. A random variable can be thought of as an instance of its distribution.

There are two important quantities associated with a random variable over a set of numbers: its mean and variance. The mean of a random variable is just its expected value (see Equation A.3). We denote the mean of a random variable X by the symbol μ_X . Restating Equation A.3 for a random variable X with distribution $D = (Q, E, P)$, we have

$$\mu_X = E(X) = \sum_{q \in Q} q \cdot P(\{q\}), \text{ or} \quad (\text{A.7})$$

$$\mu_X = E(X) = \int_Q q \cdot P(q) \cdot dq. \quad (\text{A.8})$$

The variance of a random variable is the degree to which it tends to differ from its mean. It

is denoted by σ_X^2 . Formally, the variance of a random variable X is given by:

$$\sigma_X^2 = E((X - \mu_X)^2) = E(X^2) - \mu_X^2. \quad (\text{A.9})$$

The variance is denoted by σ_X^2 in part because the square root of the variance is also a commonly used quantity, the *standard deviation*.

Definition A.2 *The standard normal distribution, denoted $N(0, 1)$, is a distribution with $Q = \mathbb{R}$ and*

$$P(E) = \frac{1}{\sqrt{2\pi}} \int_E e^{-\frac{x^2}{2}} \cdot dx.$$

The mean of this distribution is 0 and the variance is 1. The normal distribution with mean μ and standard deviation σ , denoted $N(\mu, \sigma)$ is a distribution with $Q = \mathbb{R}$ and

$$P(E) = \frac{1}{\sqrt{2\pi}} \int_E e^{-\frac{(x-\mu)^2}{2\sigma^2}} \cdot dx.$$

We now have the pieces we need to construct confidence intervals.

A.2 Markov Chains

To analyze a series of trials that are not independent, the first mathematical technology to try is Markov chains. A *Markov chain* is a set S of states together with transition probabilities $p_s(t)$ of moving from state t to state s for any two $s, t \in S$. When you use a Markov chain, you start with an initial distribution on the states of the chain. If you know in which state you are starting, then the initial distribution will have probability one of being in that starting state. If your starting state is the distribution of an initial random population yet to be created, then you may have some initial probability of being in each state. The examples in this section should help clarify this notion.

We will be dealing only with Markov chains that have *stationary transition probabilities*. In this sort of Markov chain, the numbers $p_s(t)$ are fixed constants that have no dependence on history. We restrict our focus for clarity's sake and warn you that stochastic models of evolution, a topic beyond the scope of this text, will involve Markov chains with history dependent transition probabilities.

Example A.10 *Suppose we generate a sequence of integers by the following rule. The first integer is 0 and subsequent members of the sequence are generated by flipping a coin and adding 1 to the previous number if the coin came up heads. The states of this Markov chain are $S = \{0, 1, 2, \dots\}$. The transition probabilities are:*

$$p_s(t) = \begin{cases} 0.5 & s=t \text{ or } s=t+1 \\ 0 & \text{otherwise} \end{cases},$$

and the initial distribution of states is to be in state 0 with probability 1.

It is easy to see that the integers generated are in some sense random, but the value of a member of the sequence is strongly influenced by the value of the previous member. If the current number is 5, then the next number is 5 or 6, no chance of getting a 7, even though it is very likely we will eventually get a 7. Here is a more complex example.

Example A.11 Suppose we play a game, called *Hexer*, with 4 dice as follows. Start by rolling all 4 dice. If you get no 6s, you lose. Otherwise, put the 6s aside in a “six pool” and reroll the remaining dice. Each time you make a roll that produces no 6s, you pick up a die from the six pool to be used in the next roll. If you roll no 6s with an empty six pool, you lose. When all the dice are in the six pool, you win. In all other cases, play continues.

Hexer is a Markov chain with states $\{s_0, s_1, s_2, s_3, s_4, L\}$ corresponding to losing or the number of dice in the six pool. Attaining state s_4 indicates a win. The initial distribution is to be in state s_0 with probability 1. The transition probabilities are summarized in the transition matrix.

		s					
		s_0	s_1	s_2	s_3	s_4	L
t	s_0	0	0.3858	0.1157	0.0154	0.0008	0.4823
	s_1	0.5787	0	0.3472	0.0694	0.0046	0
	s_2	0	0.6944	0	0.2778	0.0278	0
	s_3	0	0	0.8333	0	0.1666	0
	s_4	0	0	0	0	1	0
	L	0	0	0	0	0	1

Hexer transition matrix

A *transition matrix* for a Markov chain is a matrix $[a_{i,j}]$ indexed by the states of the Markov chain with $a_{i,j} = p_j(i)$.

Example A.11 gives conditions for the game *Hexer* to end. The *terminal states* in which the games ends are s_4 and L . The definition of Markov chain we are using doesn't have a notion of terminal states, so we simply assign such states a probability of 1 of following themselves in the chain and then explain separately whether a state ends the chain or is repeated indefinitely whenever we reach it. The name for such states in Markov chain theory is *absorbing states*.

If we have a Markov chain M with states S , then a subset A of S is said to be *closed* if every state that can follow a state in A is a state in A . Examples of closed subsets of the state space of *Hexer* are $\{L\}$, $\{s_4\}$, or the entire set of states.

If S does not contain two disjoint closed subsets, we say M is *indecomposable*. If for two states $x, y \in S$ it is possible for x to follow y and for y to follow x in the chain, then we say that x and y communicate. A subset A of S is a *communicating class of states*, if any two

states in A communicate. The set $\{s_0, s_1, s_2, s_3\}$ is a communicating class in the Markov chain for Hexer from Example A.11.

If there is a distribution d on the states such that for any initial distribution the limiting probabilities of being in each of the states converges to d , then we say that M is *stable* and we call d the *limiting distribution*. (The limiting probability of a state is just the limit as the number of steps goes to infinity of the number of times you've been in the state divided by the number of steps you've run the Markov chain.)

Notice that for Hexer there are two different “final” distributions as the number of steps go to infinity: probability 1 of being in state L and probability 1 of being in state s_4 . So, the Hexer Markov chain is not stable.

A *stable initial state* is a distribution d such that if you start with the distribution d on the states, you keep that distribution. If M is the transition matrix of a Markov chain and \vec{d} is the row vector of probabilities in d , then d is a stable initial distribution if

$$\vec{d} \cdot M = \vec{d}.$$

It is not hard to show that the limiting distribution of a Markov chain, if it exists, is also a stable initial distribution. The following theorems, offered without proof, go a long way toward characterizing a very nice class of Markov chains.

Theorem A.4 *An indecomposable Markov chain has at most one stable initial distribution.*

Theorem A.5 *Stable Markov chains are indecomposable.*

If there is a partition of the set of states of a Markov chain

$$\{A_0, A_2, \dots, A_{k-1}\}, k \geq 2$$

such that the only states that can follow the states in A_i are the states in A_{i+1} (addition *(mod k)*), then we say that a Markov chain is *periodic with period k*. The largest k for which a Markov chain is periodic is called the *period* of the Markov chain, and if there is no $k \geq 2$ for which a Markov chain is periodic, then we call the Markov chain *aperiodic*.

Theorem A.6 *If a Markov chain is indecomposable, aperiodic, and has states that constitute a single communicating class, then either*

- (i) *The Markov chain has no limiting distribution and the limiting probabilities of each state are zero, or,*
- (ii) *The Markov chain has a limiting distribution and is stable.*

The next two examples are Markov chains that fit (i) and (ii) of Theorem A.6 respectively.

Example A.12 Suppose we modify Example A.10 as follows. Roll 6-sided dice instead of flipping coins. Add 1 for a 5 or a 6, subtract 1 for a 1 or a 2, and otherwise leave the number unchanged. The states are now $S = \{\dots, -2, -1, 0, 1, 2, \dots\}$ and the transition probabilities become

$$p_s(t) = \begin{cases} 1/3 & s=t-1, t, \text{ or } t+1 \\ 0 & \text{otherwise} \end{cases}$$

It is not hard to see there is a single closed set of states, the whole state space, and that every state communicates with every other state. A bit of thought also shows that this Markov chain is aperiodic. This implies that Theorem A.6 applies. Since we could choose our initial distribution to have probability one on any state, it follows that each state must have the same limiting probability as each other state. As you cannot divide 1 into infinitely many equal pieces, there cannot be a limiting distribution, and, so, we are in case (i) of the theorem.

Example A.13 Suppose we have a 4-state Markov chain with states $\{a, b, c, d\}$ and transition matrix

$$\begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}.$$

It is obvious from inspection that this Markov chain satisfies the hypothesis of Theorem A.6. Since there are finitely many states, the limiting probability cannot be zero, and, so, this chain is of the type described by (ii). It is in fact easy to see that the limiting distribution is $(0.25, 0.25, 0.25, 0.25)$.

It isn't hard to approximate the stationary distribution of a Markov chain with a few states, *if you know it has one*. Suppose M is a Markov chain with n states and transition matrix T . Pick an initial distribution d and then compute the sequence $\{\vec{d}, \vec{d} \cdot T, \vec{d} \cdot T^2, \dots\}$. If M is stable, this sequence will converge to the stationary distribution. Essentially, repeated multiplication by the transition matrix will turn any initial distribution into the stationary distribution in the limit. For many choices of d (but not all), the sequence obtained by repeated multiplication by T will exhibit approximate periodicity, if M is periodic.

Let us conclude with a simple Markov chain example that solves an estimation problem in artificial life. While reading this example, keep in mind there are assumptions and estimates involved; do not accept these blindly. Any assumption, no matter how much you need it to cut down the problem to manageable size, should be repeatedly examined. With that caveat, let us proceed.

Example A.14 Suppose we are running a string evolver on the alphabet $\{0, 1\}$ that uses an evolutionary algorithm with tournament selection and tournament size 2. If we have 60 creatures of length 20 and use single point mutation, what is the expected time-to-solution?

Answer:

Assume the point mutation must change the value of the locus it mutates. Also, assume the reference string is “11111111111111111111.” (Problem 1.9 showed that the choice of reference string is irrelevant to the solution time. This choice let’s us use the results of Example A.4.) With this reference string, the creature’s fitness is the number of 1s in its gene.

The first step in solving this problem is to figure out how good the best creature in the population is. (If, for example, we had 2^{20} creatures, there would be an excellent chance the solution would exist in the initial random population.) We solved this problem in Example A.4; the answer is that the best creature has an expected fitness of 14.

The model of evolution (tournament selection) breaks the population into randomly selected sets of two creatures, copies the better over the worse in each group, and then performs a (bit flip) point mutation on the copy. This means that all creatures are following the same path to the reference string at the same rate. (Imagine how hard this example would be if we allowed crossover.) We, therefore, assume that the time-to-solution can be computed by following the best creature.

Let M be the Markov chain whose states are $\{0, 1, \dots, 20\}$ representing the fitness of the best creature. The model of evolution ensures that the best creature will survive and that improvement always comes in the form of a single 0 being transformed into a 1. From this we can compute the transition probabilities to be

$$p_s(t) = \begin{cases} (20-t)/20 & s=t+1 \\ t/20 & s=t \\ 0 & \text{otherwise} \end{cases}$$

Our current guess at an initial distribution is

$$(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0)$$

(that is, the best creature has fitness 14). Our expected time, in generations, to improve the best creature would be the reciprocal of the probability he will improve (why?). Summing over the needed improvements, this gives us an estimate of

$$\sum_{t=14}^{19} \frac{20}{20-t} = 44$$

generations.

We actually have the information needed to build the transition matrix, and the true initial distribution of the population is available; it is $\vec{d} = (p_0, p_1, \dots, p_{20})$ where

$$p_i = \binom{20}{i} \left(\frac{1}{2}\right)^{20}.$$

We could get a much better estimate of time-to-solution by taking the true initial distribution and multiplying it by the transition matrix (with a computer) until the generation in which the probability of beginning in state 20 is at least $1/60$. Keep in mind that, instead of following the best creature, we are now tracking the whole population of 60 creatures - so $1/60$ of a chance of being in state 20 gives us an expectation of one creature in state 20.

If we do this, our estimate becomes 21 generations, about half of the far cruder and easier estimate above. In any case, an estimate of this sort should never be seen as a precise answer, but, rather, as a ballpark figure that tells you if your simulation needs to be run for a few minutes, an hour, overnight, or on a generation of hardware not available until slightly after the FAA certification of a craft capable of interstellar travel.