Entropy Rates:
Some Definitions, Facts, and Examples

Mathias Winther Madsen

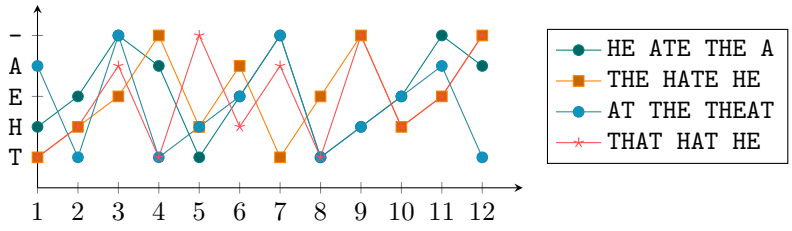
November 20, 2015

Contents

1 Definitions

1.1 Random processes

- A **word** is a finite string $x = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$.
- A **sample path** is an infinite sequence $x = (x_1, x_2, x_3, x_4, \ldots) \in \mathcal{X}^{\mathbb{N}}$.
- A **random process** is a probability distribution over sample paths.
- A random process P defines a time-indexed **family of random variables**, $X_1, X_2, X_3, \ldots$, whose values at x are the coordinates $x_1, x_2, x_3, \ldots$.
- By the **Daniel-Kolmogorov extension theorem**, a random process is uniquely defined by its word probabilities. In other words, if you know the joint distribution of any finite sub-family $X_n, X_{n+1}, \ldots, X_{n+k}$, then you know the entire distribution.



1.2 Entropy rates

- The **entropy rate** of a random process is

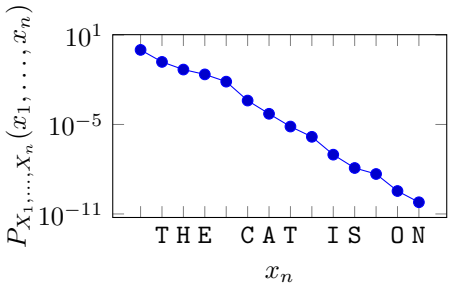
$$\lim_{n\to\infty} \frac{H(X_1, X_2, X_3, \ldots, X_n)}{n}$$

  when this limit exists.

- By the chain rule (or product rule) of entropy, this is equal to

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} H(X_{i+1} \mid X_1, X_2, \ldots, X_i),$$

  which is often easier to compute.



- Entropy rates are measured in **bits per symbol**. When the symbols are emitted regularly (e.g., 1000 symbols per second), this is proportional to an amount of **bits per time unit** (e.g., bits per millisecond).

1.3 Markov Chains

- A random process is **i.i.d.** if

$$P(X_{n+1} \,|\, X_1, X_2, \ldots, X_n) \;=\; P(X_{n+1}).$$

- A random process is a **Markov chain** if

$$P(X_{n+1} \,|\, X_1, X_2, \ldots, X_n) \;=\; P(X_{n+1} \,|\, X_n).$$

  The conditional probabilities $P(X_{n+1} \,|\, X_n)$ of a Markov chain are called its **transition probabilities**.

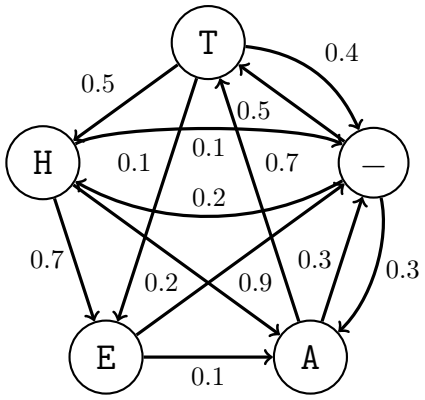
- A system of transition probabilities is consistent with several different Markov chains, but if an **initial condition** is also supplied in the form of a marginal distribution for X_1, the transitional probabilities define a unique Markov chain.

- A Markov chain is **stationary** if all its marginal distributions are identical:

$$P(X_1) \;=\; P(X_2) \;=\; P(X_3) \;=\; P(X_4) \;=\; \cdots$$

- A system of transition probabilities defines a family of Markov chains. If only one member of this family is stationary, then its marginal distributions express how much time, on average, the members of this family will spend in various states. The are the **limiting relative visiting times**.

- The entropy rate of a stationary Markov chain is the **weighted average of the conditional entropies** at each state.



1.4 Stationary random processes

- A random process is **stationary** if the probability of a word $(x_1, x_2, \ldots, x_n)$ is independent of where in the sample path we are:

$$P(X_1 = x_1, \ldots, X_n = x_n) \;=\; P(X_2 = x_1, \ldots, X_{n+1} = x_n)$$

- The n**th time-average** of a function $f : \mathcal{X}^{\mathbb{N}} \to \mathbb{R}$ along the sample path $x = x_1, x_2, x_3, \ldots$ is

$$A_n f(x) \;=\; \frac{f(x_1, x_2, \ldots) + f(x_2, x_3, \ldots) + \cdots + f(x_n, x_{n+1}, \ldots)}{n}.$$

- The **limiting time-average** of f along x is

$$\lim_{n\to\infty} A_n f(x)$$

  when this limit exist.

- By the **first part of the ergodic theorem**, the time-averages of a stationary process always converge (although possibly to $\pm\infty$).

- Reversely, there is (by the extension theorem) **at most one stationary process** that produces a certain system of limiting time-averages.

- When an ergodic process has convergent time-averages, these limiting time-averages coincide with the limiting time-averages of one and only one stationary random process P^*. We say that P^* **describes the limit behavior of** P.

- If P^* and P are random processes such that

  - P^* is stationary and ergodic;
  - $P^*(B) = 0$ implies that $P(B) = 0$;

  then P^* describes the limit behavior of P.

1.5 Ergodic random processes

- A set of sample paths set is **time-invariant** if membership of that set is decided by the tail of the sample path rather than any initial segment:

$$(x_1, x_2, x_3, \ldots) \in B \qquad \Longrightarrow \qquad (x_2, x_3, x_4, \ldots) \in B.$$

- A random process P is **ergodic** if $P(B) = 1$ or $P(B) = 0$ for all time-invariant sets B.

- By the **second part of the ergodic theorem**, the limiting time-averages of an ergodic process are deterministic random variables (if they exist).

- A non-ergodic process is a **(non-trivial) mixture of ergodic processes**.
- The probability distribution over possible limit behaviors of such a mixture is described by a **mixture of stationary and ergodic processes**.
- In general, $\lim_n A_n f(X)$ is a random variable with **one value on each mixture component**. The probabilities of these values are given by the mixture proportions.

1.6 The Shannon-McMillan-Breiman theorem

- A stationary distribution has an entropy rate (possibly $+\infty$).
- On a sample path drawn from a stationary and ergodic distribution, the **average surprisal converges to the entropy rate** with probability 1.
- For large enough n, a stationary and ergodic process with an entropy rate of H has about 2^{nH} typical sequences of length n.

2 Examples

In each of the following examples, spaces have no meaning, but are only included for readability. When spaces are counted as encodable source symbols, they are shown as underscores (_).

2.1 An i.i.d. Process

Let $X_1, X_2, X_3, \ldots$ be independent and identically distributed random variables. Then $X_1, X_2, X_3, \ldots$ defines a random process which may, for instance, have samples paths like

`ITTTSSTLCTEC_EFAIRNPEIAI_SARH_FM...`

This process is stationary and ergodic.

If each variable has an entropy of H, then the entropy rate of the process is

$$\lim_{n\to\infty} \frac{H + H + \cdots + H}{n} \;=\; H.$$

2.2 Finite Repetition

Suppose we repeatedly pick a letter and print it three times:

`LLL EEE HHH QQQ MMM QQQ OOO TTT EEE YYY XXX GGG ...`

This random process is ergodic but not stationary: the probability of encountering the word `AB` is 26^{-2} at positions $3, 6, 9, 12, \ldots$, but 0 elsewhere.

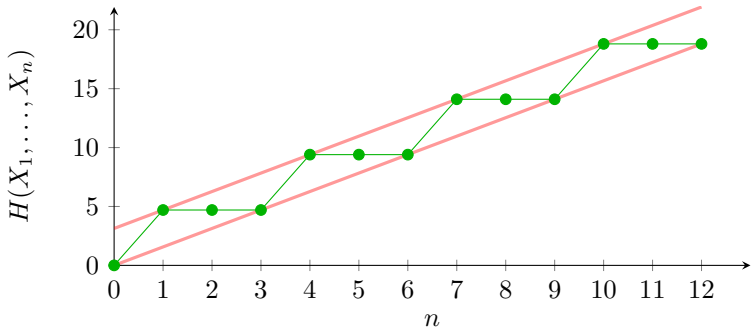
The entropy of its nth initial segment is

$$H(X_1, X_2, \ldots, X_n) \;=\; \log 26 + 0 + 0 + \log 26 + 0 + 0 + \cdots + \log 26 + 0 + 0,$$

which is sandwiched between the two bounds

$$\frac{\log 26}{3} n \quad \leq \quad H(X_1, X_2, \ldots, X_n) \quad \leq \quad \frac{\log 26}{3} n + \frac{2 \log 26}{3}.$$

When didvided by n, both of these bounds converge to $\frac{1}{3} \log 26$. This is thus the entropy rate of the process.



The limit behavior of the process is described by a stationary process which is an equal mixture of three components:

- The random process P itself, $X_1, X_2, X_3, \ldots$;
- The time-shifted distribution TP which describes the random process $X_2, X_3, X_4, \ldots$;
- The doubly time-shifted distribution T^2P which describes the random process $X_3, X_4, X_5, \ldots$.

The mixture

$$P^* \;=\; \frac{1}{3}P + \frac{1}{3}TP + \frac{1}{3}T^2P$$

is a stationary but not ergodic distribution.

2.3 Random Repetition

We repeatedly pick a letter and print it $k \sim \text{Geometric}(\frac{1}{2})$ times:

SSS P MMMMM D HHH K Z T D U C AAA I D TTT Y HHHH ...

This process is a stationary and ergodic Markov chain. Its transition probabilities are

$$P(X_{n+1} = x_{n+1} \,|\, X_n = x_n) \quad = \quad \begin{cases} (1/2) + (1/2)(1/26) & (x_{n+1} = x_n) \\ (1/2)(1/26) & (x_{n+1} \neq x_n) \end{cases}$$

The surprisal value corresponding to a repetition of the preceding letter is

$$\log\left(\frac{1}{2}+\frac{1}{2}\cdot\frac{1}{26}\right) \quad = \quad \log\frac{27}{52} \quad = \quad 0.95.$$

The surprisal value that corresponds to the choice of a new letter is

$$\log\left(\frac{1}{2}\cdot\frac{1}{26}\right) \quad = \quad \log\frac{1}{52} \quad = \quad 5.70.$$

These two surprisal values are the same in all states. The conditional entropy of the next letter given the last one is therefore always

$$H(X_{n+1}\,|\,X_1, X_2, \ldots, X_n) \quad = \quad \frac{27}{52}\log\frac{52}{27} + 25\cdot\frac{1}{52}\log\frac{52}{1} \quad \approx \quad 3.23.$$

Since this conditional entropy is the same in all states, it is also equal to the entropy rate of the process.

2.4 Eternal Repetition

Pick a letter at random and print it forever:

GGGGGGGGGGGGGGGGGGGGGGGGGGGGGG ...

This process is stationary, but not ergodic. It can be defined as the infinite extension of the word probabilities

$$P(X_1 = x_1, \ldots, X_n = x_n) \quad = \quad \begin{cases} 1/26 & (x_1 = x_2 = \cdots = x_n) \\ 0 & (\text{otherwise}) \end{cases}$$

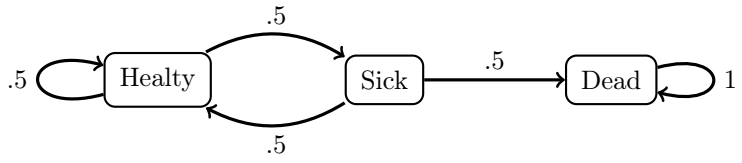
When decomposed into an ergodic mixture, this random process has 26 components, all of which are deterministic (and therefore stationary and ergodic):

$$P \;=\; \frac{1}{26}P_{\mathtt{A}} + \frac{1}{26}P_{\mathtt{B}} + \frac{1}{26}P_{\mathtt{C}} \cdots + \frac{1}{26}P_{\mathtt{Z}}.$$

Since each of these components have an entropy rate of 0, so does P. In fact, the actual cost of encoding a sample of length n from this process is $\lceil \log 26 \rceil / n$ bits per character.

2.5 Healthy-Sick-Dead

The following transition probabilities define a family of Markov chains:



We can find the stationary marginal distribution over these three states by solving the following system of linear equations:

$$\begin{aligned} h &= 0.5h + 0.5s \\ s &= 0.5h \\ d &= 0.5s + d \\ h + s + d &= 1 \end{aligned}$$

This has the unique solution $(h^*, s^*, d^*) = (0, 0, 1)$. Since this is the only solution, every Markov chain in this family is ergodic, and all of them visit the three states limiting frequencies $(h^*, s^*, d^*) = (0, 0, 1)$.

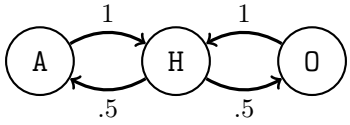
In order to compute the entropy rate, we compute the conditional entropy

$$H(X_{n+1} \,|\, X_n = \text{Dead}) \quad = \quad 0.$$

This is the only state that will recur in the long run, so the entropy rate of this process is $1 \cdot 0 = 0$.

2.6 The Santa Claus Machine

The following transition probabilities define a family of Markov processes:



To find the stationary Markov chain in this family, we solve the equations

$$\begin{aligned} a &= 0.5h \\ h &= a + o \\ o &= 0.5h \\ a + h + o &= 1 \end{aligned}$$

This has the unique solution $(a^*, h^*, o^*) = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. All Markov chains in this family are thus ergodic, and their limiting visiting frequencies are described by this stationary marginal distribution.

To compute the entropy rate, we note that

$$\begin{aligned} H(X_{n+1} \,|\, X_n = \texttt{A}) &= 0 \\ H(X_{n+1} \,|\, X_n = \texttt{H}) &= 1 \\ H(X_{n+1} \,|\, X_n = \texttt{O}) &= 0 \end{aligned}$$

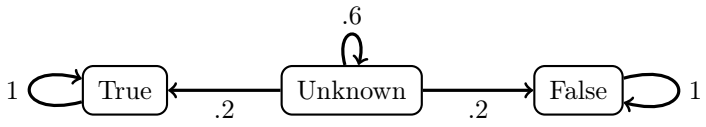
Taking the weighted average of these conditional entropies, we get

$$\frac{1}{4} \cdot 0 \; + \; \frac{1}{2} \cdot 1 \; + \; \frac{1}{4} \cdot 0 \quad = \quad \frac{1}{2}.$$

The entropy rate of this random process is therefore 1/2. This reflects the fact that we need one bit to encode the choice of `A` or `O` every other symbol.

2.7 The Decision Chain

The following transition probabilities define a family of Markov chains:



In order to find the stationary distributions in this family, we solve

$$\begin{aligned} t &= t + 0.2u \\ u &= 0.6u \\ f &= 0.2u + f \\ t + u + f &= 1 \end{aligned}$$

This problem has infinitely many solutions. They can be parametrized as

$$(t^*, u^*, f^*) = (p, 0, 1-p),$$

where p can be chosen freely from the unit interval, $p \in [0, 1]$.

Except for the two extreme cases $p = 0$ and $p = 1$, none of the Markov chains in this family are ergodic. A sample path from a Markov chain in this family can exhibit substantially different limit-behaviors depending on which trapping set it ends up in.

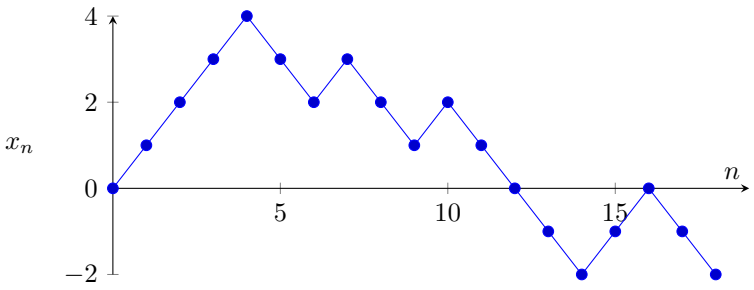
As it happens, the conditional entropy at both trapping sets is 0. The entropy rate of every process in this family is therefore 0. Note, however, that in other cases, the entropy rate might not have reflect actual average surprisals.

2.8 Random Walk

A dust particle starts at $X_1 = 0$ and then takes a unit step, either up or down, in each time period:

$$0, 1, 2, 3, 4, 3, 2, 3, 2, 1, 2, 1, 0, -1, -2, -1, 0, -1, -2, \ldots$$

This random process is a Markov chain with a countably infinite number of states. It is ergodic, but not stationary.



The system of limiting time-averages defined by this process does not define a probability distribution. This is because the marginal probability of finding the dust particle in any fixed set $A \subseteq \mathbb{Z}$ goes to zero as $n \to \infty$. The only measure consistent with these visiting frequencies is therefore the all-zero measure that sets $P(A) = 0$ for all A. Alhtough this measure is, in a certain sense, the correct expression of the limiting behavior of this random process, it is not a probability distribution.

2.9 Exponential Means

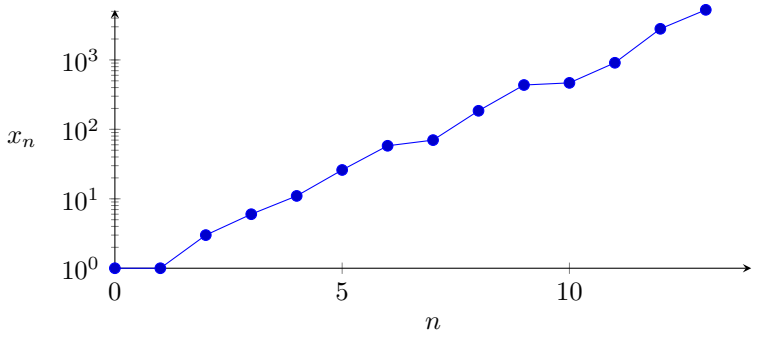
Define the random variables $X_1, X_2, X_3, \ldots$:

$$\begin{array}{rcl} X_1 & \sim & \text{Uniform}\{1,2\} \\ X_2 & \sim & \text{Uniform}\{1,2,3,4\} \\ X_3 & \sim & \text{Uniform}\{1,2,3,4,5,6,7,8\} \\ X_4 & \sim & \text{Uniform}\{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16\} \\ \vdots & \vdots & \qquad\qquad\qquad\qquad \ddots \end{array}$$

This list of random variables defines a random process. A typical sample from this process is

$$1, 1, 3, 6, 11, 26, 58, 70, 185, 435, 467, 909, 2804, 5262, \ldots$$

Note that the coordinates $X_1, X_2, X_3, \ldots$ of this random process are independent, but not identically distributed.



The process is highly non-stationary. This is apparent from the exponential growth in the marginal expectations, $E[X_n]$.

The marginal entropies $H(X_n) = n$ grow linearly, so the cumulative entropy grows roughly quadratically in n:

$$H(X_1, X_2, \ldots, X_n) \quad = \quad H(X_1) + H(X_2) + \cdots + H(X_n) \quad \approx \quad \frac{1}{2}n^2.$$

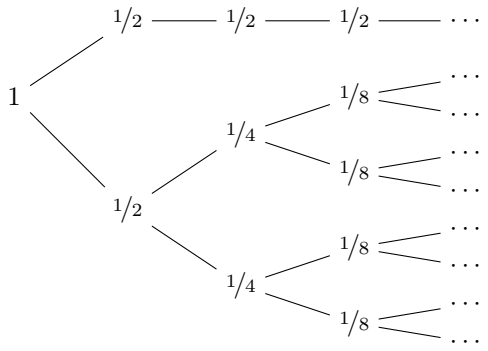
The entropy rate of this process is therefore

$$\lim_{n\to\infty} \frac{n^2/2}{n} \quad = \quad \lim_{n\to\infty} \frac{1}{2}n \quad = \quad \infty.$$

This reflects the fact that no fixed amount of bits per symbol will be sufficient to encode the sequence $X_1, X_2, \ldots, X_n$ for all n.

2.10 Half-Deterministic

The following tree defines a binary random process P:



This random process is stationary, but not ergodic. It can be decomposed into two stationary and ergodic mixture components:

1. a deterministic branch, $P_1 = \text{Bernoulli}(0)$, with entropy rate 0;
2. an i.d.d. subtree, $P_2 = \text{Bernoulli}(\frac{1}{2})$, with entropy rate 1.

We then have

$$P \quad = \quad \frac{1}{2}P_1 + \frac{1}{2}P_2.$$

The entropy rate of P is $H = \frac{1}{2}$. However, the process is not ergodic, and its entropy rate deviates from the average surprisal on all sample paths. In this case, the entropy thus has no interpretation in terms of data compression.

2.11 Nested, Paired Parentheses

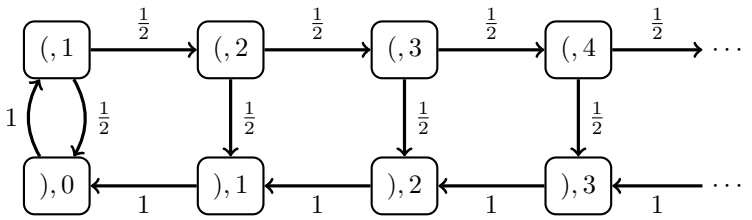
A process P repeatedly prints $k \sim \text{Geometric}(\frac{1}{2})$ opening parentheses and then immediately closes them again:

$$()((())))(((())))(())(())(((()))))(()) \ldots$$

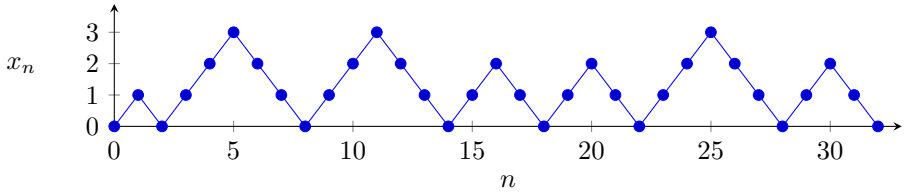
This process is ergodic. It is not stationary. P is also not a Markov chain, since checking whether all parentheses have been closed may require one to look arbitrarily far back.

However, by introducing a hidden "depth" variable that keeps track of the current level of parenthesis nesting, we can analyze P as a so-called "hidden" Markov model. This amounts to packing all the memory that P needs in order to make its next choice into a single variable.

This trick allows us to convert this non-Markovian process into a Markov chain which transitions between various memory states. In the present case, the space of hidden memory states and their accessibility relation is given by the following transition diagram:

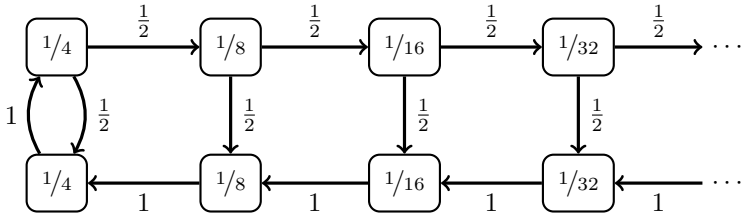


We can also think of the state of this process as a walk up and down in symmetric triangles (which then have geometrically distributed heights):



According to its definition, P starts deterministically in the bottom left state of the transition diagram, about to open its first parentheses. Over time, however, this concentration of probability mass will diffuse out through the state space. In the long run, the process will visit each state with a frequency given by the relevant stationary distribution.

This stationary marginal distribution is shown in the following diagram:



In every states in the top row of this diagram, there are two possibilities for the next symbol, and the conditional entropy is therefore 1 bit. In the bottom row, there is only one possible next symbol, and the conditional entropy is 0. The

entropy rate is therefore the weighted average

$$\begin{aligned} H &= 1 \cdot (1/4 + 1/8 + 1/16 + \cdots) + \\ &\quad 0 \cdot (1/4 + 1/8 + 1/16 + \cdots) \\ &= 1/2. \end{aligned}$$

The transition diagram with the stationary state probabilities can also be used to compute the probabilities of words according to P^*.

For instance, the word "))((" has conditional probability zero at all states except the third state in the bottom row, (")", 2). At this state, the conditional probability of observing the sequence "))((" is

$$1 \cdot 1 \cdot 1 \cdot \frac{1}{2} \;=\; \frac{1}{2}.$$

Since the stationary probability of the state (")", 2) is 1/16, it follows that P^* assigns probability 1/32 to the event that the word "))((" occurs at some given point in the sample path.

Note that the monogram "(" is emitted with probability 1 from the bottom left state, and with probability 1/2 from all states in the top row. Its marginal probability is therefore

$$\frac{1}{4} \cdot 1 \;+\; \frac{1}{4} \cdot \frac{1}{2} \;+\; \frac{1}{8} \cdot \frac{1}{2} \;+\; \frac{1}{16} \cdot \frac{1}{2} \cdots \quad = \quad \frac{1}{2}.$$

Perhaps not surprisingly, P^* thus assigns probability 1/2 to the event that a randomly plucked symbol from the sample path $x_1, x_2, x_3, \ldots$ is an opening parenthesis. Unlike the case for P, this is true even for the first symbol, X_1.

2.12 The Beta Urn Scheme

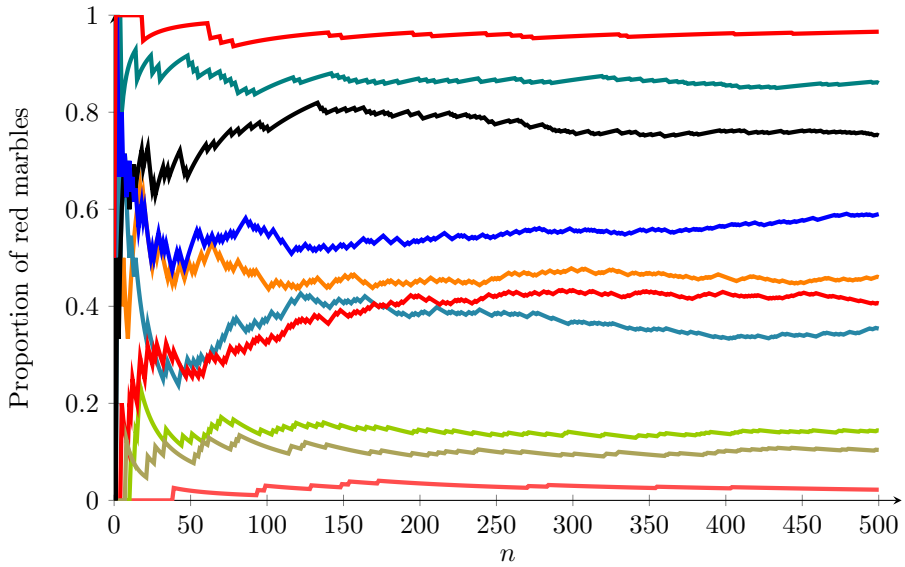
Suppose we start with an urn that contains two marbles, one red and one blue. We then repeatedly draw a marble from the urn and replace it by *two* marbles of the same color.

We define a random process $X_1, X_2, X_3, \ldots$ by recording the color of the nth marble we draw as X_n. A typical sample path from this process is then

RBRBRBRRBRBBRRRRBRBBBRRRB ...

The urn thus grows by one marble per round, and if it acquires a strong bias towards either red or blue, this bias tends to sustain itself.

However, this self-sustaining property does *not* entail that the urn will necessarily massively overrepresenting one or the other color. In fact, the limiting proportion of red marbles in the urn is a random variable which uniformly distributed over the unit interval:



The possible limit behaviors of P are therefore described by the flip flipping processes $P_\theta = \text{Bernoulli}(\theta)$ with fixed coin bias $\theta \in [0, 1]$. These processes correspond exactly to the kind of behavior you would get from an infinitely large urn with a specific proportion of red marbles.

Since Bernoulli processes are i.i.d., they are also stationary and ergodic. Each one of them defines one possible attractor for P. In fact, P can be decomposed into an overcountable mixture of Bernoulli processes:

$$P \quad = \quad \int P_\theta \, d\theta.$$

Since each Bernoulli process is stationary, this also proves that P itself is stationary, although this is by no means obvious.