

# Random drift on probability distributions

Lu Zhang Gram

August 3, 2007

## 1 Overview

This project investigates approaches to modelling random drift on a discrete probability distribution. Given such models along with methods for performing inference, neutral mutation models could be developed for a variety of stochastic grammars (regular, context-free, etc.). These could in turn be used to compute distance between grammars or to perform alignment.

## 2 Stochastic processes

This section provides a brief introduction to *stochastic processes* which are the main focus of this project. For further material consult Karlin & Taylor (1975).

We define a *stochastic process* to be a family  $\{X_t : t \in T\}$  of random variables  $X_t$  taking values in  $S$ .  $S$  is called the *state space* of our process. If  $T = \mathbb{N}$ , we call our process a *discrete-time* process. If  $T = [0, \text{inf})$ , our process is a *continuous-time* process.

In this project,  $S$  will always be  $S_n = \{v \in \mathbb{R}^{n+1} : v_i \in [0, 1] \text{ and } \sum_i v_i = 1\}$ , called the *n-dimensional simplex*.  $S_n$  specifies the region of  $\mathbb{R}^{n+1}$  where the vectors represent discrete probability distributions on  $\{1, \dots, n + 1\}$  with probability mass function  $p(i) = v_i$ . For example,  $S_1$  is a line segment in  $\mathbb{R}^2$ ,  $S_2$  an equilateral triangle in  $\mathbb{R}^3$  and  $S_3$  a tetrahedron in  $\mathbb{R}^4$ .

Now given a stochastic process  $\{X_t\}$  on  $S_n$ , the distribution for  $X_t$  given  $X_s$  where  $s < t$  has a probability density function  $f(X_t|X_s)$ . For *stationary* processes,  $f(X_t|X_s) = f(X_{t-s}|X_0)$  for all  $s < t$ . We call  $d(x|y, t) = f(X_t = x|X_0 = y)$  the *transition density*. For a discrete-time process,  $d(x|y, t)$  gives the probability of transiting from  $x$  to  $y$  in time  $t$ .

We assume our processes are *Markov*, i.e. given  $X_t, \{X_s\}$  for  $s > t$  are independent of  $\{X_u\}$  for  $u < t$ . Stationary Markov processes have independent time increments, i.e.  $\{X_{t_{i+1}} - X_{t_i} : i \in \mathbb{N}\}$  are independent for all  $t_i$  s.t.  $t_1 < t_2 < t_3 < \dots$

### 3 Aim

The aim of this project is to develop a stochastic process  $\{p_t\}$  on  $S_n$ . The process should be simple enough that we can perform simulation and preferably inference, whilst being realistic enough to be incorporated into a neutral mutation model. It should satisfy the following properties:

1. Each  $p_t$  is a discrete probability mass function on  $\{1, 2, 3, \dots, n\}$ .
2. Stationarity:  $f(p_t|p_s) = f(v_{t-s}|v_0) = d(p_t|p_s, t-s)$  for all  $s < t$ .
3. Symmetry: If  $\pi$  is a permutation on  $\{1, \dots, n\}$  and  $p_\pi(i) = p(\pi(i))$  then  $d$  satisfies  $d(p|q, t) = d(p_\pi|q_\pi, t)$ .

We prefer the boundaries to be neither absorbing nor reflecting, since absorbing boundaries would force  $p_t(i)$  to stay at 0 once it reaches 0 while with reflecting boundaries the time spent at the boundaries would have measure zero (almost surely), so  $p_t(i)$  would 'almost never' be 0. Instead, we would like there to be a positive probability of reaching the boundary and staying there for a positive amount of time. We refer to this as *globby boundaries* (the term 'sticky boundary' is already used for other types of boundary).

### 4 Normalizing approach

The obvious approach is to let the  $q_t(i)$  evolve independently as Brownian motion on  $\mathbb{R}^+$  with a reflecting barrier at 0 and setting  $p_t$  equal to  $q_t$  normalized (making sure to avoid  $q_t$  being the zero function). However, this process is non-stationary: Suppose we start at  $q = (0, 0, \dots, 0, 1)$ . Since changes in  $q_t(i)$  would have a smaller effect on  $p_t(i)$  when  $q_t(i)$  is large due to the large normalizing factor and  $q_t$  is likely to increase its distance to 0 with time, the variance of  $p$  is likely to decrease over time.

### 5 Simplex model

A direct approach is to model the process as Brownian motion on  $S_n$ . For lack of a better name, I shall refer to this as the *simplex model*. In  $S_1$  we can ensure globby boundaries this way: Use a coordinate system with the origin at  $\mathbf{O} = (\frac{1}{2}, \frac{1}{2})$  and basis vector  $\mathbf{e} = (\frac{1}{2}, -\frac{1}{2})$ .

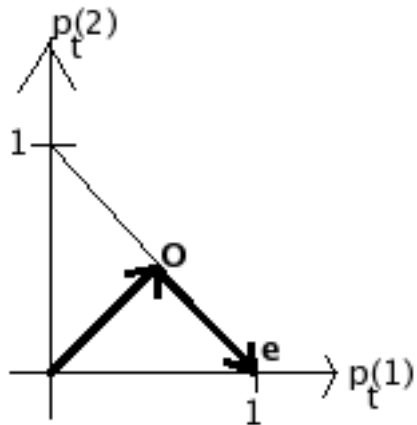


Fig. 1. The coordinate system with center  $O$  and basis vector  $e$

In this coordinate system, let  $\{x_t\}$  be ordinary Brownian motion on  $\mathbb{R}$ . Divide  $\mathbb{R}$  into regions  $D_k = [-1 + 2k, 1 + 2k]$  for  $k$  in  $\mathbb{Z}$ . These regions arise when we reflect  $[-1, 1]$  repeatedly. Thus we can obtain a mapping,  $h$  say, that maps  $\mathbb{R}$  to  $[-1, 1]$  by mapping every point in  $[-1 + 2k, 1 + 2k]$  to the corresponding point in  $[-1, 1]$ . We can then construct a 1-dimensional Brownian motion with reflecting boundaries at  $+1$  and  $-1$ ,  $\{y_t\}$  by letting  $y_t = h(x_t)$ :

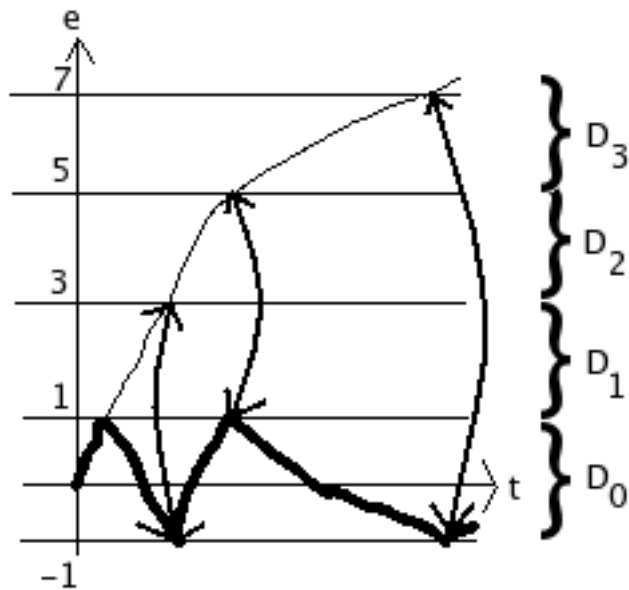


Fig. 2.  $\{x_t\}$  and  $\{y_t\}$ ,  $\{y_t\}$  indicated in bold, arrows point out corresponding points.

$h$  is algebraically defined by

$$h(x) = \begin{cases} g(x) & \text{if } x > 1 \\ x & \text{if } -1 \leq x \leq 1 \\ -g(|x|) & \text{if } x < -1 \end{cases}$$

where  $g(x) = (x - 2\lfloor \frac{x+1}{2} \rfloor)(-1)^{\lfloor \frac{x+1}{2} \rfloor}$ .

Now, we scale  $x_t$  up to reflecting Brownian motion between  $-c$  and  $+c$  for some constant  $c > 1$  by letting  $z_t = cx_t$ . Truncate the parts of our Brownian motion that move outside  $+1$  and  $-1$  using  $f$  defined by:

$$f(x) = \begin{cases} 1 & \text{if } x > 1 \\ x & \text{if } -1 \leq x \leq 1 \\ -1 & \text{if } x < -1 \end{cases}$$

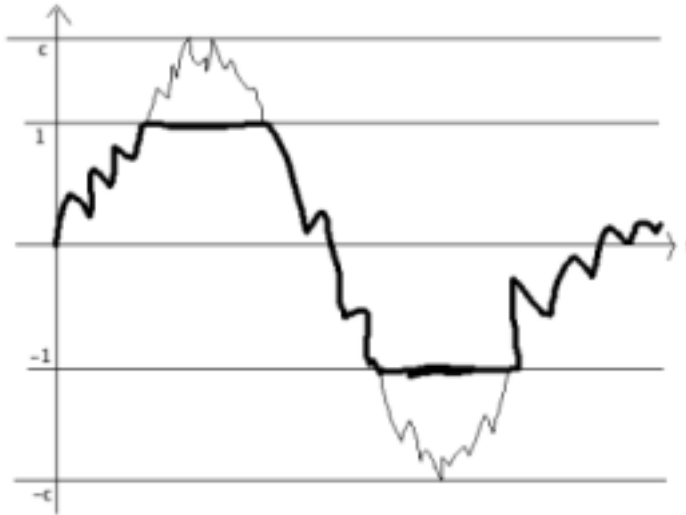


Fig. 3. The processes  $\{f(z_t)\}$  and  $\{z_t\}$ .  $\{f(z_t)\}$  is marked in bold

Finally convert back to probability mass functions by letting our final process be  $\{\mathbf{O} + f(z_t)\mathbf{e}\}$ .

This generalizes nicely to  $S_2$ , since  $S_2$  forms an equilateral triangle and we can tile  $\mathbb{R}^2$  by repeatedly reflecting equilateral triangles in their sides. Thus we can obtain a Brownian motion in  $\mathbb{R}^2$  with reflecting boundaries at the sides of the triangle,  $\{y_t\}$ , using ordinary Brownian motion in  $\mathbb{R}^2$ ,  $\{x_t\}$ :

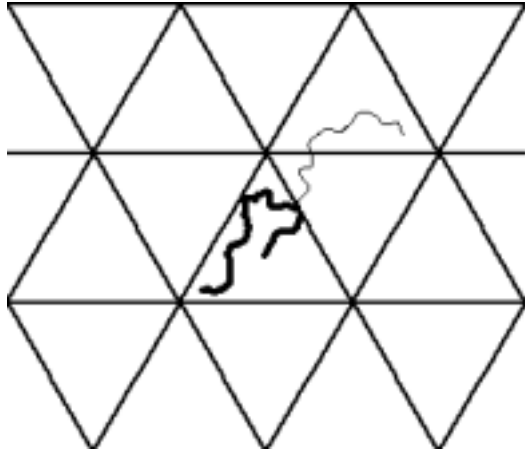


Fig. 4. Ordinary and reflecting Brownian motion,  $\{x_t\}$  and  $\{y_t\}$  respectively.  $\{y_t\}$  shown in bold.

We can then ensure globby boundaries by embedding  $S_1$  in a larger equilateral triangle and using appropriate projections to 'truncate' movement outside of  $S_1$  to movement inside  $S_1$ .

However, in higher dimensions,  $S_n$  does not tile  $\mathbb{R}^n$  through reflections (there is no way to tile  $\mathbb{R}^3$  using the tetrahedron *at all*). As a consequence, there is no unique way of mapping a point in  $R^n$  to  $S_{n-1}$  since the reflected point depends on the exact path taken. Since computer simulation must invariably happen in discrete time steps, this means that even in principle, we cannot simulate the motion correctly. Further, even in a discrete-time version of our model, computing the appropriate reflections seems mathematically intractable at the moment.

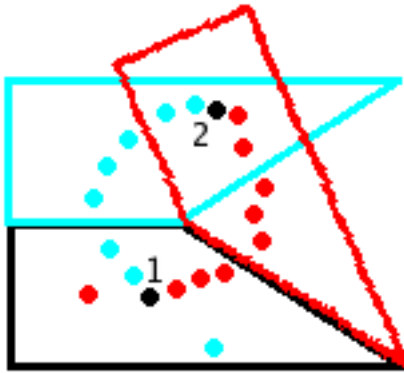


Fig. 5. 2-D example of the non-uniqueness problem with reflections. The black trapezium when reflected in the horizontal and slanted edges yield the blue and red trapezia respectively. Note that in the area of overlap between the blue and the red trapezia, we cannot uniquely map points to a point in the black trapezium:

If point 1 moved to point 2 via the blue path, point 2 should be mapped to the blue point, whereas if it moved to point 2 via the red path, it should be mapped to the red point, which is distinct. This problem does not occur if we used an equilateral triangle instead of the trapezium.

## 6 Non-orthogonal coordinate systems

One might try to model the random drift by using a non-orthogonal coordinate system with origin  $\mathbf{O}$  at the centre of  $S_n$  and  $n - 1$  basis vectors,  $\mathbf{e}_k$  going from  $\mathbf{O}$  to  $n - 1$  of the vertices of  $S_n$ . If  $x_t^i$  for  $i = 0, \dots, n - 1$  are  $n - 1$  independent processes moving between 0 and 1 (e.g. by using the 1D process in the above section), we can let  $p_t = \mathbf{O} + \sum_i x_t^i \mathbf{e}_k$ . However, this approach fails to be symmetrical in the arguments to  $p_t$ .

## 7 Mass point representation

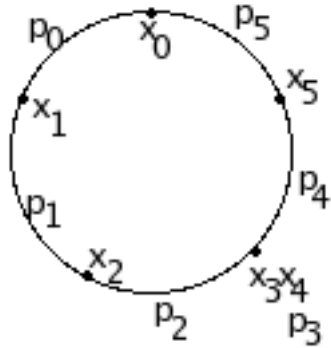


Fig. 6. Mass points  $x_i$  and probabilities  $p_i$ . Note that  $p_3 = 0$ .

A different approach to modelling random drift is to use a discrete-time model with  $n$  mass points arranged on a circle of circumference 1. If we label the points counterclockwise as  $x_0, \dots, x_n$ , we define the probability  $p_i (= p_t(i + 1))$  to be the length of the arc going counterclockwise from  $x_i$  to  $x_{i+1 \bmod n}$ . Two mass points can occupy the same location, in which case they form a *cluster*. In general, we define a *cluster* to be any maximal set of mass points occupying the same location: In fig. 6, the clusters are  $\{x_0\}$ ,  $\{x_1\}$ ,  $\{x_2\}$ ,  $\{x_3, x_4\}$  and  $\{x_5\}$ .

Ideally at every time step, we would like our process to add a Gaussian perturbation to every cluster while simultaneously tossing a coin for every cluster to see if it should be split and if so split it using two Gaussian perturbations in each direction. However, since this is difficult at best to simulate, we use an approximate process:

Let  $c$  be the number of clusters at the beginning of our time step. Pick two random (not necessarily distinct) clusters  $x$  and  $y$ , generate a Gaussian( $0, \sigma_1^2$ ) perturbation,  $\varepsilon$ , and add it to all clusters on the arc going counterclockwise

from  $x$  to  $y$ . To ensure globby boundaries, if  $x$  hits a neighbouring cluster,  $z$  say, then we only move the arc from  $x$  to  $y$  far enough for  $x$  to merge with  $z$ . Similarly, if  $y$  hits a neighbouring cluster. Repeat this  $c$  times.

Now let  $k$  be the number of probabilities equal to zero at this point.  $k$  times we pick a random zero probability  $p_i$  and toss a coin that lands heads with probability  $\tau$ . If it lands heads, we split the cluster containing  $x_i$  in two by generating two Gaussian( $0, \sigma_2^2$ ) perturbations,  $\varepsilon_1$  and  $\varepsilon_2$ , and moving all points in the cluster  $x_j$  for  $j \leq i$   $|\varepsilon_1|$  away from the centre and all points  $x_j$  in the cluster for  $j > i$   $|\varepsilon_2|$  the other way. Any collisions occurring as a result of this movement are resolved in the same way as above.

Now the above satisfies both the symmetry and stationarity properties required and simulation is easy, but inference is difficult and no inference has been attempted. Below are some simulation runs using this process ( $\sigma_1 = 0.005$ ,  $\sigma_2 = \sigma_1/2$ ):

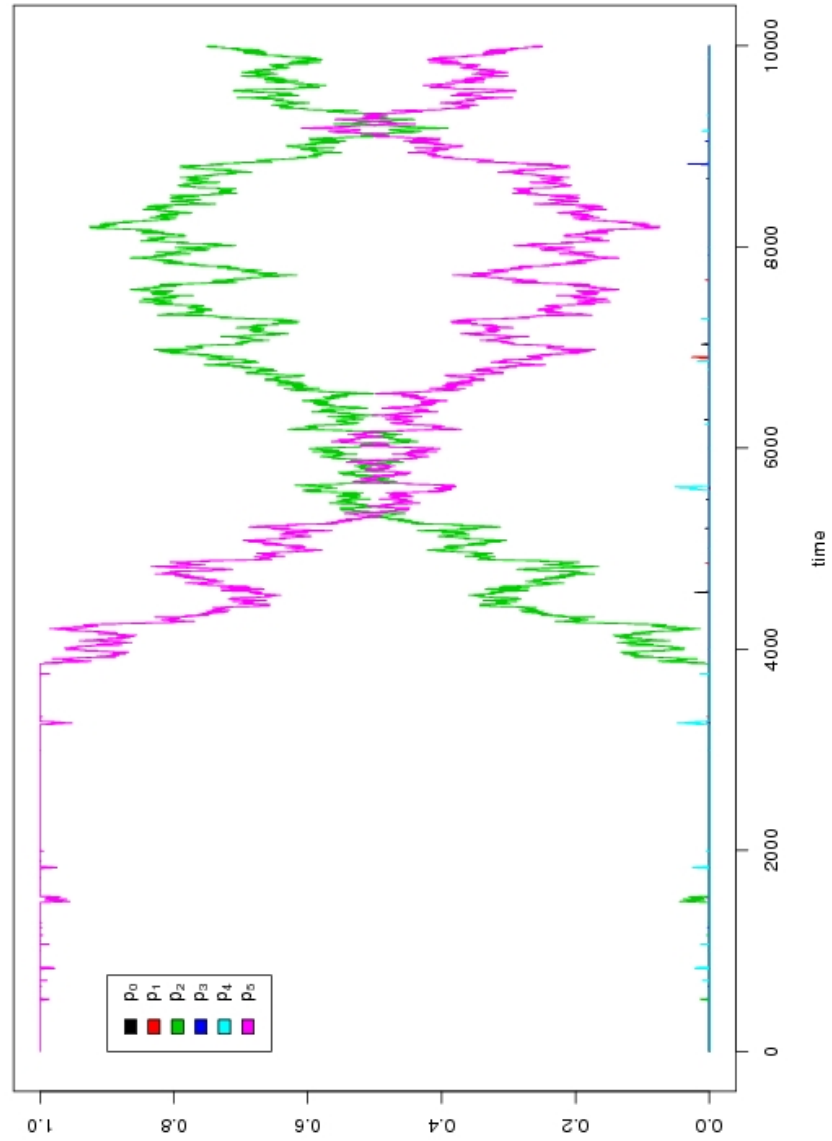


Fig. 7. Initial settings  $p_i = 0$  for  $i < 5$  and  $p_5 = 1$ ,  $\tau = 0.001$ .

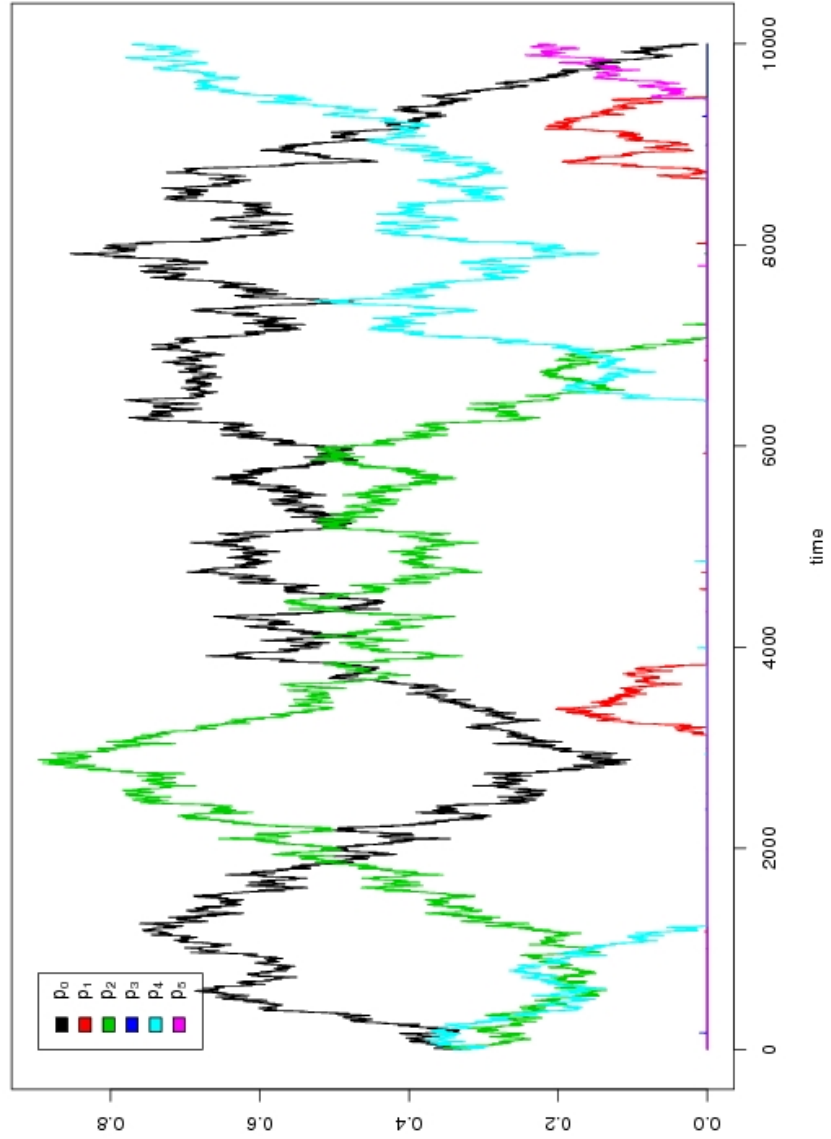
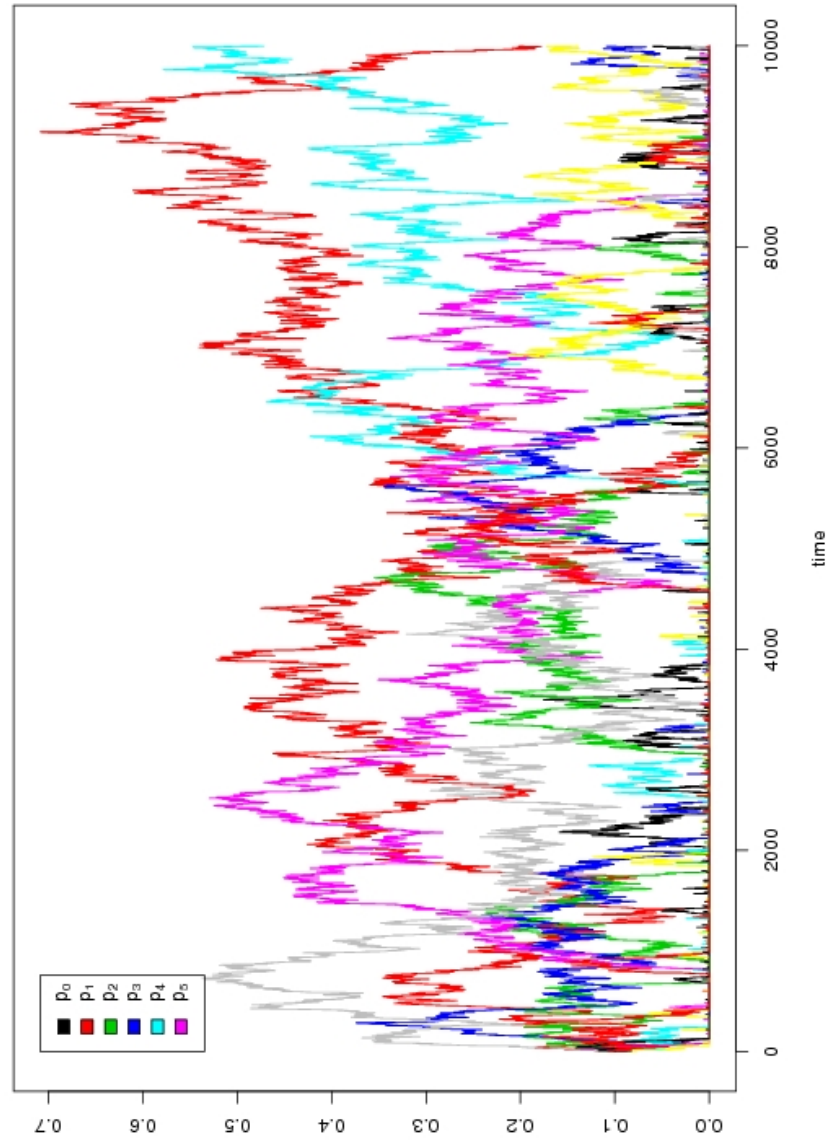
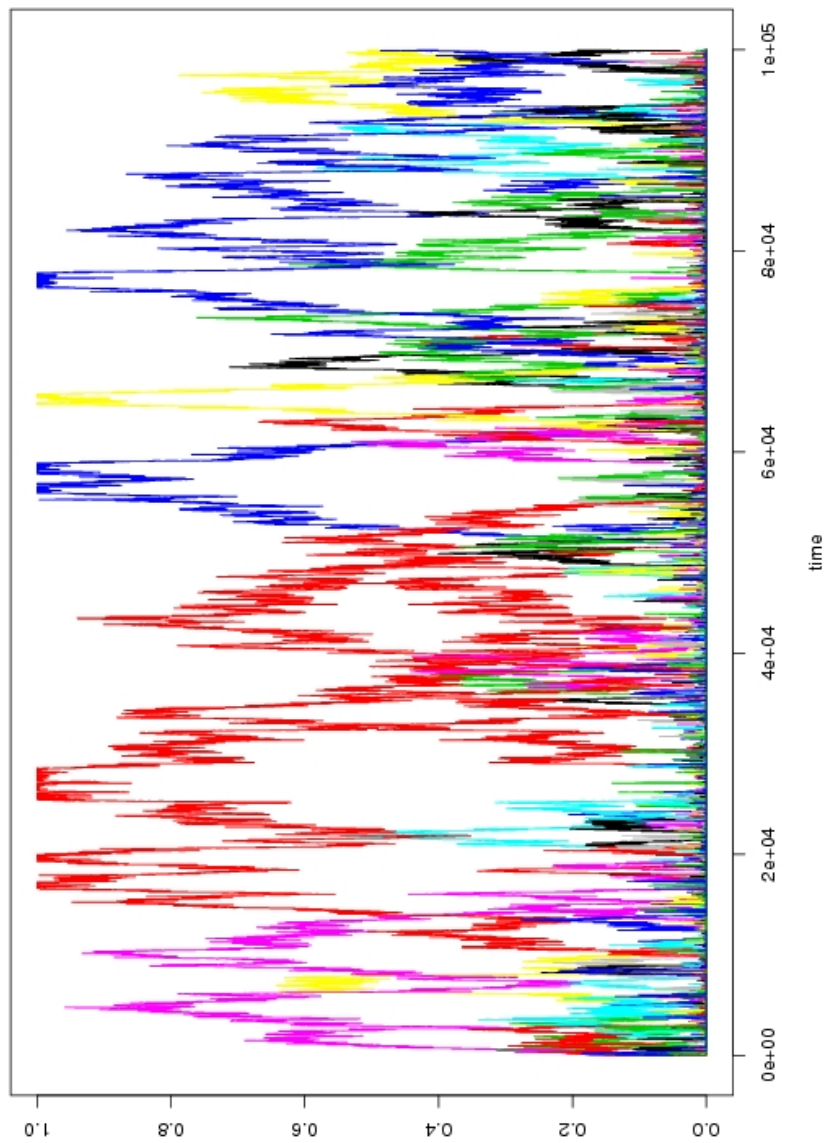


Fig. 8. Initial settings  $p_0 = 0.33, p_1 = 0, p_2 = 0.33, p_3 = 0, p_4 = 0.34, p_5 = 0,$   
 $\tau = 0.001.$



*Fig. 9. Initial settings  $p_i = 0.1$  for  $0 \leq i < 10$ ,  $\tau = 0.01$ . Only 6 probabilities shown in legend.*



*Fig. 10. Initial settings  $p_i = 0.05$  for  $0 \leq i < 20$ ,  $\tau = 0.002$ .*

## 8 Inference

While we can simulate random drift on probability distribution with an arbitrary number of probabilities using our mass-point representation, inference appears extremely difficult. Suppose e.g. we want to estimate the probability of distribution  $\mathbf{p}$  evolving into a member of  $\Lambda$  in time  $t$ , where  $\Lambda$  is a family of distributions (of non-zero measure). Finding a tractable mathematical expression for this probability seems difficult, so we would be forced to run a large number of simulations starting with  $\mathbf{p}$  and seeing whether after  $t$  time steps,  $\mathbf{p}$  fell into  $\Lambda$ .

However, in the **simplex model** in 1 dimension, we can write down the density for transiting from position  $x$  to  $y$  in time  $t$  as an infinite sum. Using suitable truncations of this sum, we can use numerical methods for inference. If we further change our underlying motion model from Brownian motion to *Cauchy processes*. Let  $\{x_t\}$  be a reflecting Cauchy process constructed in the same manner as the reflecting Brownian motion in the simplex model (section 5). The transition density for a Cauchy process is  $p(y|x, t) = \frac{1}{\pi} \frac{t}{t+(y-x)^2}$ . The transition density for the reflecting Cauchy process can then be shown to be:

$$d(y|x, t) = \sum_{n \in \mathbb{Z}} \frac{1}{\pi} \frac{t}{t + (y - x + 4n)^2} + \sum_{n \in \mathbb{Z}} \frac{1}{\pi} \frac{t}{t + (2 - y - x + 4n)^2}$$

And using contour integration from complex analysis (see e.g. Priestley (2003) for an introduction to complex analysis), we can evaluate this to:

$$\frac{1}{8} \frac{\sinh(\frac{\pi t}{2})}{\cosh^2(\frac{\pi t}{4}) - \cos^2(\frac{\pi}{4}(y-x))} + \frac{1}{8} \frac{\sinh(\frac{\pi t}{2})}{\cosh^2(\frac{\pi t}{4}) - \sin^2(\frac{\pi}{4}(y-x))}$$

We can generalize this to the simplex model in 2 dimensions. In general this model becomes difficult to deal with.

Evans (2002), in fact, investigates a general approach to constructing Brownian motion on the simplex by tiling hypersurfaces with simplexes and mapping the whole hypersurface to one simplex. A simple  $n$ -dimensional way of constructing random drift on the  $n$ -dimensional simplex using Brownian motion will be to use Brownian motion on the  $n$ -dimensional unit sphere and then since for all vectors  $(x_1, \dots, x_n)$  on the unit sphere,  $x_1^2 + \dots + x_n^2 = 1$ ,  $\{(x_1^2, \dots, x_n^2)\}$  will automatically give us the desired motion. Since any given probability distribution  $(p_1, \dots, p_n)$  corresponds to the  $2^n$  points  $(\pm\sqrt{p_1}, \dots, \pm\sqrt{p_n})$ , i.e. a finite number of points, this gives a fairly tractable model for inference.

In general, March (1995)<sup>1</sup> has shown that any semimartingale process (quite a large class of processes) on the simplex can be characterized in a form similar to that of *Fleming-Viot processes*.

---

<sup>1</sup>[http://www2.economia.unimi.it/PAS/Letters/letter\\_26.shtml](http://www2.economia.unimi.it/PAS/Letters/letter_26.shtml)