Discrete random structures whose limits are described by a PDE: 3 open problems

David Aldous

April 3, 2014

Much of my research has involved study of $n \to \infty$ limits of "size n" random structures. There are many techniques one can try. This talk is about **models where it's easy to see (heuristically) there should be some kind of limit process which is determined (somehow) by a specific PDE.**

In such cases I want to start with the explicit solution of the PDE – then the "hard work" comes when we try to formalize heuristics. (In principle not need explicit solution, but in practice . . . . . .). This is a very old-fashioned attitude, to PDE theorists!

The most familiar interface between Probability and PDEs is the theory of finite- or infinite-dimensional diffusions. My examples are (mostly) different.

The only example where I could carry this program through was in Aldous-Diaconis (1995) *Hammersley's Interacting Particle Process and Longest Increasing Subsequences*.

Limit was determined by a function $F(t, x)$ on $0 < t < \infty, 0 < x < \infty$ satisfying (writing F_t for $\frac{\partial F}{\partial t}$)

$$F_t = 1/F_x$$

$$F(t, 0) = F(x, 0) = 0$$

Here we knew the answer in advance (because re-proving known result)

$$F(t, x) = 2\sqrt{tx}.$$

The talk will describe 3 examples where I can't solve the PDE explicitly and so haven't got started ......

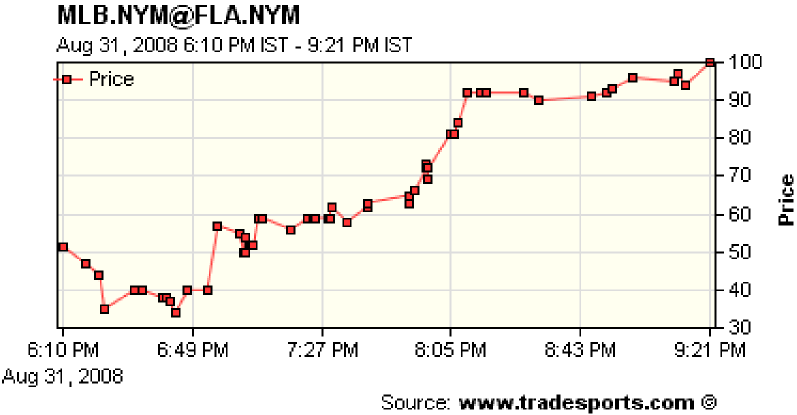
1. How to make a spectator sport exciting

You are watching a match in real time.

$p(t)$ = chance home team wins, given what has happened so far.

Several ways to think about the process $p(t)$

(a) via real-time gambling odds



(b) Via a math model for the point difference process
$X(t) =$ (points by home team) - (points by visiting team)

e.g. for soccer: team i scores at times of a Poisson (rate λ_i) process.

Given any model for $X(t)$, and taking match duration as 1 time unit, we have the derived "price process"

$$p(t) = \mathbb{P}(X(1) > 0 | \mathcal{F}(t)). \tag{1}$$

If designing a new sport, what would we want the process $p(t)$ to be? Imagine equally good teams, so

$$p(0) = 1/2; \quad p(1) = 1 \text{ or } 0, \text{ equally likely.}$$

What would we like the distribution of $p(1/2)$ to be? To a spectator:

- if $p(1/2)$ typically close to $1/2$ then only the second half of the match is important
- if $p(1/2)$ typically close to 1 or 0 then only the first half of the match is important

Note paradox: an individual match is exciting if result open until near the end, but we don't want that to happen in every match.

Rules of the sport determine point difference process
$X(t) =$ (points by home team) - (points by visiting team)
which then determines "price process"

$$p(t) = \mathbb{P}(X(1) > 0 | \mathcal{F}(t)).$$

Clearly $p(t)$ must be a martingale. Simplify by assuming $p(t)$ is a continuous-path martingale and a (time-inhomogeneous) Markov process (roughly, this is saying $X(t)$ is continuous and (time-inhomogeneous) Markov). So we have

$$dp(t) = \sigma(p(t), t)\ dB(t)$$

for some variance rate $\sigma^2(x, t)$.

Question. We can in principle (cf. Jeopardy) design a game to have any chosen $\sigma(x, t)$, up to two integrals being finite/infinite. How do we choose?

Under the simplest model (e.g. the soccer model) for (continuized) point difference

$$p(1/2) \text{ has } U(0,1) \text{ distribution.} \qquad (2)$$

We have a small data-set consistent with this theory. See 2013 Monthly paper *Using Prediction Market Data to Illustrate Undergraduate Probability*.

Is (2) desirable?

- from entropy viewpoint ???
- from analysis of variance viewpoint, $1/3$ of variance is resolved in first half, $2/3$ in second half ???

Return to previous slide. We study the question there by considering maximizing entropy for the whole process $(p(t), 0 \leq t \leq 1)$. Not clear how to do this directly in continuous time/space; we first discretize then pass to a limit.

Take integer parameters (T, N). Take discrete state space $\{-N, -N+1, \dots, N-1, N\}$. We can construct the discrete time process $(X_s,\ s = 0, 1, 2, \dots, T)$ which is the maximum entropy process satisfying

$$X(0) = 0; \quad X(T) = N \text{ or } -N \tag{3}$$

and the martingale property. The process is the time-inhomogeneous Markov chain whose transition probabilities $p_s(i,j) = P(X_{s+1} = j | X_s = i)$ are defined by backwards induction as follows. Clearly for $s = T-1$ we must have

$$p_{T-1}(i, N) = \tfrac{i+N}{2N}, \quad p_{T-1}(i, -N) = \tfrac{N-i}{2N}.$$

Define

$$e_{T-1}(i) = -\tfrac{i+N}{2N} \log \tfrac{i+N}{2N} - \tfrac{N-i}{2N} \log \tfrac{N-i}{2N}$$

that is the entropy of the distribution $p_{T-1}(i, \cdot)$.

Now inductively for $s = T-2, T-3, \ldots, 0$, for each i we define $p_s(i, \cdot)$ as the distribution $q(\cdot)$ on $[-N, N]$ which maximizes

$$-\sum_j q(j) \log q(j) + \sum_j q(j) e_{s+1}(j) \tag{4}$$

subject to having mean $= i$, and let $e_s(i)$ be the corresponding maximized value of (4). So this "dynamic programming" construction inductively specifies the maximum entropy process, starting at state i at time s, satisfying (3) and the martingale property.

Now a back-of-an-envelope calculation shows what the continuous limit should be. Rescale (s, i) to (t, x) and rescale $e_s(i)$ to

$$e(t, x), \quad 0 < t < 1, -1 < x < 1$$

We can "solve" the maximization problem to get the PDE

$$e_t = \frac{1}{2} \log(-e_{xx}) \tag{5}$$

with boundary conditions

$$e(t, \pm 1) = 0,\ 0 \leq t < 1; \quad e(1, x) = 0, -1 < x < 1;$$

and we find that

$$\sigma^2(t, x) = \frac{-1}{e_{xx}(t, x)} \tag{6}$$

STOP.

2. The Lake Wobegon Publishing Group.

Recall that in Lake Wobegon

> *all the women are strong, all the men are good looking, and all the children are above average.*

In this spirit the Lake Wobegon Math Society journal only publishes papers whose quality is **above the average** quality of their previously-published papers. We model the quality $U_1, U_2, \ldots$ of successive submissions as IID.

Remark. Analogous to "record process" but not distribution-free. Use $U(0,1)$ or Exponential(1) distributions.

Exercise (graduate stochastic processes course):
study $Z_n :=$ number of accepted papers among first n submissions,

Now imagine a second journal that considers all papers rejected by first journal, and uses the same "better than average" rule for acceptance. And so on ......

This is a "putting objects into piles" random model. There are two well known such models that are corners of large topics.

Chinese Restaurant Process.

Patience sorting. Cards with IID U(0,1) labels put into piles so we see the label on the top card in each pile. These visible labels are in increasing order, for instance

$$0.15, 0.33, 0.40, 0.54, 0.71, 0.83$$

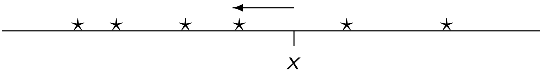
If next card has label 0.45 we put it on top of the 0.40 card (largest label smaller than the card we are placing); if it had label 0.09 we would start a new pile on the left.

Our "Lake Wobegon Publishing Group" model is the variant of "patience sorting" where we use the *average* of the labels in each pile (instead of the maximum) to determine where the next card is placed.

To study patience sorting informally consider

$$F(t,x) = \mathbb{E}(\text{ number of piles with top card label } \le x).$$

At given t the process of top card labels is approximately a spatial Poisson process, rate $\lambda(x) = F_x$. So chance that next card increases the number of piles with top card label $\le x$ is the chance the next label is in the interval



This chance $= 1/\lambda(x)$, so this gives the PDE

$$F_t = 1/F_x$$

stated at start of talk, whose solution is $F(t,x) = 2\sqrt{tx}$. So the process has $\sim 2n^{1/2}$ piles with order $n^{1/2}$ cards in each.

Background: number of piles = length of longest increasing subsequence of a random permutation.

To study our "Lake Wobegon Publishing Group" model informally, consider

$$F(t,x) = \mathbb{E}(\text{ number of piles with average card label } \leq x).$$

A slightly more elaborate calculation gives

$$\frac{1}{F_x} = \left(\frac{1}{2F_tF_x}\right)_t$$

which can be rearranged to

$$F_{tt}F_x + F_{xt}F_t + 2F_t^2F_x = 0.$$

Boundary conditions

$$F(t,0) = F(0,x) = 0.$$

STOP.

3. Online $\zeta(3)$. The setting is:

- complete graph on n vertices
- assign IID Exponential(mean n) edge-lengths
- $L_n =$ length of minimum spanning tree (MST).

A remarkable "offline" result of Frieze (1985) shows

$$n^{-1}\mathbb{E}L_n \sim \zeta(3).$$

The formula can be seen as arising from two simple relations. First

$$\lim_n n^{-1}\mathbb{E}L_n = \tfrac{1}{2}\int_0^\infty \lambda p(\lambda) d\lambda$$

where $p(\lambda)$ is the (limit) probability that a length-λ edge is in the MST.

[repeat] $p(\lambda)$ is the (limit) probability that a length-λ edge is in the MST. Second

$$1 - p(\lambda) = q^2(\lambda) \tag{7}$$

where $q(\lambda)$ is the probability that a Galton-Watson Branching Process with Poisson(λ) offspring survives forever.

Here (7) holds because
(i) a length-λ edge is **not** in the MST iff there is an alternative path between its end-vertices using only edges of length $< \lambda$.

(ii) relative to a given vertex, the process of edges of length $< \lambda$ is (in $n \to \infty$ limit) a Galton-Watson BP with Poisson(λ) offspring.

(iii) Both BPs surviving forever ($n = \infty$) corresponds for large finite n to having overlapping giant components.

In project with Omer Angel and Nathanael Berestycki (moribund since 2008) we study the corresponding "online" spanning tree problem where the edge-lengths are shown in random order, and one must decide immediately whether to use the edge in the tree.

Fix n. Any scheme for constructing a tree will build up a forest of tree-components. Represent state as a point $\mathbf{x} = (x_1, \ldots, x_n)$ in the simplex Δ_{n-1}

$$x_i = \text{ proportion of vertices in size-}i\text{ components.}$$

Initial state is $(1, 0, 0, 0, \ldots)$ but we consider

$$F_n(\mathbf{x}) = n^{-1}\mathbb{E}(\text{length of optimal online tree starting from } \mathbf{x}).$$

Use classical method: get equation for $F_n(\cdot)$ by conditioning on first step.

$$F_n(\mathbf{x}) = n^{-1}\mathbb{E}(\text{length of optimal online tree starting from } \mathbf{x}).$$

Write $\mathbf{x}^{ij}$ for the vector obtained from $\mathbf{x}$ if we accept an edge linking a size-i and a size-j component. Clearly the optimal rule is:

- accept such an edge iff its length is $< F_n(\mathbf{x}) - F_n(\mathbf{x}^{ij})$.

Then by considering the cost of the first accepted edge from $\mathbf{x}$ we find

$$F_n(\mathbf{x}) = n^2 \sum_{i,j} x_i x_j \frac{(F_n(\mathbf{x}) - F_n(\mathbf{x}^{ij}))^2}{4} \qquad (8)$$

which is in fact exact for a Poissonized arrival process.

We now just write down the heuristic $n \to \infty$ limit of (8) as a PDE for a function $G(\mathbf{y})$ on the infinite-dimensional simplex.

$$G = \frac{1}{4} \sum_{i,j} y_i y_j (iG_i + jG_j - (i+j)G_{i+j})^2$$

$$G_i = \frac{\partial G(\mathbf{y})}{\partial y_i}$$

A PDE for a function $G(\mathbf{y})$ on the infinite-dimensional simplex.

$$G = \frac{1}{4} \sum_{i,j} y_i y_j (iG_i + jG_j - (i+j)G_{i+j})^2$$

A boundary condition is

$$G(\mathbf{y}^n) \to 0 \text{ whenever } y_i^n \to 0 \ \forall i$$

We also have, from the fact $F_n(\mathbf{x}) - F_n(\mathbf{x}^{ij}) > 0$, that

$$iG_i + jG_j - (i+j)G_{i+j} > 0.$$

So we conjecture there is a unique solution to the PDE and that the "online $\zeta(3)$ constant" is $G(1, 0, 0, 0, \ldots)$.

STOP.