Rotor-routing, smoothing kernels, and reduction of variance: breaking the O(1/n) barrier

Jim Propp (UMass Lowell)

ICERM Computational Challenges in Probability Seminar
September 11, 2012

Slides for this talk are on-line at
http://jamespropp.org/icerm12.pdf

Acknowledgments

This talk describes work that evolved from conversations and collaborations with David Einstein, Ander Holroyd, and Lionel Levine, as well as answers I received to questions I posted on MathOverflow (see http://mathoverflow.net).

I. Introduction

Monte Carlo simulation (the two-minute course)

Consider a sequence of identically distributed r.v.'s $X_1, X_2, \ldots$ with $\mathrm{Law}(X_n) = \mathrm{Law}(X)$ for all n.

Let $S_n := X_1 + \cdots + X_n$.

$\mathrm{Exp}(S_n/n) = \mathrm{Exp}(X)$, and if the X_n's are independent, the LLN says $S_n/n \to \mathrm{Exp}(X)$ almost surely.

Monte Carlo simulation (the two-minute course)

When it's easy to generate IID samples from $\mathrm{Law}(X)$ but hard to compute $\mathrm{Exp}(X)$, this gives a way to estimate $\mathrm{Exp}(X)$ empirically (the "Monte Carlo method'):

Estimate $\mathrm{Exp}(X)$ by $S_n/n \sim \mathrm{Exp}(X)$ for some large n.

If $\mathrm{Var}(X)$ is finite, then $\mathrm{Var}(S_n/n) = \mathrm{Var}(S_n)/n^2 = n\,\mathrm{Var}(X)/n^2 = \mathrm{Var}(X)/n$, so the standard deviation $\sigma(S_n/n)$ of our estimate is typically $O(1/\sqrt{n})$.

Quasi-Monte Carlo, aka "Casablanca simulation" (Customer: "Are you sure this place is honest?")

Idea: We make $\sigma(S_n/n)$ smaller than $O(1/\sqrt{n})$ by "cheating", i.e., by using NON-independent, identically distributed r.v.'s.

Here we are working in the space of all joinings $\mathrm{Law}(X_1, X_2, \dots)$ whose nth marginal is $\mathrm{Law}(X)$ for all n, and we are trying to simultaneously minimize the functionals $\mathrm{Var}(S_n/n)$, or equivalently minimize the functionals $\mathrm{Var}(S_n)$.

Note that S_n/n is automatically an unbiased estimator of $\mathrm{Exp}(X)$.

The Bernoulli case is special

For most r.v.'s, these goals conflict.

E.g., when $\mathrm{Law}(X)$ is uniform on $[0, 1]$,
there are joinings for which $\mathrm{Var}(S_2)$ is zero (easy),
and other joinings for which $\mathrm{Var}(S_3)$ is zero (a fun puzzle),
but there's no joining that achieves both at the same time.

But when $\mathrm{Law}(X)$ is $\mathrm{Bernoulli}(p)$, there's a law for $X_1, \ldots, X_n$
that simultaneously minimizes all of the $\mathrm{Var}(S_n)$'s (with $n \geq 2$).

The first "half" of this talk

Such "maximally anticorrelated Bernoulli sequences" can be used as components of networks ("randomized rotor-router networks") that compute anticorrelated sequences of r.v.'s of various kinds.

I'll apply this to a random variable associated with absorbing Markov chains (namely the indicator r.v. of absorption in a particular state) and show that with randomized rotor-routers we get $\mathrm{Var}(S_n/n) = O(1/n^2)$ (cf. $\mathrm{Var}(S_n/n) = O(1/n)$ for IID).

That is, the typical difference between S_n/n and $\mathrm{Exp}(X)$ is $O(1/n)$ rather than $O(1/\sqrt{n})$.

Moreover, I'll show that the tail of the distribution of $S_n - \mathrm{Exp}(X)$ isn't just small; it vanishes beyond a certain point.

The second "half" of this talk

The preceding result is clearly best possible; $\mathrm{Var}(S_n)$ can't be $o(1)$, since S_{n-1} and S_n differ by an instance of X.

So $\mathrm{Var}(S_n/n)$ can't be $o(1/n^2)$ and $\sigma(S_n/n)$ can't be $o(1/n)$.

But in the last part of this talk, I'll describe how to use $X_1, \ldots, X_n$ to get estimates of $\mathrm{Exp}(X)$ with typical error $o(1/n)$.

II. Rotor-routing

"MAID" (Maximally Anticorrelated, Identically Distributed) Bernoulli sequences

With $0 < p < 1$, and U uniform in $[0, 1)$, let
$X_n = \lfloor U + np \rfloor - \lfloor U + (n-1)p \rfloor \in \{0, 1\}$ for all $n \geq 1$.

Then $\mathrm{Law}(X_n)$ is $\mathrm{Bernoulli}(p)$, and

$$S_n = \lfloor U + np \rfloor - \lfloor U \rfloor \in \{\lfloor np \rfloor, \lceil np \rceil\},$$

so S_n has variance as small as it can be, subject to $\mathrm{Exp}(S_n) = np$.

$(X_1, X_2, \dots)$ is a random Sturmian sequence of density p.

A physical model

We have a spinner consisting of an arrow pinned to a disk; the center of the arrow is pinned to the center of a disk and the arrow is free to rotate. The disk has a black sector occupying a proportion $0 < p < 1$ of the disk.

If we were to repeatedly randomize the arrow, outputting a 1 or a 0 according to whether the arrow pointed into the black sector or not, we would get an IID $\mathrm{Bernoulli}(p)$ process.

But instead we randomize the arrow just once at the start, and on subsequent turns we merely rotate it counterclockwise p turns around the circle, so that our sequence of 1's and 0's is a MAID $\mathrm{Bernoulli}(p)$ process.

Irrational p vs. rational p

When p is irrational, there is a measure-preserving almost-bijection between the circle $\mathbf{R}/\mathbf{Z}$ and the set of Sturmian sequences of density p.

When p is rational, say $p = a/b$ in lowest terms, then there are just b Sturmian sequences of density p, and the MAID $\mathrm{Bernoulli}(p)$ process gives them equal likelihood.

E.g., for $p = 2/5$, the law of $(X_1, X_2, \dots)$ is supported on five infinite sequences of period 5, each of which has probability $1/5$:

$(1, 0, 1, 0, 0, \dots)$
$(0, 1, 0, 1, 0, \dots)$
$(0, 0, 1, 0, 1, \dots)$
$(1, 0, 0, 1, 0, \dots)$
$(0, 1, 0, 0, 1, \dots)$

Markov chains and hitting probabilities

Consider a finite-state absorbing Markov chain with a designated source-state s and absorbing states ("target-states") $t_1, \ldots, t_m$, such that Prob(the chain eventually enters $\{t_1, \ldots, t_m\}$ | the chain starts at state s) $= 1$.

Let t^* be one of the target states, and let p^* be the probability that the chain eventually enters state t^*.

Then $p^* = \mathrm{Exp}(X)$, where X is 1 or 0 according to whether a run of the Markov chain leads to absorption at t^* or absorption elsewhere.

Particles on a directed graph

It's convenient to imagine that the states of the chain are vertices in a directed graph, and that the evolution of the chain corresponds to the trajectory of a particle that travels from the ith vertex to the jth vertex when the Markov chain goes from state i to state j.

Gambler's ruin

For simplicity, assume that each state i of the Markov chain has just two successors, with probabilities p_i and $1 - p_i$.

We'll focus on the Markov chain with state-space $\{0, 1, 2, 3\}$ where 1 is the source and 0 and 3 are the targets, with all allowed transitions of the form $i \to i + 1$ and $i \to i - 1$, with $\text{Prob}(1 \to 2) = \text{Prob}(2 \to 3) = p$.

(This corresponds to the purse-size of a gambler who starts out with \$1 and makes a succession of bets, gaining \$1 with probability p and losing \$1 with probability $1 - p$, until he either has \$3 and leaves happy or has \$0 and leaves broke.)

Let $t^* = 3$. It's easy to prove that $p^* = p^2/(1 - p + p^2)$. But what if we want to estimate p^* empirically by running the chain?

Monte Carlo made complicated

To simulate the gambler's ruin Markov chain repeatedly (in "supertime") we use random variables $U_{i,k}$ ($i \in \{1, 2\}$, $k \in \{1, 2, 3, \dots\}$) where $U_{i,k}$ is the source of randomness we use when, for the kth supertime, a run of the Markov chain has put us in state i.

All the $U_{i,k}$'s are uniform in $[0, 1]$ and are independent of one another.

If the chain is in non-absorbing state i at time t, let its state at time $t + 1$ be $i + 1$ if $U_{i,k} \in [0, p_i)$ and $i - 1$ otherwise.

Monte Carlo made complicated

1	2	$\underline{3}$	1	$\underline{0}$	1	2	1	2	...
$U_{1,1}$			$U_{1,2}$		$U_{1,3}$		$U_{1,4}$		...
	$U_{2,1}$					$U_{2,2}$		$U_{2,3}$	...

Monte Carlo made complicated

(It may be helpful to imagine a spinner at i that we spin each supertime the Markov chain is in state i.)

When we reach a target state, we start again from the source state.

Each run will result in absorption at 0 or absorption at 3, outputting a $\mathrm{Bernoulli}(p^*)$ bit.

The successive $\mathrm{Bernoulli}(p^*)$ bits will be IID.

Anticorrelated quasi-Monte Carlo

Instead of having each sequence $U_{i,1}, U_{i,2}, \ldots$ be IID, have each sequence be MAID.

That is, $U_{i,k+1} = U_{i,k} + p_i \pmod 1$.

Once again there is a spinner at i, but instead of randomizing it each time we use it, we rotate it by p_i turns (i.e., by an angle of $2\pi p_i$).

We call this kind of spinner a rotor.

Rotor-routing

(see Holroyd, Levine, Mészáros, Peres, P., and Wilson, "Chip-Firing and Rotor-Routing on Directed Graphs," arXiv:0801.3306)

The particle starts at $i_1 = s$.
For $t = 1, 2, \ldots$ in succession, we update the rotor at i_t (incrementing it by $p_{i_t} \pmod 1$) and use it to decide which i_{t+1} the particle should go to, until the particle gets absorbed at a target.

Rotors can be used for general finite-state Markov chains (even when states have more than two successors), and indeed for some infinite-state Markov chains; see Holroyd and Propp, "Rotor Walks and Markov Chains", arXiv:0904.4507.

An example

Consider \$3 gambler's ruin with $p = 1/2$.

1st particle:	$1 \to 0$
2nd particle:	$1 \to 2 \to 3$
3rd particle:	$1 \to 0$
4th particle:	$1 \to 2 \to 1 \to 0$
5th particle:	$1 \to 2 \to 3$
6th particle:	$1 \to 0$
7th particle:	$1 \to 2 \to 1 \to 0$
etc.	(with period 3)

“Let’s see that again in configuration space”

$i = 1$	$i = 2$	
$\rightarrow$	$\leftarrow$	
		$\Downarrow$ absorption at 0
$\leftarrow$	$\leftarrow$	
		$\Downarrow$ absorption at 3
$\rightarrow$	$\rightarrow$	
		$\Downarrow$ absorption at 0
$\leftarrow$	$\rightarrow$	
		$\Downarrow$ absorption at 0
$\leftarrow$	$\leftarrow$	
		$\Downarrow$ absorption at 3
$\rightarrow$	$\rightarrow$	
		$\Downarrow$ absorption at 0
$\leftarrow$	$\rightarrow$	
		$\Downarrow \ldots$

Cycles go away

Note that initially the rotors at 1 and 2 form a 2-cycle (with each pointing toward the other), but that thereafter the graph formed by the rotors is acyclic.

More generally, for rotor-routing on any finite graph, the initial configuration of the rotors may contain cycles, but eventually all the cycles go away.

Stationarity

Theorem (P.): The operation of updating the rotors by sending a particle through the network preserves the restriction of product measure to the set of "acyclic rotor configurations".

(Note: It's easy to sample from this conditional distribution; use Wilson's partial rejection sampling scheme, aka cycle-popping, whereby you repeatedly rerandomize rotors that participate in cycles until there aren't any.)

ID-ness

If the initial setting of the rotors is governed by the conditional measure on acyclic configurations, then the successive runs are identically distributed.

More specifically, the outcome of the kth run (1 if the kth run leads to t^*, 0 otherwise) has law $\mathrm{Bernoulli}(p^*)$.

That is, the sequence of bits arising from anticorrelated Monte Carlo will be a sequence of (identically distributed) $\mathrm{Bernoulli}(p^*)$ bits, where p^* is the probability of absorption at target t^* for true Monte Carlo.

(See Holroyd and Propp, "Rotor Walks and Markov Chains", arXiv:0904.4507).

Confluence

Imagine that instead of just one particle in the directed graph we have many; at each step we can advance only one particle (by updating the rotor at the vertex it occupies and then moving the particle in the direction indicated by the rotor), but we get to choose which particle to move, until all particles have been absorbed at target vertices.

Confluence Property (aka abelian property): The number of particles absorbed at t_i does not depend on the choices we make.

Parallel rotor-routing

So, when sending n particles through the network, instead of sending one particle through the network at a time ("sequential rotor-routing"), we may start with all n particles at the source s, advance each particle one step, advance each not-yet-absorbed particle another step, advance each not-yet-absorbed particle another step, etc.

The number of particles absorbed at t^* will be the same for sequential rotor-routing and parallel rotor-routing.

Rotors achieve constant discrepancy

Theorem (Holroyd-P.): If one sends n particles through a finite network of rotors, the number that get absorbed at t^* differs from np^* by $O(1)$ (so that the number of particles that get absorbed at t^*, divided by n, differs from p^* by $O(1/n)$).

Illustration of proof in a simple case

Consider \$3 gambler's ruin with $p = 1/2$, $p^* = 1/3$.

Use parallel rotor-routing.

The sum of the positions of the particles is initially $n \times 1 = n$.

When there are an even number of particles at i, half go to $i - 1$ and half go to $i + 1$, so the sum of the positions isn't changed.

When there are an odd number of particles at i, the sum of the positions changes by ± 1, but the next time there are an odd number of particles at vertex i, the sum of the positions will change in the opposite direction.

So at each stage t the sum of the positions changes by ± 2 (since there are two sites i), but the cumulative changes never exceed ± 2.

Punchline of proof

When all the particles have been absorbed at 0 or 3,

$$A \times 0 + B \times 3 = n \pm 2$$

where A and B denote the number of particles absorbed at 0 and 3 respectively; hence

$$B = (1/3)n \pm 2/3.$$

(The general case involves extra technology and in particular uses harmonic functions on directed graphs, but no new ideas are required.)

III. Kernel smoothing

The $O(1/n)$ barrier

$\mathrm{Var}(S_n)$ can't be $o(1)$, since S_{n-1} and S_n differ by an instance of X.

So $\sigma(S_n/n)$ can't be $o(1/n)$.

Breaking the $O(1/n)$ barrier

To get past the barrier, we consider weighted sums

$$S_n^* = (a_1 X_1 + a_2 X_2 + \cdots + a_n X_n)/(a_1 + a_2 + \cdots + a_n)$$

with the a_i's not equal to one another (and depending on n).

(If the X_i's were IID, this wouldn't help at all!)

Note that such a weighted average is automatically an unbiased estimator of $\mathrm{Exp}(X)$.

I'll call the family of weights $a_{n,i}$ ($n \geq 1$, $1 \leq i \leq n$) a smoothing kernel, since a similar device in nonparametric estimation goes by that name.

Kernel smoothing: a digression and advertisement

Kernel smoothing can be successfully applied to estimating quantities in analysis

(the mean value of an almost periodic function),

geometry

(the density of a point-set with discrete spectrum),

and number theory

(the asymptotic average of an arithmetic function);

see the slides for my talk "How well can you see the slope of a digital line? (and other applications of averaging kernels)", http://jamespropp.org/Slope.cdf.

But I haven't got a general idea for when kernel smoothing should work and how much improvement it should yield;
there doesn't seem to be literature on it.

Parabolic weights

A good choice of weights for many applications, and rotor-routing in particular, is

$$a_{n,i} = i(n + 1 - i),$$

which I'll adopt hereafter.

The resulting weighted average is the slope of the least-squares regression line through the points

$$(1, S_1),\ (2, S_2),\ \ldots, (n, S_n)$$

where $S_i = X_1 + X_2 + .. + X_i$.

Rational p^*

Example: Suppose all transition probabilities in the Markov chain are rational, so that the sequence $X_1, X_2, \ldots$ is periodic, and p^* is rational.

Theorem (Einstein): The difference between the weighted sum S_n^* and the absorption probability p^* is $O(1/n^2)$.

(Note: The constant implicit in the $O(1/n^2)$ depends on the Markov chain.)

Random p

Theorem (P.): If $X_1, X_2, \ldots$ is a random Sturmian sequence of density p, where p is chosen uniformly at random in $[0, 1]$, then the standard deviation of the difference between the weighted sum S_n^* and the density p is $O(1/n^{3/2})$.

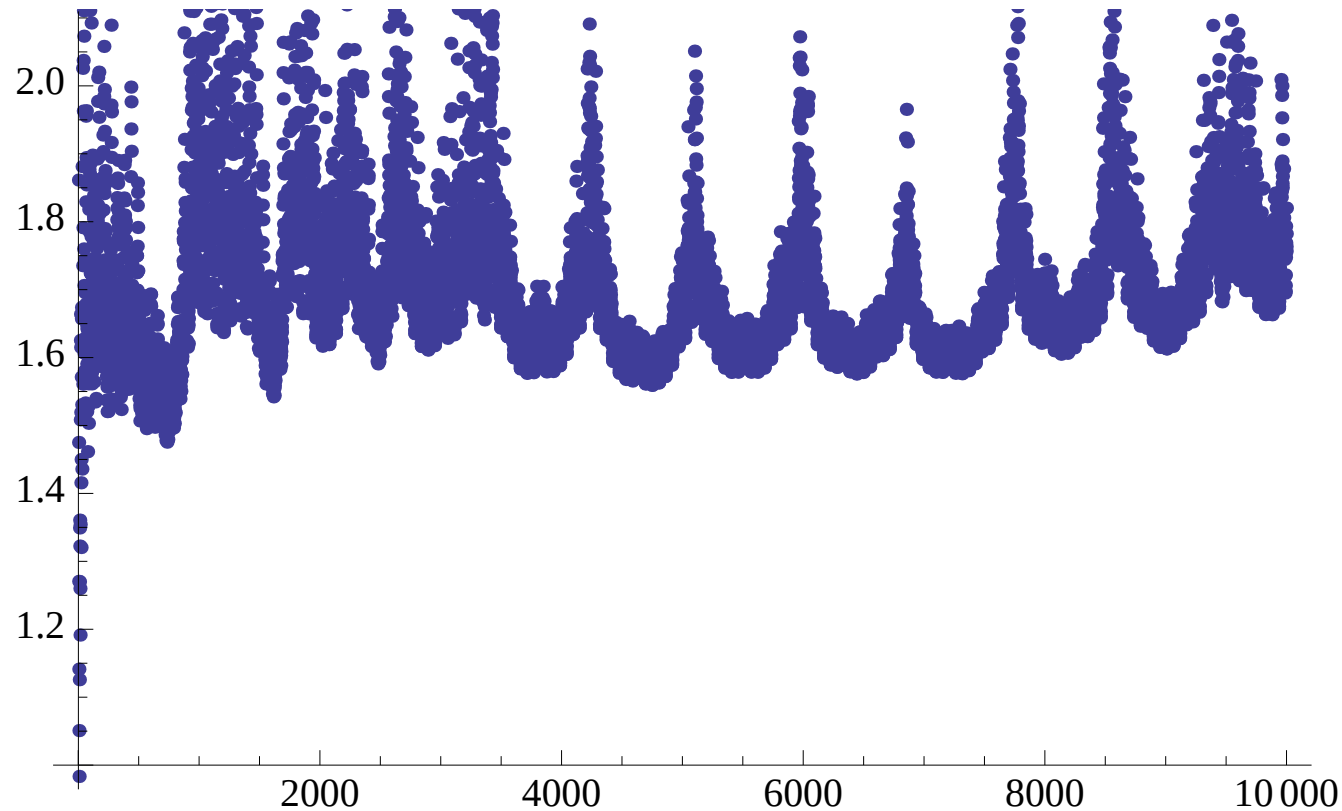
Does kernel smoothing help rotor-routing?

The figure on the next slide shows what we get when we do 10^4 rounds of Casablanca simulation of the \$3 gambler's ruin process with $p = 1/\pi$ with randomized rotors.

The horizontal axis records n,
and the vertical axis records the exponent α such that the discrepancy between S_n^* and p^* equals $1/n^\alpha$.

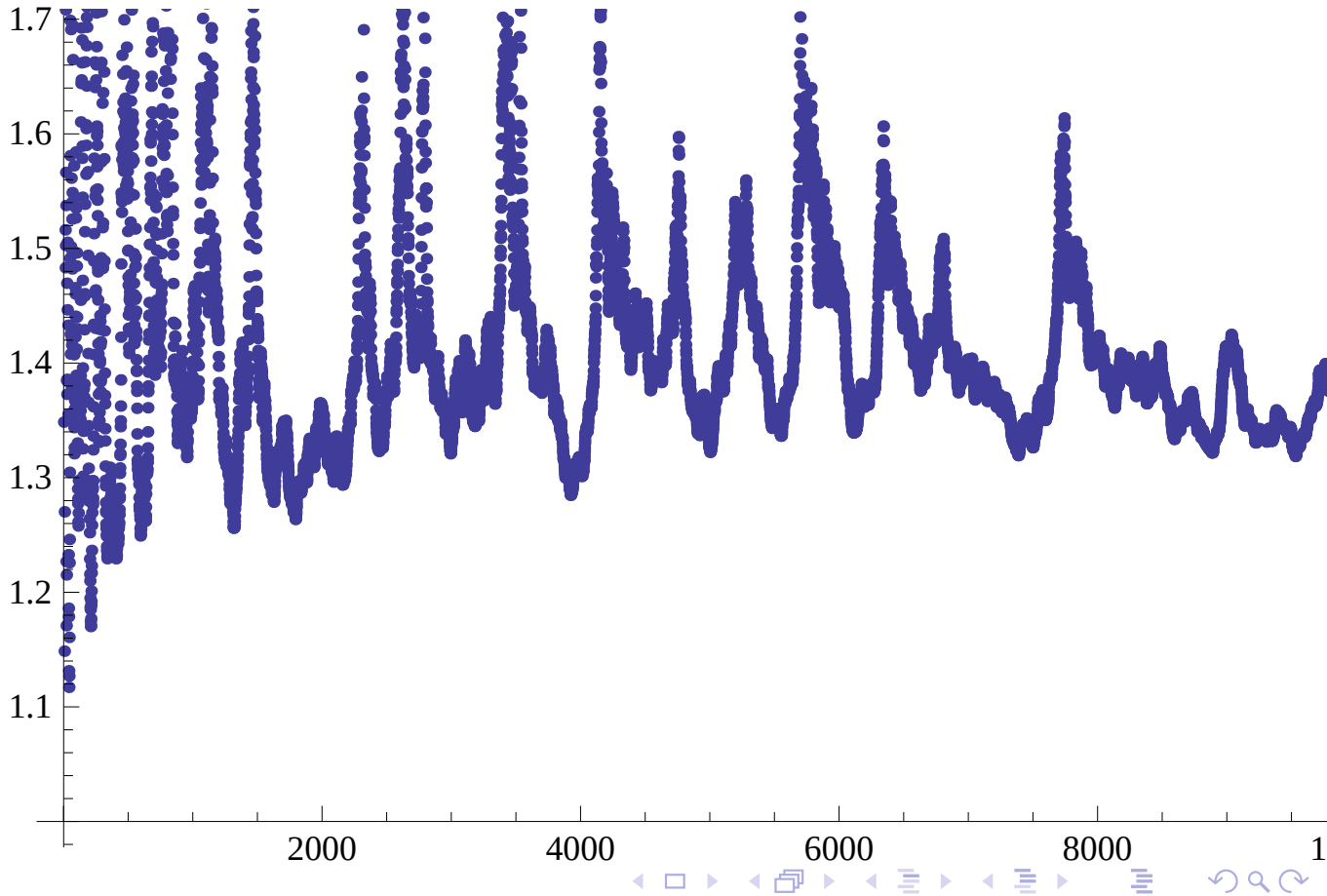
The line $\alpha = 1$ represents the $O(1/n)$ barrier; we want to go above it.

Out[21]=



Does kernel smoothing help rotor-routing?

Empirically, here's what we get when we do 10^4 rounds of Casablanca simulation of unbiased random walk on $\mathbf{Z} \times \mathbf{Z}$ with source $s = (0, 0)$ and targets $t_1 = (0, 0)$ and $t_2 = t^* = (1, 1)$ and $p^* = \pi/8$, with **non-randomized** rotors initialized in the "fylfot" configuration (see Holroyd and Propp, "Rotor Walks and Markov Chains", arXiv:0904.4507 for an explanation):


1.7
1.6
1.5
1.4
1.3
1.2
1.1
2000
4000
6000
8000

Non-randomized rotors?

Note that in the preceding case we are no longer doing probability theory; everything is deterministic.

However, it seems likely that probabilistic intuitions and even probabilistic theorems can be applied to the analysis of the asymptotic ($n \to \infty$) behavior of these systems, and be used to rigorously prove that kernel smoothing applied to rotor-routing breaks the $O(1/n)$ barrier.

(See Levine and Peres, "Strong Spherical Asymptotics for Rotor-Router Aggregation and the Divisible Sandpile", arXiv:0704.0688 for an example of how to use probabilistic theorems to prove results about a deterministic rotor-router process.)

IV. Conclusion

Derandomization of Monte Carlo with rotor-routers (for finite Markov chains) improves discrepancy from $O(1/\sqrt{n})$ to $O(1/n)$.

(I've discussed this in the context of absorption probabilities; for applications to absorption times, steady-state probabilities, etc., see Holroyd and Propp, "Rotor Walks and Markov Chains", arXiv:0904.4507.)

The use of kernel smoothing appears to give discrepancy $o(1/n)$ in many situations, but I don't have proofs of this outside the "instant absorption" case (Markov chains in which absorption always occurs on the first step) and a few other easy-to-analyze special cases. Nor do I understand what the relevant asymptotic for the kernel-smoothed discrepancy should be, even heuristically.

The problem

How can we get actual proofs that kernel smoothing breaks the $O(1/n)$ barrier for generic finite-state Markov chains?

Note that prior presults (Holroyd and P.) bounding the discrepancy between the absorption probability p^* and the ordinary average S_n/n (i.e, the relative frequency of absorption at target t^*) all required the confluence property.

However, the confluence property can't be used here, since we can't compute S_n^* without knowing the order in which the respective absorptions at, and not at, t^* occur.

Challenges

Find a method that will let us prove things about the estimate for p^* obtained by doing a rotor-router derandomized simulation of the absorption process and applying kernel smoothing to the resulting sequence of 0's and 1's.

A brand new idea is needed!
(Maybe some sort of nonabelian version of the confluence property?)

Also: Might there be some (nonlinear) method of estimating p^* that's better than (linear) kernel-smoothing?
Geometry of numbers says you can't hope (generically) to do better than $O(1/n^2)$, but David Einstein has almost proved (?) that for the seeing-the-slope-of-a-line problem you can get within a power-of-log factor of this bound.

And the big question...

Do these ideas (derandomization, kernel-smoothing) scale up for any actual applications?

If anyone can think of a good target application, please let me know!

Slides for this talk are on-line at
http://jamespropp.org/icerm12.pdf