Circumventing Schmidt's bound on discrepancy using tapered estimators

James Propp (UMass Lowell, Tufts University, and Microsoft Research New England)

Institute for
Computational and Experimental Research
in Mathematics,
October 29, 2014

Slides at http://jamespropp.org/discrep.pdf

Note: I don't work in this field, so some of what I've found may be old, or worse. Please let me know!

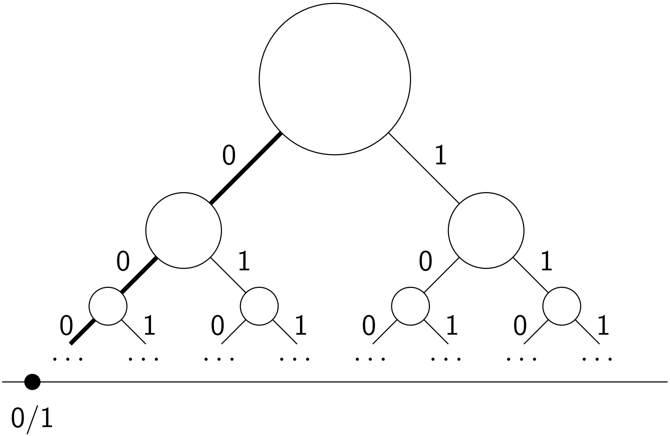
Van der Corput's sequence

Van der Corput's sequence (van der Corput, 1935), with a 0 prepended, is

$$\begin{aligned}
x_0 &= 0 = .0000\ldots_2, \\
x_1 &= 1/2 = .1000\ldots_2, \\
x_2 &= 1/4 = .0100\ldots_2, \\
x_3 &= 3/4 = .1100\ldots_2, \\
x_4 &= 1/8 = .0010\ldots_2, \\
x_5 &= 5/8 = .1010\ldots_2, \\
& \text{etc.}
\end{aligned}$$

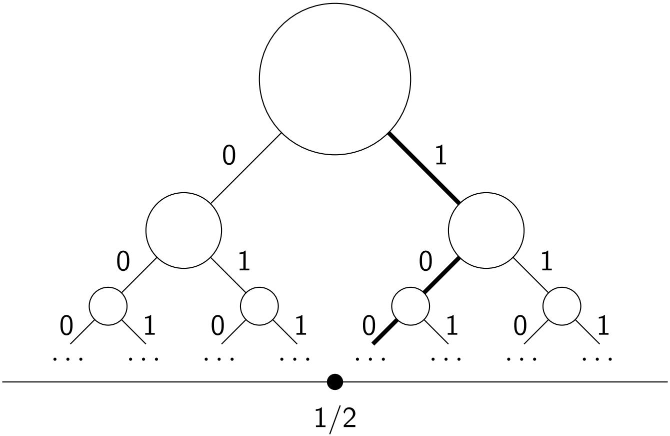
Quasirandom walks in a binary tree

This corresponds to a sequence of quasirandom walks down an infinite binary tree: each particle starts at the root, and the kth particle to leave a vertex goes left if k is odd and right if k is even.



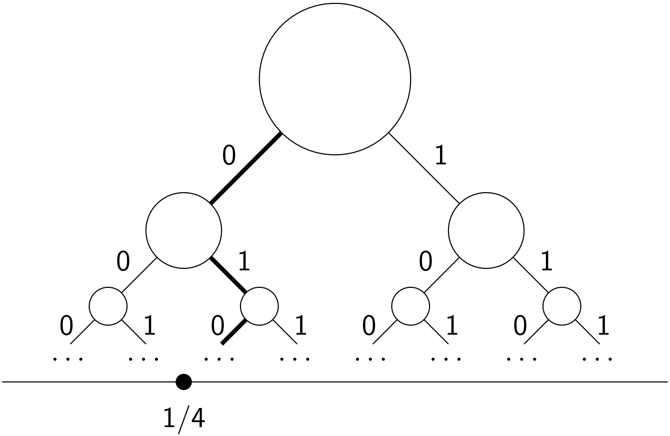
Quasirandom walks in a binary tree

This corresponds to a sequence of quasirandom walks down an infinite binary tree: each particle starts at the root, and the kth particle to leave a vertex goes left if k is odd and right if k is even.



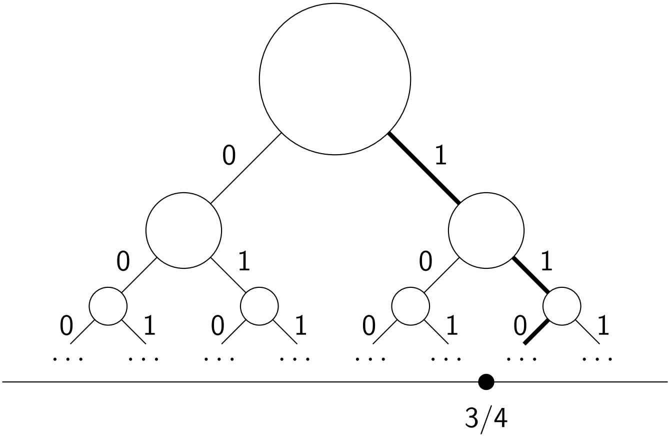
Quasirandom walks in a binary tree

This corresponds to a sequence of quasirandom walks down an infinite binary tree: each particle starts at the root, and the kth particle to leave a vertex goes left if k is odd and right if k is even.



Quasirandom walks in a binary tree

This corresponds to a sequence of quasirandom walks down an infinite binary tree: each particle starts at the root, and the kth particle to leave a vertex goes left if k is odd and right if k is even.



Discrepancy

The sequence $(x_n)_{n\geq 1} = (1/2, 1/4, 3/4, 1/8, 5/8, \ldots)$ is uniformly distributed in $[0, 1)$; specifically, for all α in $[0, 1)$,

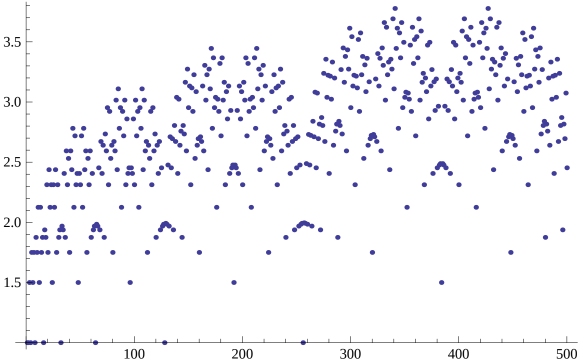
$$d(n, \alpha) := \#(\{1 \leq i \leq n : \ x_i < \alpha\})/n - \alpha = O(\log n \ / \ n).$$

That is,

$$D(n, \alpha) := \#(\{1 \leq i \leq n : \ x_i < \alpha\}) - n\,\alpha = O(\log n).$$

$D(n)$, plotted

Here is a plot of $D(n) = \sup_\alpha D(n, \alpha)$ for $n \leq 500$:



One can "see" that it is increasing without bound.

Can we do better?

In 1935, van der Corput raised the question of whether there could exist a sequence of numbers $x_1, x_2, \ldots$ in $[0, 1)$ such that the discrepancy

$$D(n) = \sup_{\alpha \in [0,1)} \left| \#(\{1 \leq i \leq n : x_i < \alpha\}) - n\,\alpha \right|$$

is bounded as $n \to \infty$.

In 1945, Van Aardenne-Ehrenfest showed that the answer is "no", with later improvements by Roth and by Schmidt showing that van der Corput's sequence is essentially optimal.

Paraphrasing the question

Let f be the indicator function of $[0, \alpha)$, so that $\int_0^1 f(t)\, dt = \alpha$.

Then $\#(\{1 \le i \le n : x_i < \alpha\}) = \sum_{i=1}^n f(x_i)$ so we are estimating the **integral** $\int_0^1 f(t)\, dt$ using the **average** $\frac{1}{n} \sum_{i=1}^n f(x_i)$, in QMC.

E.g., for $\alpha = (-1+\sqrt{5})/2 \approx .618$, our 100th estimate of α is $1/100$ times the sum of the 100 bits

$$\begin{matrix} 11010101110101011101 \\ 01011101010111010101 \\ 11010101110101011101 \\ 01011101010111010101 \\ 11010101110101011101 \end{matrix}$$

These bits are highly patterned (look down the columns).

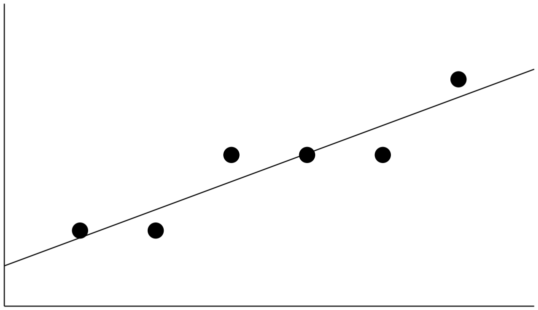
Exploiting patterns

Precisely because the van der Corput sequence is highly patterned, the sequence of bits $f(x_1), f(x_2), \ldots$ is highly patterned, too, and it turns out that there is a linear combination of those bits that agrees **more closely** with α than the unweighted average $(f(x_1) + f(x_2) + \ldots + f(x_n))/n$, namely, the weighted average $(1)(n)f(x_1) + (2)(n-1)f(x_2) + \ldots + (n)(1)f(x_n)$ divided by $(1)(n) + (2)(n-1) + \ldots + (n)(1) = n(n+1)(n+2)/6$.

Question: How well?

But first: Where do those weights come from?

Plot the n points (i, s_i) where $s_i = f(x_1) + ... + f(x_i)$.



$((1)(n)f(x_1)+(2)(n-1)f(x_2)+...+(n)(1)f(x_n))/(n(n+1)(n+2)/6)$ is the slope of the least squares regression line through the points (i, s_i).

A new question

Van der Corput put

$$e(n,\alpha) = \sum_{i=1}^{n} \frac{1}{n} 1[x_i < \alpha],$$

$d(n) = \sup_{\alpha \in [0,1)} |e(\alpha) - \alpha|$, and $D(n) = n\, d(n)$ and asked, Could $D(n)$ stay bounded for suitable $x_1, x_2, \ldots$?

Let's replace the unweighted average $e(n,\alpha)$ by the weighted average

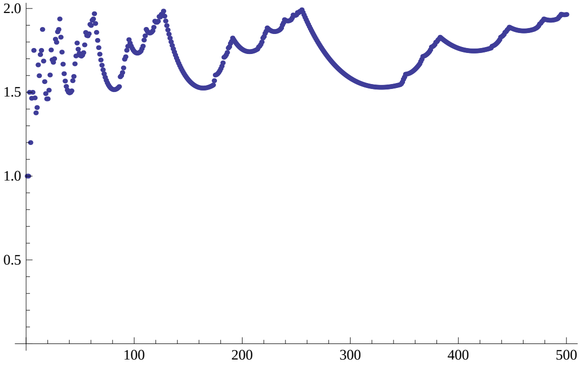
$$e^\sharp(n,\alpha) = \sum_{i=1}^{n} \frac{i(n+1-i)}{n(n+1)(n+2)/6} 1[x_i < \alpha]$$

and put $d^\sharp(n) = \sup_{\alpha \in [0,1)} |e^\sharp(\alpha) - \alpha|$ and $D^\sharp(n) = n\, d^\sharp(n)$.

How big is $D^\sharp(n)$?

Conjecture

For all n, $D^\sharp(n) \leq 2$, and the 2 is best possible; indeed, $\limsup D^\sharp(n) = 2$ and $\liminf D^\sharp(n) \approx 1.5$.



Conjecture, continued

Moreover, for all $n < 2^k - 2$, $D^\sharp(n) < D^\sharp(2^k - 2) = (2^k - 1)/2^{k-1}$ (verified for $k \leq 9$).

$(D^\sharp(n))_{n\geq 1} = (\frac{1}{2}, \mathbf{1}, \frac{3}{4}, 1, \frac{5}{4}, \mathbf{\frac{3}{2}}, \frac{7}{6}, \frac{16}{15}, \frac{9}{8}, \frac{61}{44}, \frac{11}{8}, \frac{3}{2}, \frac{13}{8}, \mathbf{\frac{7}{4}}, \ldots, \mathbf{\frac{15}{8}}, \ldots)$

So the answer to van der Corput's question changes from "no" to "yes"!

(Can we prove this by Friday?)

Why should tapering reduce discrepancy?

A related fact in signal processing:

For any function $f(t)$ expressible as a constant f_0 plus a linear combination of finitely many sinusoids,

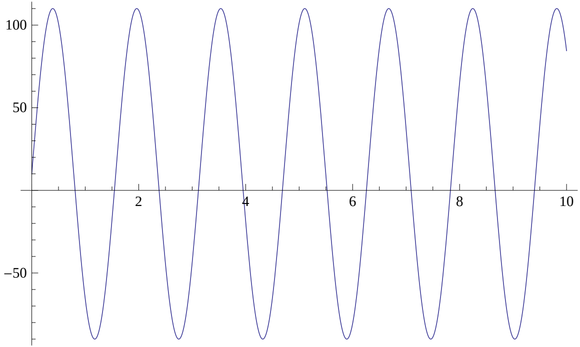
$$\frac{1}{T}\int_0^T f(t)\,dt = f_0 + O(1/T)$$

but

$$\frac{1}{T/6}\int_0^T \left(\frac{t}{T}\right)\left(1-\frac{t}{T}\right) f(t)\,dt = f_0 + O(1/T^2).$$

Unsmoothed averaging

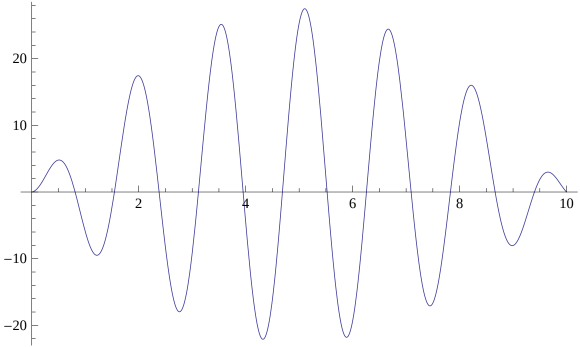
Let $f_0 = 10,\ f(x) = 10 + 100 \sin 4x,\ T = 10$:



$$\frac{1}{10}\int_0^{10} f(t)\,dt \approx 14$$

Smoothed averaging

Let $f_0 = 10,\ f(x) = 10 + 100 \sin 4x,\ T = 10$:



$$\frac{1}{10/6} \int_0^{10} \left(\frac{t}{10}\right) \left(1 - \frac{t}{10}\right) f(t)\, dt \approx 9.8$$

Why it works

By taking a weighted average that assigns high weight near $T/2$ and low weight near 0 and T, we remove the part of the error that's due to truncation of the integral at the endpoints, and we exploit the oscillatory character of the integrand to get self-cancellation in between.

(Where the analogy breaks down:

In the case of estimating the mean value f_0 of an almost periodic function f, we can do even better by using powers of the integration kernel that fall off more quickly at 0 and T.

But for our problem of estimating the integral of the indicator function of $[0, \alpha)$ using the values of this function at quasirandom points, this doesn't work.)

Discrepancy viewed as error of an estimator

I've been exploring a "statistical" viewpoint on discrepancy theory.

Given some discrepancy theorem bounding some quantity $|x - y|$, view it as a problem in estimation, where y is some quantity we're trying to estimate and x is an estimator.

Once you've adopted this point of view, take a look at the building blocks of which x is composed, and ask, Is there a way to combine them that gives a better estimate of y?

Examples

Example 1: Van der Corput's question

Example 2: Estimating the density of a Sturmian sequence, or equivalently the slope of a digital line

Example 3: Estimating the sizes of features in models exhibiting proportionate growth

Example 4: Estimating the density of a quasiperiodic point-packing

Example 5: Estimating π

Example 2: Sturmian sequences

A **Sturmian bit-string** is a sequence of 0's and 1's such that for all n, there exists k such that the sum of any consecutive n consecutive terms of the sequence is k or $k+1$.

A sequence of bits $x_0, x_1, \dots$ is Sturmian iff there exist α, β in $[0,1]$ such that for all but at most one value of n, $x_n = \mathrm{nint}(\alpha n + \beta) - \mathrm{nint}(\alpha(n-1) + \beta)$, where $\mathrm{nint}(x)$ equals the integer nearest to x with $\mathrm{nint}(n + \frac{1}{2}) = n$.

α equals the asymptotic density of 1's.

Random Sturmian sequences

Theorem (Propp): Let A and B be independent random variables uniform in $[0, 1]$, and let

$$X_n = \mathrm{nint}(A(n) + B) - \mathrm{nint}(A(n-1) + B)$$

so that $X_1, X_2, \ldots$ is a random Sturmian sequence of 0's and 1's of random density A.

Then the variance of the discrepancy between A and $((1)(n)X_1 + (2)(n-1)X_2 + \ldots + (n)(1)X_n)/(n(n+1)(n+2)/6)$ equals $1/n(n+1)(n+2)$.

Digital lines

Equivalently, let A and B be independent and uniform in $[0, 1]$, take the random digital line segment consisting of the points (x, y) with x in $\{1, 2, ..., n\}$ and $y = \mathrm{nint}(Ax + B)$, and let A_n be the slope of the least squares approximation to the digital line segment.

Then the mean of the squared difference between A_n and A equals $1/(n-1)(n)(n+1)$.

Can we do better?

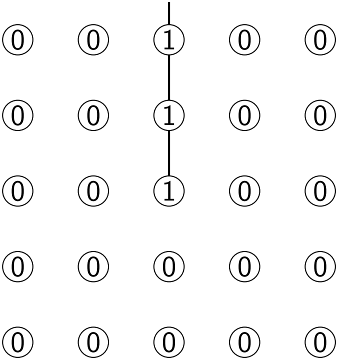
The RMS error $1/n^{3/2}$ is qualitatively better than $1/n$, but can we reconstruct the slope of a digital line even more accurately?

Open question for problem session: How well can we expect to reconstruct A if we use the optimal method (linear programming)? The answer appears to be close to $O(1/n^2)$ (for (A, B) uniform in $[0, 1] \times [0, 1]$).

(Exercise: Use geometry of numbers to show that you can't reconstruct B with error less than $O(1/n)$, and you can't reconstruct A with error less than $O(1/n^2)$.)

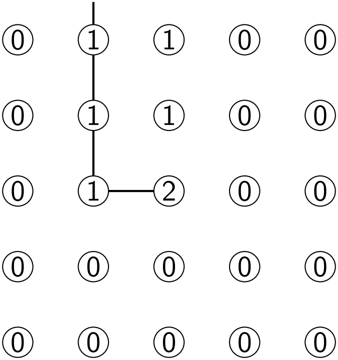
Example 3: Proportionate growth for quasirandom walk

Consider a walk-process on $\mathbb{Z}^2$: Each particle starts at $(0,0)$, and the particle that leaves a vertex for the kth time exits going North, West, South, or East according to whether k is 1, 2, 3, or 0 mod 4.



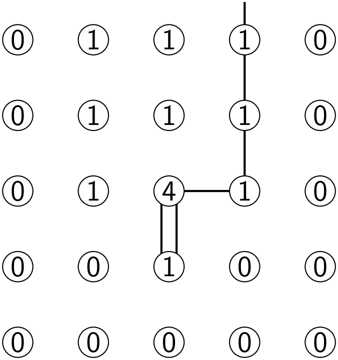
Proportionate growth for quasirandom walks in $\mathbb{Z}^2$

Consider a walk-process on $\mathbb{Z}^2$: Each particle starts at $(0,0)$, and the particle that leaves a vertex for the kth time exits going North, West, South, or East according to whether k is 1, 2, 3, or 0 mod 4.



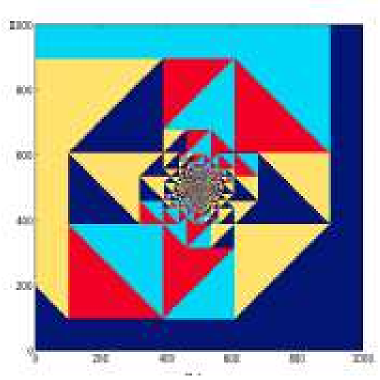
Proportionate growth for quasirandom walks in $\mathbb{Z}^2$

Consider a walk-process on $\mathbb{Z}^2$: Each particle starts at $(0,0)$, and the particle that leaves a vertex for the kth time exits going North, West, South, or East according to whether k is 1, 2, 3, or 0 mod 4.



An interesting continuum limit

If we color each node according to the number of visits it has had mod 4, we see growing structures that, if rescaled, seem to be converging to a continuum limit with fractal structure near the origin; see Figure 1 in `arXiv:1312.6888` by Dandekar and Dhar.



The model appears to be exactly solvable

Dandekar and Dhar have computed what the limit structure should be (though their inventive and compelling analysis makes some unproved assumptions).

E.g., a certain visual feature in the rescaled limit should have size $\frac{3}{4} - \frac{2}{\pi}$.

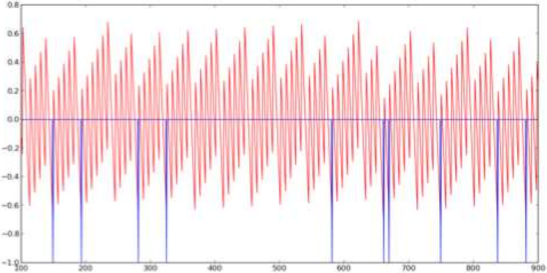
Convergence to the limit

If $f(n)$ denotes the size of such a visual feature after n particles have walked through the system, we expect $f(n)/n$ to converge to $\frac{3}{4} - \frac{2}{\pi}$.

Since $f(n)$ is always an integer, we can't expect the discrepancy between $f(n)/n$ and its limit c to go to zero faster than $1/n$.

Patterns in the scaled-up discrepancy

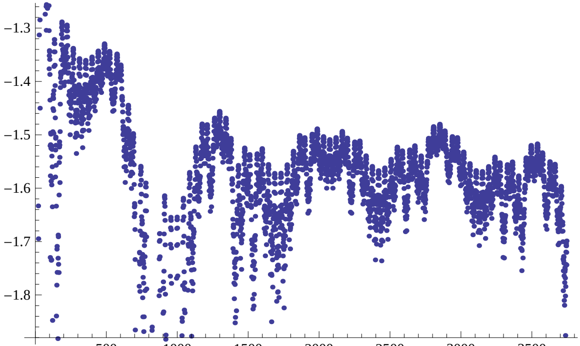
The discrepancy between $f(n)$ and cn exhibits almost-periodicity (see Figure 15 in `arXiv:1312.6888` by Dandekar and Dhar):



So, even though the differences $f(n) - f(n-1)$ aren't Sturmian, we might hope that the smoothing trick we used before will help us.

Exploiting the patterns

Empirically, it seems that if we let $x_n = f(n) - f(n-1)$, then the discrepancy between $((1)(n)x_1 + (2)(n-1)x_2 + ... + (n)(1)x_n)/(n(n+1)(n+2)/6)$ and $\frac{3}{4} - \frac{2}{\pi}$ goes to zero like $n^{-3/2}$.



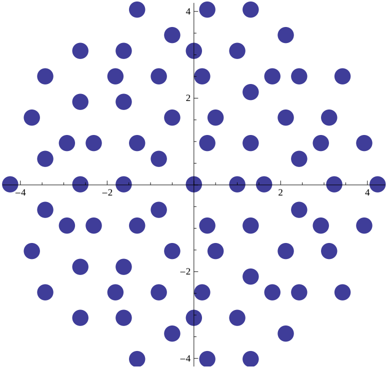
Moral for Quasi Monte Carlo estimation

If your sequence of estimates exhibits patterns, consider exploiting those patterns to derive better estimates.

In particular, don't assume unweighted averages are optimal for Quasi Monte Carlo just because they're optimal for ordinary Monte Carlo!

Example 4: Quasiperiodic point-sets

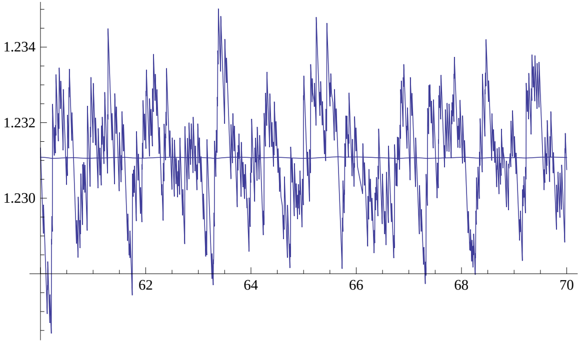
Here is a point-set derived from a Penrose tiling.



How accurately can we estimate the density of the point-set given a finite disk-shaped excerpt?

Ordinary vs. tapered (or “averaged”) averaging

The graph below shows, for various values of R, the estimate using ordinary averaging (i.e., counting the points in a disk of radius R and dividing by the area of the disk) versus the estimate using tapered averaging (where a point at distance r from the center gets weight $R - r$).



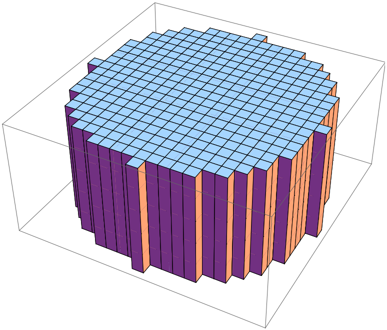
(The true density is $(5 + \sqrt{5})/(\sin \pi/5) \approx 1.231$.)

Example 5: Estimating π

How accurately can we estimate π by drawing a circle of radius $R = \sqrt{n}$ centered at $(0, 0)$ in a square grid?

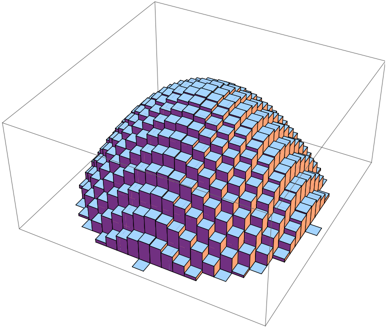
It depends on what we do with the grid-points.

If we just count them (and divide by n), our estimate has error on the order of n^c where $c \approx -.6$ (finding c exactly is Gauss' Circle Problem, still unsolved).



Same grid, smaller discrepancy

If we weight the points so that a point at distance r from the origin gets weight $1 - (r/R)^2$ and divide the sum of the weights by $R^2/2$, our estimate appears to have error on the order of $n^{-1.2}$.



(Is it just a coincidence that -1.2 is about twice $-.6$?)

For more information

Slides for this talk are at http://jamespropp.org/discrep.pdf

Slides for a related talk are at
http://jamespropp.org/tapering.pdf

For information on quasirandom walks, quasirandom Markov chains, quasirandom aggregation, etc., see:
http://jamespropp.org/mcqmc08-01.pdf;
Holroyd and Propp, "Rotor walks and Markov chains" (arXiv:0904.4507);
Kleber, "Goldbug Variations" (arXiv:math/0501497);
Holroyd et al., "Chip-firing and rotor-routing on directed graphs" (arXiv:0801.3306); and
Dandekar and Dhar, "Proportionate growth in patterns formed in the rotor-router model" (arXiv:1312.6888).