Known Results for the Line-Picking Problem

Matthew Roughan Eric Parsonage Jonathon Tuke
School of Mathematical Sciences
University of Adelaide
<{matthew.roughan,eric.parsonage,}@adelaide.edu.au>

August 28, 2012

Abstract

1 Introduction

The *line-picking* problem is a standard problem in stochastic geometry, where we pick lines at random from some region. The typical questions one asks are what will then mean line length be? What will the Probability Density Function (PDF) be?

This brief note describes the current list of known PDFs, and where they were derived, as well as the current set of code to calculate these.

The code is written in C with minimal external dependencies, and with suitable wrapper functions for Matlab and R, to allow it to be run on a wide variety of systems.

Motivation: examples of uses

- Modelling a PNNI hierarchical routing protocol for ATM packet telecommunication networks [1] (length of lines in rectangles);
-

2 Problem Definition

Start with a space Ω, from which we can draw points x at random. Typical examples include a rectangle in R^2, or a hyperball in R^n.

Draw two IID (Independently, Identically Distributed) points from the space, and draw a line between the points. The line could refer to the natural geodesic on the space, or something more complicated such as a geodesic in a higher dimensional Euclidean space in which Ω is embedded. For instance, we might consider simple straight lines between points chosen in a rectangle, or straight lines in R^n between points chosen on the surface of a sphere, or geodesics on the surface of the same sphere.

Another way to frame this is to assume we have a distance metic $d(\cdot, \cdot)$ on the space Ω. The typical distance metric used in these problems is the Euclidean distance, but others are possible.

So a *line-picking problem* consists of 3 components:

- Ω (typically a subset of R^n)
- The measure μ on Ω describing the choice of points (typically uniform)
- The space in which we draw lines (geodesics) and its related distance metric.

so we describe these problems by the triple: (Ω, μ, d). Given that the typical case of μ is uniform, and of d is Euclidean, we often ommit these from the problem statement.

EXAMPLE FIGURES: square, surface of sphere, manhattan distance

Bertrand's Paradox of 1889, "Does a random chord on a circle has length exceeding the length of a side of an inscribed equilateral triangle?"

3 Region Transforms

Many of the standard problems use a uniform point distribution and Euclidean distances on some convex region of R^n. In these cases, there is are easy scaling and trasnlation laws that means once we know the distribution for some size region, we can compute it for any size.

Translation is trivial. Given the uniform distribution of point, translation has no affect on the line-length distribution.

Scaling is almost as simple: if the region Ω is scaled in all dimensions by L, then the problem is identical under a scaling of the distance metric by L. We can therefore scale the density functions as follows

$$g_L(t) = \frac{1}{L} g_1 \left(\frac{t}{L} \right). \tag{1}$$

Non-uniform scaling is not so simple as we shall see in the results below.

4 Known Results

The first result we present is that for the rectangle: given in [2, Theorem 2.4.4] and [3, Theorem 2]

$$g^{\text{rect}}_{a,b}(t) = \frac{4t}{a^2 b^2} \phi_{a,b}(t), \tag{2}$$

where

$$\phi_{a,b}(t) = \begin{cases} \frac{ab\pi}{2} - (a+b)t + \frac{t^2}{2}, & \text{for } t \leq a, \\ ab \sin^{-1}(a/t) - \frac{a^2}{2} - bt + b\sqrt{t^2 - a^2}, & \text{for } a \leq t \leq b, \\ ab \left[\sin^{-1}(a/t) - \sin^{-1} \sqrt{1 - \frac{b^2}{t^2}} \right] - \frac{a^2+b^2+t^2}{2} + a\sqrt{t^2 - b^2} + b\sqrt{t^2 - a^2}, & \text{for } b \leq t \leq \sqrt{a^2 + b^2}, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where the rectangle has sides of length $a \leq b$. Figure 1a shows these for various cases, chosen such that $\sqrt{a^2 + b^2} = 1$ to allow comparison. We label these rectangles by their aspect ratio $a : b$.

This is a rather complicated expression, but is easily evaluated numerically. Naldi [4] approximated this expression with a β function, though given the requirements to numerically evaluate that function there hardly seems any advantage, though we shall see later that this would have been completely appropriate if the region have been a circle.

There are two obvious special cases of the above – the line, and the square – both of which have been addressed separately (e.g., see [5,6]), but which also result from limits of the above formula. The probability density function of distances between two (uniformly) randomly chosen points on the unit square is given in [6], as

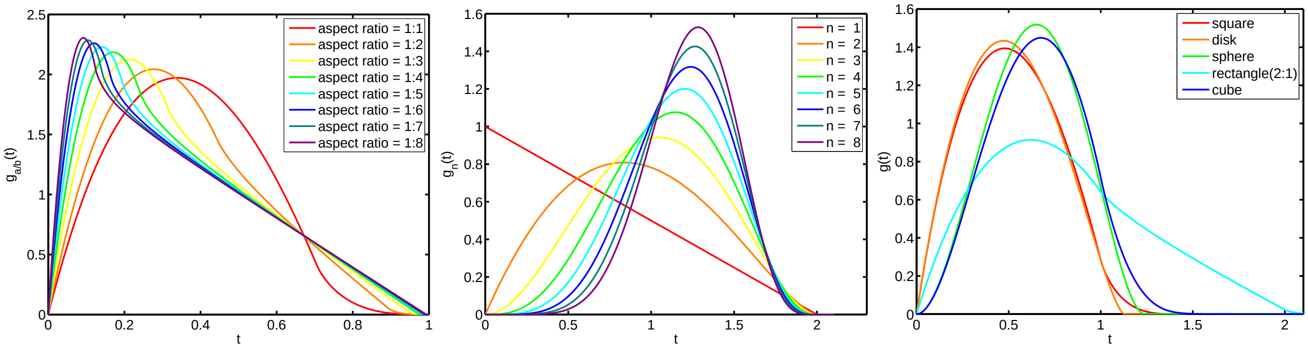
$$g^{\text{square}}(t) = \begin{cases} 2t(t^2 - 4t + \pi), & \text{for } 0 \leq t \leq 1, \\ 2t \left[4\sqrt{t^2 - 1} - (t^2 + 2 - \pi) - 4 \tan^{-1} \left(\sqrt{t^2 - 1} \right) \right], & \text{for } 1 \leq t \leq \sqrt{2}. \end{cases} \tag{4}$$

The probability density function of distances between two (uniformly) randomly chosen points on the unit line is given in [3, 7], as

$$g^{\text{line}}(t) = 2(1 - t), \tag{5}$$

or for a line of length L as

$$g^{\text{line}}_L(t) = \frac{2}{L} \left(1 - \frac{t}{L} \right). \tag{6}$$



(a) Density distribution of rectangles, each chosen with a different aspect ratio a/b, but fixed diagonal distance $\sqrt{a^2+b^2}=1$.

(b) Density distribution of n-D balls.

(c) Various densities distributions for shapes with area fixed to 1.

Figure 1: Example distance densities.

The results have also been extended into 3D, with the probability density function of distances between two (uniformly) randomly chosen points in the unit cube is given in [8, 9], by a yet more complicated, but again easily evaluated formula. Likewise the formula have been calculated for a box (with sides a, b and c) [5] and 4- and 5-Cubes [10]. Other results are also known, for instance the distribution when the points are chosen on the sides of the square (but lines are drawn across it) or faces of a cube! [8], and the distribution of distances between points chosen in two different rectangles [3].

The other obvious region on which to solve the line-picking problem is the ball in n-dimensions [11] (equations (27-31)). For a n-dimensional ball of radius R,

$$g_R^{nD-\mathrm{ball}}(t) = n\frac{t^{n-1}}{R^n} I_x\left(\frac{1}{2}(n+1), \frac{1}{2}\right), \tag{7}$$

where

$$x = 1 - \frac{t^2}{4R^2}, \tag{8}$$

and $I_x(p,q)$ is a *regularized beta function*

$$I_x(p,q) = \frac{B(x;p,q)}{B(p,q)}, \tag{9}$$

where $B(x;p,q)$ is an incomplete beta function, and $B(p,q)$ is a beta function, i.e.,

$$\begin{aligned} B(p,q) &= \int_0^1 t^{p-1}(1-t)^{q-1}\,dt, & (10)\\ B(x;p,q) &= \int_0^x t^{p-1}(1-t)^{q-1}\,dt. & (11)\end{aligned}$$

The first few of these are [11] (P_2 in (5) and (17), P_3 in (9) and (19), P_4 in (18) and P_5 in (20), general even form in (15), general odd form in (16)):

$$\begin{aligned}
g_R^{1D-\mathrm{ball}}(t) &= \frac{1}{R} - \frac{t}{2R}, & (12)\\
g_R^{2D-\mathrm{ball}}(t) &= \frac{4t}{\pi R^2}\cos^{-1}\left(\frac{t}{2R}\right) - \frac{2t^2}{\pi R^3}\sqrt{1-\frac{t^2}{4R^2}}, & (13)\\
&= \frac{2t}{R^2} - \frac{2t^2}{\pi R^3}\sqrt{1-\frac{t^2}{4R^2}} - \frac{4t}{\pi R^2}\sin^{-1}\left(\frac{t}{2R}\right), & (14)\\
g_R^{3D-\mathrm{ball}}(t) &= \frac{3t^2}{R^3} - \frac{9t^3}{4R^4} + \frac{3t^5}{16R^6}, & (15)\\
g_R^{4D-\mathrm{ball}}(t) &= \frac{8t^3}{\pi R^4}\cos^{-1}\left(\frac{t}{2R}\right) - \frac{8t^4}{3\pi R^5}\left(1-\frac{t^2}{4R^2}\right)^{3/2} - \frac{4t^4}{\pi R^5}\sqrt{1-\frac{t^2}{4R^2}} & (16)\\
g_R^{5D-\mathrm{ball}}(t) &= \frac{5t^4}{R^5} - \frac{75t^6}{16R^6} + \frac{25t^7}{32R^8} - \frac{15t^9}{256R^{10}}. & (17)
\end{aligned}$$

Tu and Fischbach [11] also extend these results to cases with non-uniform point distributions.

Figure 1b shows a comparison of line picking on balls of various dimensions, and Figure 1c shows a comparison of the 2D and 3D balls to the square and cube. We can see that as long as the areas (volumes) are matched, they appear quite similar, respectively, though the rectangle varies considerably more.

4.1 Moments

Ghosh [3] gives the first four moments of the line-length distribution for the rectangle.

$$\begin{aligned}
\alpha_1 &= \frac{1}{6}\left[\frac{b^2}{a}\cosh^{-1}(M/b) + \frac{a^2}{b}\cosh^{-1}(M/a)\right] + \frac{1}{15}\left[\frac{a^3}{b^2}+\frac{b^3}{a^2}\right] - \frac{M}{15}\left[\frac{a^2}{b^2}+\frac{b^2}{a^2}-3\right], & (18)\\
\alpha_2 &= \frac{1}{6}M^2, & (19)\\
\alpha_3 &= \frac{1}{20}\left[\frac{b^4}{a}\cosh^{-1}(M/b) + \frac{a^4}{b}\cosh^{-1}(M/a)\right] + \frac{2}{105}\left[\frac{a^5}{b^2}+\frac{b^5}{a^2}\right] - \frac{2M}{105}\left[\frac{a^4}{b^2}+\frac{b^4}{a^2}\right] - \frac{5}{84}M^3, & (20)\\
\alpha_4 &= \frac{1}{15}a^4 + \frac{1}{18}a^2b^2 + \frac{1}{15}b^4, & (21)
\end{aligned}$$

where $M = \sqrt{a^2+b^2}$, from which we can derive the special cases of the square and line (though these can also be derived directly). Obvious central moments such as mean, and variance, etc., can be derived from these, though forumlas will be complex. Rosenberg [1] derives a similar result, but under different assumptions of the random selection of lines (he assumes we first choose a random angle, then choose the line.

The mean for the cube, known as the *Robbins constant*, is given in [9, 12] as

$$\mu^{\mathrm{cube}} = \frac{1}{105}\left[4 + 17\sqrt{2} - 6\sqrt{3} + 21\ln(1+\sqrt{2}) + 42\ln(2+\sqrt{3}) - 7\pi\right] = 0.66170... \tag{22}$$

but a closed form for the variance does not appear (only even moments are reported). Even more complicated results appear for 4- and 5-Cubes in 5-Cubes [10]:

$$\begin{aligned}
\mu^{4-\mathrm{cube}} &= 0.7776656535..., & (23)\\
\mu^{4-\mathrm{cube}} &= 0.8785309152.... & (24)
\end{aligned}$$

The means for the n-dimensional ball (with radius 1) are given in [13] as

$$\begin{aligned}
\mu^{1D-\text{ball}} &= \frac{2}{3}, &(25)\\
\mu^{2D-\text{ball}} &= \frac{128}{45\pi}, &(26)\\
\mu^{3D-\text{ball}} &= \frac{36}{35}, &(27)\\
\mu^{4D-\text{ball}} &= \frac{16384}{4725\pi}. &(28)
\end{aligned}$$

The more general form for higher order moments is given in [11] (equation (138-141)) as

$$\begin{aligned}
\alpha_m^{nD-\text{ball}} &= \frac{n2^{m+n}}{m+n} \frac{B\left(\frac{n+1}{2}, \frac{n+m+1}{2}\right)}{B\left(\frac{n+1}{2}, \frac{1}{2}\right)} R^m, &(29)\\
&= \left(\frac{n}{n+m}\right)^2 \frac{\Gamma\left(n+m+1\right)\Gamma\left(n/2\right)}{\Gamma\left((n+m)/2\right)\Gamma\left(n+1+m/2\right)} R^m. &(30)
\end{aligned}$$

for and n-dimensional ball of radius R.

4.2 CDF

Cumulative distribution functions are also known ...

Can be calculated by integration, but ...

4.2.1 Hyperball

As before

$$g_R^{nD-\text{ball}}(t) = n\frac{t^{n-1}}{R^n} I_x\left(\frac{1}{2}(n+1), \frac{1}{2}\right), \tag{31}$$

where

$$x = 1 - \frac{t^2}{4R^2}, \tag{32}$$

and $I_x(p,q)$ is a *regularized beta function*

$$I_x(p,q) = \frac{B(x;p,q)}{B(p,q)}, \tag{33}$$

where $B(x;p,q)$ is an incomplete beta function, and $B(p,q)$ is a beta function, i.e.,

$$\begin{aligned}
B(p,q) &= \int_0^1 t^{p-1}(1-t)^{q-1}\,dt, &(34)\\
B(x;p,q) &= \int_0^x t^{p-1}(1-t)^{q-1}\,dt. &(35)
\end{aligned}$$

So we need to calculate terms like

$$
\begin{aligned}
\int_0^t \tau^{n-1} B(\tau;p,q)\, d\tau &= \int_0^t \tau^{n-1} \int_0^\tau s^{p-1}(1-s)^{q-1}\, ds\, d\tau \\
&= \int_0^t \int_0^\tau \tau^{n-1} s^{p-1}(1-s)^{q-1}\, ds\, d\tau \\
&= \int_0^t \int_s^t \tau^{n-1} s^{p-1}(1-s)^{q-1}\, d\tau\, ds \\
&= \int_0^t s^{p-1}(1-s)^{q-1} \int_s^t \tau^{n-1}\, d\tau\, ds \\
&= \frac{1}{n} \int_0^t s^{p-1}(1-s)^{q-1} \left[\tau^n\right]_s^t\, ds \\
&= \frac{t^n}{n} \int_0^t s^{p-1}(1-s)^{q-1}\, ds - \frac{1}{n} \int_0^t s^{n+p-1}(1-s)^{q-1}\, ds \\
&= \frac{t^n}{n} B(t;p,q) - \frac{1}{n} B(t;p+n,q). \qquad (36)
\end{aligned}
$$

Now we need adapt this to the above case, where we calculate $g_R^{nD-\text{ball}}(t)$ ...

5 New Results

ERIC todo

6 Approximations

For some problems, it may be interesting to understand the behaviour of these distributions for small t.

6.1 Examples

6.1.1 Line

For the line we need no approximation:

$$g_L^{\text{line}}(t) = \frac{2}{L} - \frac{2}{L^2} t. \qquad (37)$$

6.1.2 Rectangle (and square)

For the rectangle (and square) we need only restrict our attention to the case $t < a$ (the smaller side length), to get a simple polynomial expression:

$$
\begin{aligned}
g_{a,b}^{\text{rect}}(t) &= \frac{4t}{a^2b^2} \left[\frac{ab\pi}{2} - (a+b)t + \frac{t^2}{2} \right], \\
&= \frac{2\pi}{A} t - \frac{2P}{A^2} t^2 + \frac{2}{A^2} t^3, \qquad (38)
\end{aligned}
$$

where $A = ab$ is the area of the rectangle, and $P = a + b$ is the perimeter.

The affect of changing the scale must be obvious, in that we know area scales quadratically, and perimeter linearly with the size of the region, so simple dimensional analysis suggests a form such as that above must occur, but it is perhaps suprising that it is so clean.

6.1.3 Cube (and box)

For the cube [9], for small t,

$$\begin{aligned} g_1^{\mathrm{cube}}(t) &= -t^2\left[(t-8)t^2+\pi(6t-4)\right], \\ &= 4\pi t^2 - 6\pi t^3 + 8t^4 - t^5. \end{aligned} \tag{39}$$

For the box [5] with $t < a \le b \le c$, we first see the distribution of $u = t^2$ (using $g(t) = 2th(t^2)$),

$$\begin{aligned} h^{\mathrm{box}}_{a,b,c}(u) &= \frac{1}{6a^2b^2c^2}\left[-6\pi bcu + 8bu^{3/2} + 12\pi abc\sqrt{u} - 6\pi a(b+c)u + 8(a+c)u^{3/2} - 3u^2\right], \\ g^{\mathrm{box}}_{a,b,c}(t) &= \frac{t}{3a^2b^2c^2}\left[-6\pi bct^2 + 8bt^3 + 12\pi abct - 6\pi a(b+c)t^2 + 8(a+c)t^3 - 3t^4\right], \\ &= \frac{4\pi}{V}t^2 + \frac{\pi}{3V^2}\left[-6bc - 6a(b+c)\right]t^3 + \frac{1}{3V^2}\left[8b + 8(a+c)\right]t^4 + \frac{1}{3V^2}\left[-3\right]t^5, \\ &= \frac{4\pi}{V}t^2 - \frac{2\pi}{V^2}\left[ab+bc+ac\right]t^3 + \frac{8}{3V^2}\left[a+b+c\right]t^4 - \frac{1}{V^2}t^5, \\ &= \frac{4\pi}{V}t^2 - \frac{\pi S}{V^2}t^3 + \frac{2P}{3V^2}t^4 - \frac{1}{V^2}t^5, \end{aligned} \tag{40}$$

for volume $V = abc$, surface area $S = 2(ab + bc + ac)$ and edge perimeter $P = 4(a + b + c)$¿

6.1.4 Hyperball

For the n-dimensional hyperball, simple approximations are available by use of Taylor series. Ignoring the 1D case (which is the same as the line), we see the 2D disk from [11, 14]:

$$g_R^{2D-\mathrm{ball}}(t) = \frac{4t}{\pi R^2}\cos^{-1}\left(\frac{t}{2R}\right) - \frac{2t^2}{\pi R^3}\sqrt{1-\frac{t^2}{4R^2}}$$

and we use

$$\begin{aligned} \cos^{-1}(x) &= \frac{1}{2}\pi - x - \frac{1}{6}x^3 + \cdots \\ \sqrt{1-x^2} &= 1 - \frac{x^2}{2} - \frac{x^4}{8} + \cdots. \end{aligned}$$

to derive

$$\begin{aligned} g_R^{2D-\mathrm{ball}}(t) &= \frac{4t}{\pi R^2}\cos^{-1}\left(\frac{t}{2R}\right) - \frac{2t^2}{\pi R^3}\sqrt{1-\frac{t^2}{4R^2}} \\ &= \frac{4t}{\pi R^2}\left(\frac{1}{2}\pi - \frac{t}{2R}\right) - \frac{2t^2}{\pi R^3} + O(t^4) \\ &= \frac{2t}{R^2} - \frac{4t^2}{\pi R^3} + O(t^4) \\ &= \frac{2\pi}{A}t - \frac{2P}{A^2}t^2 + O(t^4) \end{aligned} \tag{41}$$

where once again $A = \pi R^2$ and $P = 2\pi R$ are area and perimeter, respectively. We note that this is very similar to the formula obtained for the square for small t. The first two terms are, in fact, identical.

For the 3D ball we already have the expression as a polynomial:

$$\begin{aligned} g_R^{3D-\mathrm{ball}}(t) &= \frac{3t^2}{R^3} - \frac{9t^3}{4R^4} + \frac{3t^5}{16R^6}, \\ &= \frac{4\pi}{V}t^2 - \frac{\pi S}{V^2}t^3 + \frac{3}{16R^6}t^5, \end{aligned} \tag{42}$$

where again $V = 4\pi R^3/3$ and $S = 4\pi R^2$ are the volume and surface area. Interestingly, these are identical to those terms for the cube, and even the forth order term is the same if we say that the sphere has zero edges.

The more general expression for the n-D ball can be derived by noting that for a n-dimensional ball of radius R,

$$g_R^{nD-\mathrm{ball}}(t) = n\frac{t^{n-1}}{R^n} I_x\left(\frac{1}{2}(n+1), \frac{1}{2}\right), \tag{43}$$

rememeber $x = 1 - t^2/4R^2$, $a = (n+1)/2$, and $b = 1/2$, and from [15, 26.5.4]

$$\begin{aligned} I_x(a,b) &= 1 - I_{1-x}(b,a) \\ &= 1 - \frac{(1-x)^b x^a}{bB(b,a)}\left\{1 + \sum_{i=0}^{\infty}\frac{B(b+1,i+1)}{B(a+b,i+1)}(1-x)^{i+1}\right\} \\ &= 1 - \frac{2(t/2R)(1-t^2/4R^2)^a}{B(b,a)}\left\{1 + \sum_{i=0}^{\infty}\frac{B(b+1,i+1)}{B(a+b,i+1)}(t/2R)^{2(i+1)}\right\} \\ &\simeq 1 - \frac{t/R}{B(1/2,(n+1)/2)}. \end{aligned} \tag{44}$$

The Gamma function satisfies [15, 6.1.12]

$$\Gamma(1/2) = \pi^{1/2}, \tag{45}$$

$$\Gamma(3/2) = \frac{1}{2}\pi^{1/2}, \tag{46}$$

$$\Gamma(k+1) = k!, \tag{47}$$

$$\Gamma(k+1/2) = \frac{(2k)!\pi^{1/2}}{2^{2k}k!}, \tag{48}$$

$$\tag{49}$$

so we can derive

$$\begin{aligned} B(1/2,(n+1)/2) &= \frac{\Gamma(1/2)\Gamma((n+1)/2)}{\Gamma(n/2+1)} \\ &= \begin{cases} \dfrac{\pi^{1/2}\Gamma(k)}{\Gamma(k+1/2)} & n \text{ odd, i.e., } n = 2k-1, \\ \dfrac{\pi^{1/2}\Gamma(k+1/2)}{\Gamma(k+1)} & n \text{ even, i.e., } n = 2k. \end{cases} \\ &= \begin{cases} \dfrac{\pi^{1/2}2^{2k}k!(k-1)!}{(2k)!\pi^{1/2}} & n \text{ odd, i.e., } n = 2k-1, \\ \dfrac{\pi^{1/2}(2k)!\pi^{1/2}}{2^{2k}k!k!} & n \text{ even, i.e., } n = 2k. \end{cases} \\ &= \begin{cases} \dfrac{2^{2k}k!(k-1)!}{(2k)!} & n \text{ odd, i.e., } n = 2k-1, \\ \dfrac{\pi(2k)!}{2^{2k}k!k!} & n \text{ even, i.e., } n = 2k. \end{cases} \end{aligned} \tag{50}$$

Thus the asymptotic form of these distributions for $t \to 0$, i.e.,

$$g_R^{nD-\mathrm{ball}}(t) = n\frac{t^{n-1}}{R^n} + n\frac{t^n}{R^{n+1}}\left\{\begin{array}{ll} \dfrac{(2k)!}{2^{2k}k!(k-1)!} & \text{for } n = 2k-1, \\ \dfrac{2^{2k}k!k!}{\pi(2k)!} & \text{for } n = 2k. \end{array}\right\} + O(t^{n+1}). \tag{51}$$

For example: when $n = 1, 2, 3$ and 4:

$$\begin{aligned} g_R^{1D-\mathrm{ball}}(t) &\simeq n\frac{t^{n-1}}{R^n} - n\frac{t^n}{R^{n+1}}\left\{\frac{(2k)!}{2^{2k}k!(k-1)!}\right\} \\ &\simeq \frac{1}{R} - \frac{1}{2}\frac{t}{R^2} \qquad (52) \\ g_R^{2D-\mathrm{ball}}(t) &\simeq n\frac{t^{n-1}}{R^n} - n\frac{t^n}{R^{n+1}}\left\{\frac{2^{2k}k!k!}{\pi(2k)!}\right\} \\ &\simeq 2\frac{t^1}{R^2} - \frac{4}{\pi}\frac{t^2}{R^3} \qquad (53) \\ g_R^{3D-\mathrm{ball}}(t) &\simeq n\frac{t^{n-1}}{R^n} - n\frac{t^n}{R^{n+1}}\left\{\frac{(2k)!}{2^{2k}k!(k-1)!}\right\} \\ &\simeq \frac{3t^2}{R^3} - \frac{9}{4}\frac{t^3}{R^4} \qquad (54) \\ g_R^{4D-\mathrm{ball}}(t) &\simeq n\frac{t^{n-1}}{R^n} - n\frac{t^n}{R^{n+1}}\left\{\frac{2^{2k}k!k!}{\pi(2k)!}\right\} \\ &\simeq \frac{4t^3}{R^4} - \frac{2^5}{3\pi}\frac{t^4}{R^5}. \qquad (55) \end{aligned}$$

which obviously agree with the previous formula.

The generalized volume and surface are of the n-D hyperball are

$$\begin{aligned} V_n(R) &= C_n R^n, \qquad (56) \\ S_n(R) &= \frac{dV_n}{dR} = nC_nR^{n-1}. \qquad (57) \end{aligned}$$

where

$$C_n = \frac{\pi^{n/2}}{\Gamma(n/2+1)}, \qquad (58)$$

so we can write

$$\begin{aligned} g_R^{nD-\mathrm{ball}}(t) &\simeq \frac{nC_n}{V_n}t^{n-1} - \frac{C_nS_n}{B(1/2,(n+1)/2)V_n^2}t^n \\ &\simeq \frac{nC_n}{V_n}t^{n-1} - \frac{\pi^{n/2}}{\Gamma(n/2+1)}\frac{\Gamma(n/2+1)}{\Gamma(1/2)\Gamma((n+1)/2)}\frac{S_n}{V_n^2}t^n \\ &\simeq \frac{n\pi^{n/2}}{\Gamma(n/2+1)V_n}t^{n-1} - \frac{\pi^{(n-1)/2}}{\Gamma((n+1)/2)}\frac{S_n}{V_n^2}t^n. \qquad (59) \end{aligned}$$

6.2 General pattern

It seems as if there is a general pattern here. That makes a good deal of sense because for small t, the boundaries have little affect. Hence, we would expect the shape of the region to only affect the (small) line lengths through macro-properties such as the area and perimeter.

6.2.1 2D case

Formally, consider a convex region in 2D, for t small, we might first assume that a random point is unlikely to be within distance t of the boundary of the region, and so the probability density function, to first order, is simply the probability that the second point lies on the circle around this of radius t.

$$g(t) \sim \frac{2\pi}{A}t, \qquad (60)$$

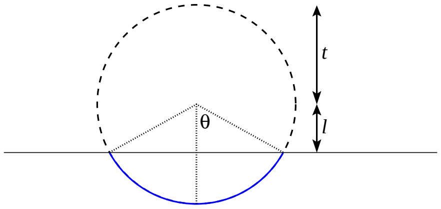


Figure 2: .

We can obtain a more accurate approximation by again assuming t is small, so that we can approximate the boundary by a straight line. In this case, imagine the initial point was chosen (at random) to lie with distance $t/2$ of the boundary. We can see that we over-estimate the possible second points that lie at distance t. We can correct by subtracting the number of such points that would actuall lie outside the region.

Formally, in 2D, consider the first point lies at distance $\ell < t$ from the boundary, then there will be an arc of the circle of radius t that lies outside the circle (see Figure 2). Then the length of the arc in question will be

$$s = 2\theta = 2\cos^{-1}(\ell/t). \tag{61}$$

We then have to integrate over the possible distances ℓ, i.e.,

$$\begin{aligned} h(t) &= \int_0^t 2\cos^{-1}(\ell/t)\, d\ell \\ &= -2t \int_{\pi/2}^0 \theta \sin\theta \, d\theta \\ &= 2t \int_0^{\pi/2} \theta \sin\theta \, d\theta \\ &= 2t \left\{ \left[-\theta\cos\theta\right]_0^{\pi/2} + \int_0^{\pi/2} \cos\theta \, d\theta \right\} \\ &= 2t \left[\sin\theta\right]_0^{\pi/2} \\ &= 2t, \end{aligned} \tag{62}$$

using the substitution $\theta = \cos^{-1}(\ell/t)$, or $\ell = t\cos(\theta)$, so that $d\ell = -t\sin\theta$, and we integate by parts. This must be multiplied by the probability that the first point lies in the boundary region — the area of the boundary divided by the total boundary, i.e., tP/A, and divided by A again to get the probability that the second point would lie on the arc, if there were no boundary, so the corrected form of $g(t)$ in 2D is

$$g(t) \sim \frac{2\pi}{A} t - \frac{2P}{A} t^2, \tag{63}$$

which again matches the results for the rectangle and disk.

The next order correction would then build in the fact that the border is not straight, e.g., in the case of a disk it is round, and in the case of the square it has corners. However, as this term is very dependent on shape and not general characteristics we won't calculate it here.

6.2.2 3D case

Likewise for a region of 3D Euclidean space, we look at the probability that the point lies on the surface of a sphere of radius t, i.e.,

$$g(t) \sim \frac{4\pi}{V} t^2, \tag{64}$$

We can naturally generalize to n-D Euclidean spaces, by noting that the generalized surface are of the n-D hyperball is

$$S_n = \frac{dV_n}{dR} = nC_n R^{n-1}. \tag{65}$$

where V_n is its volume, and C_n is defined in (**??**), so the first order term of any expansion (for small t) of the density function will take the form

$$g(t) \sim \frac{nC_n}{V} t^{n-1}, \tag{66}$$

where V is the n-D volume of the region of interest.

However, calculating the integral for the next term (correcting for the perimeter would be tiresome, and in any case will be derived below.

6.2.3 Higher dimensions

We can go through exactly the same process in higher dimensions, to obtain the same types of approximations, but computing the integrals is somewhat redundant as we can see the correct forms of approximatiosn directly from the hyperball distribution. Intuitively, in high dimensions, the points are spread "more thinly" in the sense that their density on n-dimensional volumes will be lower, and hence, the distances between points will be larger. In particular the chance of very short lines decreases, and we see this in the fact that the first non-zero term in the Taylor series above has order $n-1$ for the n-dimensional problem.

By the previous argument, we know that the result for the ball should extend to arbitrary shapes, where S_n is the general form of the surface area, and V_n the arbitrary form for the volume, so we could for instance estimate the formula for small t for a 4-rectangle with sizes a, b, c and d as

$$\begin{aligned} g_{a,b,c,d}^{4D-\text{rectangle}}(t) &\simeq \frac{nC_n}{V_n} t^{n-1} - \frac{\pi^{(n-1)/2}}{\Gamma((n+1)/2)} \frac{S_n}{V_n^2} t^n \\ &\simeq \frac{2\pi^2}{abcd} t^3 - \frac{8\pi(abc+abd+acd+bcd)}{3(abcd)^2} t^4. \end{aligned} \tag{67}$$

6.3 Non-Euclidean Problems

We can extend some of this insight into other problems, for instance, the *circle line picking* problem, where pairs of points are chosen on a circle, but the lines cross the circle. Here, the probability distribution for line length (with a unit circle) is [14]

$$g(t) = \frac{1}{\pi}\left(1 - \frac{s^2}{4}\right)^{-1/2} \simeq \frac{1}{\pi}\left(1 - \frac{s^2}{4} + \cdots\right) \tag{68}$$

whereas in *sphere line picking* [16], there distribution takes the form

$$g(t) = \frac{1}{2} t. \tag{69}$$

Clearly in this type of case, the dimension of the space is not the critical factor, because the space on which the points are chosen is embedded in a larger space from which lines are chosen, and the geometry of the relationship is important.

INCIDENTALLY – limit n -¿ infty for balls, has almost fixed distances between nodes, so it approaches ER graph

7 Numerical Computation by Simulation

It may be possible that one wishes to compute distributions, on irregular regions, for which there is no closed form solution. Numerically it is straight-forward to calculate the function $g(t)$. There are two obvious approaches:

- Numerical computation of a n-dimensional integral over Ω, or
- Simulation of the problem, and estimation of the density from simulated results.

The two approaches have different advantages and disadvantages. The former approach has no stochastic component, and so errors are predictable and regular.

The later approach allows complex, potentially non-convex, non-uniform, problems to be solved as long as they can be simulated. Given the stochastic nature of the latter, it may help to say a little more:

The general process is as follows:

1. Simulate a set of $2N$ points in the region of interest, and calculate the distances between successive pairs. The region may be irregular, or even non-convex; decisions may be made about some lines being inadmissable (because, for instance, they are exterior to the region for a non-convex region), or distances may be non-Euclidean, or the point distribution can be non-uniform. All that is needed is a set of output distances $\{t_i\}_{i=1}^N$.

2. The density could then be approximate through binning, or a kernal smoothing technique, but in fact, we don't need direct access to the density as the estimator uses the Laplace transform.

We have tested the above approach, running it 30 times (with different seeds), in Matlab for various values of N. The results are shown in Figure **??**. The first plot shows estimates of the mean relative absolute error of the estimated Laplace transforms over the range $S \in [0, 50]$. We can see from the fitted straight line, that the errors decrease as $1/\sqrt{N}$, dropping to around 1% at around $N = 100,000$.

....

The second plot shows the computation times[1] relative to the computation times for the "exact" method[2] We can immediately notice that computation times are roughly linear in N, as one might expect. that around the range $N = 100,000$, the simulation approach is competitive with the exact approach.

The simulation-based approach is not as accurate as the exact numerical approach, however, it accuracy should be sufficient for most estimation problems, without increasing the computational workload unduly. make

8 Programs

8.1 A Rough Guide

The code is arranged to be usable as

1. Directly, as a command-line function;

2. By linking into a larger set of code;

3. Called through a `Matlab` MEX wrapper; or

4. Called through a `R` wrapper function.

It is designed to be as independent of external libraries as possible, needing only the C standard libraries. So compilation should be straight forward on the majority of machines.

Ideally, typing `make` in the top level directory should make all of the targets, however, `R` users may find it easier to install using standard `R` installation procedures (but not that these won't necessarily construct the other components, need for instance for Matlab).

The makefiles in the subdirectories are named `gMakefile` to avoid conflicts with the way `R` interprets them, so if you wish to remake a specific subdirectory, enter the directory and type: `make -f gMakefile`.

There are a large number of functions defined in the code, for each of the cases discussed above, however, there are a small set of functions that you may need to be aware of, that allow one to call all of the others through a simple, uniform interface.

....

8.2 Numerical Issues

Most of the computations in the code involve simple calculations, with no obvious numerical issues (other than the obvious fact that floating point arithmetic is being used).

[1] Both algorithms were implemented in Matlab, the exact method using Matlab's `quadqk` function.

[2] Note that both techniques are in some respect numerical, because even when we have a closed form solution for the density, we still typically need to numerically integrate this to obtain the Laplace transform, but we shall refer to this solution as "exact" for the sake of clarity in the following results, and because in the following we perform numerical integration with error tolerances of 10^{-6}, which means the errors in this approach are significantly smaller than those of the simulation-based approach, at least for the ranges of N tested here.

The computations on the n-D ball, however, require calculation of the incomplete beta function. We have provided a separate library to perform this computation, but users may find they can obtain more accurate results using third party library functions.

...

Estimates of errors ...

8.3 Tests

The tools come with a set of tests to compare performance on your system with ours, and ensure everything is working ...

9 Correlations

Correlations between distances [17]

(0) correlation between a pair 1/10

(i) n nodes, then $N = n(n-1)/2$ pairs of nodes, and so this many pairs of distances

(ii) the $N(N-1)/2 = n(n-1)(n(n-1)-1)/8 =$ possible pairs of correlations

(iii) but only $n(n-1)(n-2)/2$ of the correlations are positive, because they share a node so we get average correlation between all pairs

$$\frac{1}{10}\frac{n(n-1)(n-2)/2}{n(n-1)(n(n-1)-1)/8} = \frac{2}{5}\frac{(n-2)}{(n(n-1)-1)} \simeq \frac{2}{5n}$$

for large n

Empirical measurement (see triples.m)

$$r = 0.114865 \pm 0.000037$$

10 Conclusion and Future Work

References