# 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 23, 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.

## 2 Problem Definition

Start with a space $\Omega$, 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 $\Omega$ 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 $\Omega$. 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:

- $\Omega$ (typically a subset of $R^n$)
- The measure $\mu$ on $\Omega$ 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: $(\Omega, \mu, d)$. Given that the typical case of $\mu$ is uniform, and of $d$ is Euclidean, we often ommit these from the problem statement.

EXAMPLE FIGURES: square, surface of sphere, manhattan distance

# 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 $\Omega$ 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 [**?**, Theorem 2.4.4] and [1, Theorem 2]

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

where

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

where the rectangle has sides of length $a \le b$. 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 [2] approximated this expression with a $\beta$ 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 [3, 4]), 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 [4], as

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

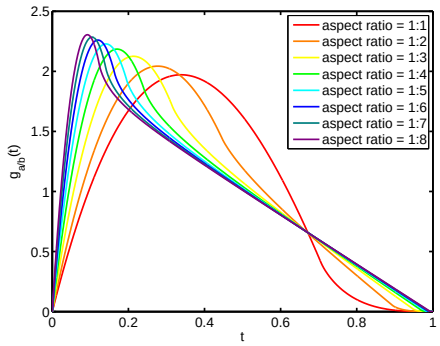
The probability density function of distances between two (uniformly) randomly chosen points on the unit line is given in [5, 1], as

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

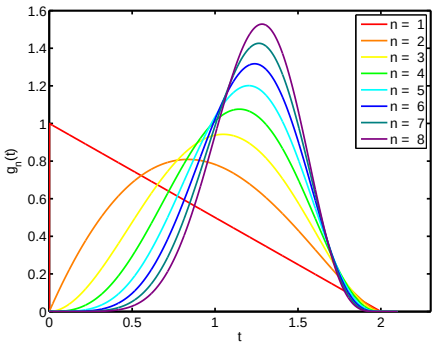
or for a line of length $L$ as

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

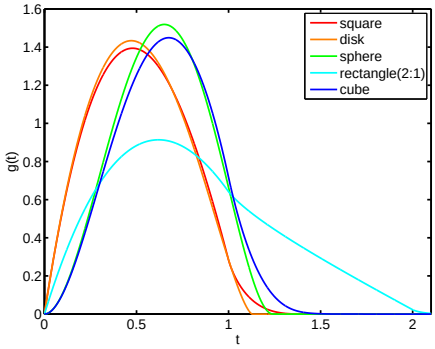
The results have also been extended into 3D, with the probability density function of distances between two (uniformly) randomly chosen points in the unit cube is given in [6, 7], by a yet more complicated, but again easily evaluated formula. Likewise the formula have been calculated for a box (with sides $a, b$ and $c$) [3] and 4- and 5-Cubes [8]. 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![6], and the distribution of distances between points chosen in two different rectangles [1].




(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 other obvious region on which to solve the line-picking problem is the ball in $n$-dimensions [9] (equations (27-31)). For a $n$-dimensional ball of radius $R$,

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

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.,

$$B(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1}\,dt, \tag{10}$$

$$B(x;p,q) = \int_0^x t^{p-1}(1-t)^{q-1}\,dt. \tag{11}$$

The first few of these are [9] ($P_2$ 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)):

$$g_R^{1D-\text{ball}}(t) = \frac{1}{R} - \frac{t}{2R}, \tag{12}$$

$$g_R^{2D-\text{ball}}(t) = \frac{4t}{\pi R^2}\cos^{-1}\left(\frac{t}{2R}\right) - \frac{2t^2}{\pi R^3}\sqrt{1-\frac{t^2}{4R^2}}, \tag{13}$$

$$= \frac{2t}{R^2} - \frac{2t^2}{\pi R^3}\sqrt{1-\frac{t^2}{4R^2}} - \frac{4t}{\pi R^2}\sin^{-1}\left(\frac{t}{2R}\right), \tag{14}$$

$$g_R^{3D-\text{ball}}(t) = \frac{3t^2}{R^3} - \frac{9t^3}{4R^4} + \frac{3t^5}{16R^6}, \tag{15}$$

$$g_R^{4D-\text{ball}}(t) = \frac{8t^3}{\pi R^4}\cos^{-1}\left(\frac{t}{2R}\right) - \frac{8t^4}{3\pi R^5}\left(1-\frac{t^2}{4R^2}\right)^{3/2} - \frac{4t^4}{\pi R^5}\sqrt{1-\frac{t^2}{4R^2}} \tag{16}$$

$$g_R^{5D-\text{ball}}(t) = \frac{5t^4}{R^5} - \frac{75t^6}{16R^6} + \frac{25t^7}{32R^8} - \frac{15t^9}{256R^{10}}. \tag{17}$$

Tu and Fischbach [9] 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 [1] gives the first four moments of the line-length distribution for the rectangle.

$$\alpha_1 = \frac{1}{6}\left[\frac{b^2}{a}\cosh^{-1}(M/b) + \frac{a^2}{b}\cosh^{-1}(M/a)\right] + \frac{1}{15}\left[\frac{a^3}{b^2}+\frac{b^3}{a^2}\right] - \frac{M}{15}\left[\frac{a^2}{b^2}+\frac{b^2}{a^2}-3\right], \tag{18}$$

$$\alpha_2 = \frac{1}{6}M^2, \tag{19}$$

$$\alpha_3 = \frac{1}{20}\left[\frac{b^4}{a}\cosh^{-1}(M/b) + \frac{a^4}{b}\cosh^{-1}(M/a)\right] + \frac{2}{105}\left[\frac{a^5}{b^2}+\frac{b^5}{a^2}\right] - \frac{2M}{105}\left[\frac{a^4}{b^2}+\frac{b^4}{a^2}\right] - \frac{5}{84}M^3, \tag{20}$$

$$\alpha_4 = \frac{1}{15}a^4 + \frac{1}{18}a^2b^2 + \frac{1}{15}b^4, \tag{21}$$

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.

The mean for the cube, known as the *Robbins constant*, is given in [10, 7] as

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

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 [8]:

$$\mu^{4-\text{cube}} = 0.7776656535..., \tag{23}$$

$$\mu^{4-\text{cube}} = 0.8785309152.... \tag{24}$$

The means for the $n$-dimensional ball (with radius 1) are given in [11] as

$$\mu^{1D-\text{ball}} = \frac{2}{3}, \tag{25}$$

$$\mu^{2D-\text{ball}} = \frac{128}{45\pi}, \tag{26}$$

$$\mu^{3D-\text{ball}} = \frac{36}{35}, \tag{27}$$

$$\mu^{4D-\text{ball}} = \frac{16384}{4725\pi}. \tag{28}$$

The more general form for higher order moments is given in [9] (equation (138-141)) as

$$\alpha_m^{nD-\text{ball}} = \frac{n2^{m+n}}{m+n}\frac{B\left(\frac{n+1}{2},\frac{n+m+1}{2}\right)}{B\left(\frac{n+1}{2},\frac{1}{2}\right)}R^m, \tag{29}$$

$$= \left(\frac{n}{n+m}\right)^2 \frac{\Gamma(n+m+1)\,\Gamma(n/2)}{\Gamma((n+m)/2)\,\Gamma(n+1+m/2)}R^m. \tag{30}$$

for and $n$-dimensional ball of radius $R$.

## 4.2 CDF

Cumulative distribution functions are also known ...

Can be calculated by integration, but ...

## 4.3 Approximations

Note that once again we can derive a small $t$ approximation, namely,

$$g_{a,b}(t) \simeq \frac{4t}{a^2b^2}\frac{ab\pi}{2} = \frac{2\pi t}{ab}$$

though here we must assume that the scale-length implied by $\alpha$ is smaller than $a$, the minimum dimension of the rectangle. Once again, the approximation is the same as that of the rectangle and disk, with respect to unit area.

The universality of the small $t$ expansions (used for deriving small $\alpha$ approximations to the length distributiohn) suggests a general form

$$g(t) \simeq \frac{2\pi t}{A} = 2\pi d,$$

where $A$ is the area of the region of interest and $d$ is the scaled distance $d = t/A$. Notice that this scaling is different from the one used in the Waxman distance function where we scale by the maximum distnace $L$.

It is also numerically straight-forward to compute $g(t)$ where the points are placed non-uniformly in the region of interest. Some analytic cases are treated in [9], but for instance dealing with an inhomogeneous Poisson process used to model "burstiness" or clustering of points, and thence with a somewhat different structure than the standard Waxman graph.

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

$$g_n(t) \simeq n\frac{t^{n-1}}{R^n} \tag{31}$$

because we can write

$$\begin{aligned}
I_x(a,b) &= \sum_{j=a}^{a+b-1} \frac{(a+b-1)!}{j!(a+b-1-j)!}x^j(1-x)^{a+b-1-j} \\
&= \sum_{j=a}^{a+b-1} \frac{(a+b-1)!}{j!(a+b-1-j)!}\left(1-\frac{t^2}{4R^2}\right)^j\left(\frac{t^2}{4R^2}\right)^{a+b-1-j} \\
&= \left(1-\frac{t^2}{4R^2}\right)^{a+b-1} + \sum_{j=a}^{a+b-2} \frac{(a+b-1)!}{j!(a+b-1-j)!}\left(1-\frac{t^2}{4R^2}\right)^j\left(\frac{t^2}{4R^2}\right)^{a+b-1-j} \\
&= 1 + O(t),
\end{aligned}$$

and this is directly useful for calculating large $s$ (small $\alpha$) approximations of the estimator.

For instance, from before, we know that if the $j$th derivative of $g(t)$ at $t = 0$ is the first non-zero derivative, then the ratio of Laplace transforms takes the form $\sim (j+1)/s$, so here, it is immediately obvious that

$$-G_n(s)/G_n(s) \to n/s,$$

for large $s$.

This result is now obvious for the sphere, but for non-spherical (but convex) regions, we simply need to consider $s$ large enough that the probability of links longer than $\epsilon$ is negligable, where $\epsilon$ is chosen so that the chance of intersecting with a boundary is negligable, and so we can use a spherical approximation. For instance, the *cube line picking* problem has probability distribution given by [7], which has first non-zero term of $O(t)$, which is correct for a 2D space.

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 [12]

$$g(t) = \frac{1}{\pi}\left(1-\frac{s^2}{4}\right)^{-1/2} \simeq \frac{1}{\pi}\left(1-\frac{s^2}{4}+\cdots\right) \tag{32}$$

whereas in *sphere line picking* [13], there distribution takes the form

$$g(t) = \frac{1}{2}t. \tag{33}$$

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

# 5 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 $\Omega$, 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

# 6 Programs

## 6.1 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).

[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 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.

....

### 6.2 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).

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 ...

### 6.3 Tests

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

## 7 Correlations

Correlations between distances [14]

(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)

$$r = 0.114865 \pm 0.000037$$

## 8 Conclusion and Future Work

## References

[1] B.Ghosh, "Random distance within a rectangle and between two rectangles," *Bulletin of the Calcutta Mathematical Society*, vol. 43, no. 1, pp. 17–24, 1951.

[2] M.Naldi, "Connectivity of Waxman graphs," *Computer Communications*, vol. 29, pp. 24–31, 2005.

[3] J. Philip, "The probability distribution of the distance between two random points in a box." `http://www.math.kth.se/~johanph/hadc.pdf`.

[4] E. W. Weisstein, "Square line picking." From MathWorld–A Wolfram Web Resource. `http://mathworld.wolfram.com/SquareLinePicking.html`.

[5] E. W. Weisstein, "Line line picking." From MathWorld–A Wolfram Web Resource. `http://mathworld.wolfram.com/LineLinePicking.html`.

[6] A. Mathai, P. Moschopoulos, and G. Pederzoli, "Distance between random points in a cube," *Statistica*, vol. 59, no. 1, pp. 61–81, 1999.

[7] E. W. Weisstein, "Cube line picking." From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/CubeLinePicking.html.

[8] J. Philip, "The distance between two random points in a 4- and 5-cube." http://www.math.kth.se/~johanph/h45.pdf.

[9] S.-J. Tu and E. Fischbach, "A new geometric probability technique for an N-dimensional sphere and its applications." arXiv:math-ph/0004021v3, http://arxiv.org/abs/math-ph/0004021, 2000.

[10] D. Robbins, "Average distance between two points in a box," *American Mathematical Monthly*, vol. 85, p. 278, 1978.

[11] E. W. Weisstein, "Ball line picking." From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/BallLinePicking.html.

[12] E. W. Weisstein, "Circle line picking." From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/CircleLinePicking.html.

[13] E. W. Weisstein, "Sphere line picking." From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/SphereLinePicking.html.

[14] M. Bartlett, "The spectral analysis of two-dimensional point processes," *Biometrika*, vol. 51, no. 3/4, pp. 299–311, 1964.