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 22, 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 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

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

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

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

4 Known Results

The Waxman graph has been analysed in some detail, starting first with analysis of points chosen randomly in the plane. Such points form a spatial Poisson process across the region of interest. Much analysis has considered such processes, and we will only repeat directly relevant results here. The most recondite result concerns the distribution of distances between two points chosen randomly in the unit square. The probability density function of distances between two (uniformly) randomly chosen points on the unit square is given in [1], as

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

This is a rather complicated expression, but is easily evaluated numerically. Alternatively, in [2] this expression was approximated with a β function, though given the requirements to numerically evaluate that function there hardly seems any advantage. Figure **??** shows the shape of this distribution. It is noteworthy that the distance distribution between two points will also give the distribution of the distances of links in an Erdős-Rényi random graph constructed on a set of random points on the square.

The general problem of choosing a line from some space has been called a *line picking* problem, with many examples having been studied.

Numerically it is straight-forward to calculate the function $g(t)$ and its Laplace transform where the nodes are placed on irregular regions of the plane. It simply requires computation of a double integral. Note however, that even if we constrain nodes to lie in the region, if the region is non-convex the links may cross the boundaries of the regions. A simple example is the Waxman graph on a circular disc of radius R. It has the advantage of isometry, or rotational symmetry. When we restrict the vertices (and edges) to a circle the problem is still relatively easy to solve. The distance distribution between random points in the a circle radius R is given by [3] as

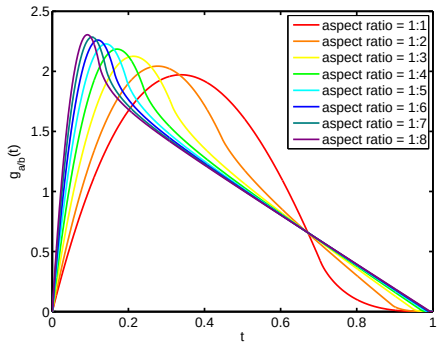
$$g_R(t) = \frac{4t}{\pi R^2}\cos^{-1}\left(\frac{t}{2R}\right) - \frac{2t^2}{\pi R^3}\sqrt{1 - \frac{t^2}{4R^2}}. \tag{2}$$

This function is also shown (for $R = 0.5$) in Figure **??** for comparison with the function on the unit square. Interesting, despite noticable differences in these functions, we found little difference between the critical ration $-G'(s)/G(s)$ for large s (small α) suggesting the in these cases the exact form of the region may be fairly unimportant.

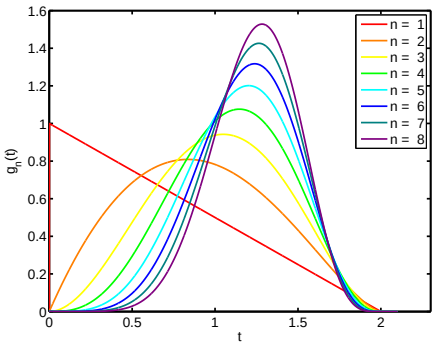
Note that the asymptotic expansion of $\cos^{-1}(x) = \frac{\pi}{2} - O(x)$, so the small t expansion of $g_R(t) = 2t/R^2$, which is identical per unit area to the small α approximation for the square.

Another example, an approximation of which appears in [2] is the rectangle. The precise form of the distance density for this case is given in [**?**, Theorem 2.4.4] as

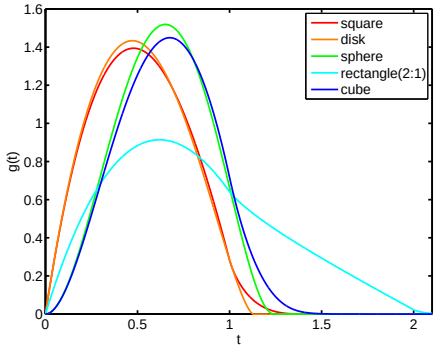
$$g_{a,b}(t) = \frac{4t}{a^2 b^2}\phi_{a,b}(t), \tag{3}$$



(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.

where

$$\phi_{a,b}(t)=\begin{cases}\frac{ab\pi}{2}-(a+b)t+\frac{t^2}{2}, & \text{for } t\le a,\\ ab\sin^{-1}(a/t)-\frac{a^2}{2}-bt+b\sqrt{t^2-a^2}, & \text{for } a\le t\le b,\\ ab\left[\sin^{-1}(a/t)-\sin^{-1}\sqrt{1-\frac{b^2}{t^2}}\right]-\frac{a^2+b^2+t^2}{2}+a\sqrt{t^2-b^2}+b\sqrt{t^2-a^2}, & \text{for } b\le t\le\sqrt{a^2+b^2},\\ 0, & \text{otherwise,}\end{cases} \tag{4}$$

where the rectangle has sides of length $a\le b$. Figure 1a

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 α 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 α 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.

The above result generalize immediately to Waxman graphs on n dimensional spaces. We simply need to calculate $g(t)$ on the space in question. For instance, generalizations of $g(t)$ on the n-dimensional ball exist [3], called the "ball line picking", i.e., for a n-dimensional ball of radius R,

$$g_n(t)=n\frac{t^{n-1}}{R^n}I_x\left(\frac{1}{2}(n+1),\frac{1}{2}\right),$$

where

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

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

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

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

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

The first few of these are [3] (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)):

$$g_1(t) = \frac{1}{R} - \frac{t}{2R} \tag{5}$$

$$g_2(t) = \frac{4t}{\pi R^2}\cos^{-1}\left(\frac{t}{2R}\right) - \frac{2t^2}{\pi R^3}\sqrt{1-\frac{t^2}{4R^2}} \tag{6}$$

$$= \frac{2t}{R^2} - \frac{2t^2}{\pi R^3}\sqrt{1-\frac{t^2}{4R^2}} - \frac{4t}{\pi R^2}\sin^{-1}\left(\frac{t}{2R}\right) \tag{7}$$

$$g_3(t) = \frac{3t^2}{R^3} - \frac{9t^3}{4R^4} + \frac{3t^5}{16R^6} \tag{8}$$

$$g_4(t) = \frac{8t^3}{\pi R^4}\cos^{-1}\left(\frac{t}{2R}\right) - \frac{8t^4}{3\pi R^5}\left(1-\frac{t^2}{4R^2}\right)^{3/2} - \frac{4t^4}{\pi R^5}\sqrt{1-\frac{t^2}{4R^2}} \tag{9}$$

$$g_5(t) = \frac{5t^4}{R^5} - \frac{75t^6}{16R^6} + \frac{25t^7}{32R^8} - \frac{15t^9}{256R^{10}} \tag{10}$$

The above cases for which we know the form of the distribution are straight-forward to compute, but note that the derivation of $g(t)$ just involves a set of integrals that could be numerically pre-computed for any region of interest.

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.

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 [3], 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.,

$$g_n(t) \simeq n\frac{t^{n-1}}{R^n} \tag{11}$$

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 α) approximations of the estimator.

For instance, from before, we know that if the jth 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 ϵ is negligable, where ϵ 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 [4], 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 [5]

$$g(t) = \frac{1}{\pi}\left(1 - \frac{s^2}{4}\right)^{-1/2} \simeq \frac{1}{\pi}\left(1 - \frac{s^2}{4} + \cdots\right) \tag{12}$$

whereas in *sphere line picking* [6], there distribution takes the form

$$g(t) = \frac{1}{2}t. \tag{13}$$

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, and resulting Laplace transforms on irregular regions, for which there is no closed form solution. This can be easily accomplished numerically, by simulating a cases.

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.

3. We can estimate the Laplace transform as follows:

$$\begin{aligned} G(s) &= \int_0^t g(t)\,dt &&= \frac{1}{n}\sum_{i=1}^{N} e^{-st_i}, \\ G'(s) &= \int_0^t tg(t)\,dt &&= \frac{1}{n}\sum_{i=1}^{N} t_i e^{-st_i}, \end{aligned} \tag{14}$$

Note that only one set of points need be generated to estimate the Laplace transforms for a large set of values of s.

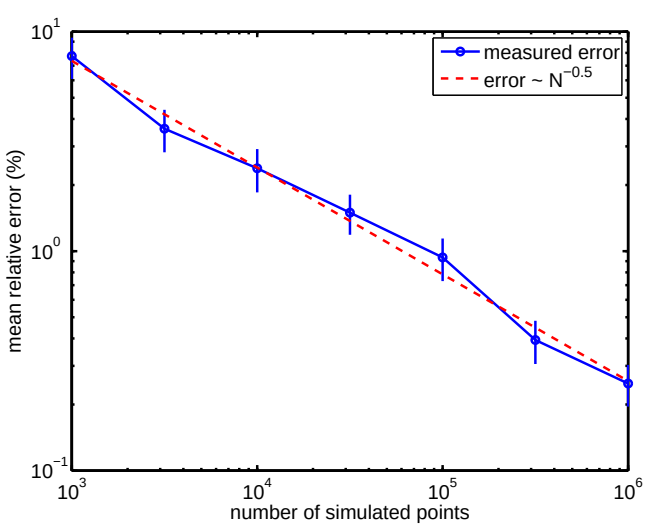
We have tested the above approach, running it 30 times (with different seeds), in Matlab for various values of N. The results are shown in Figure 2. The first plot shows estimates of the mean relative absolute error of the estimated Laplace transforms over the range $S \in [0, 50]$. We can see from the 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.

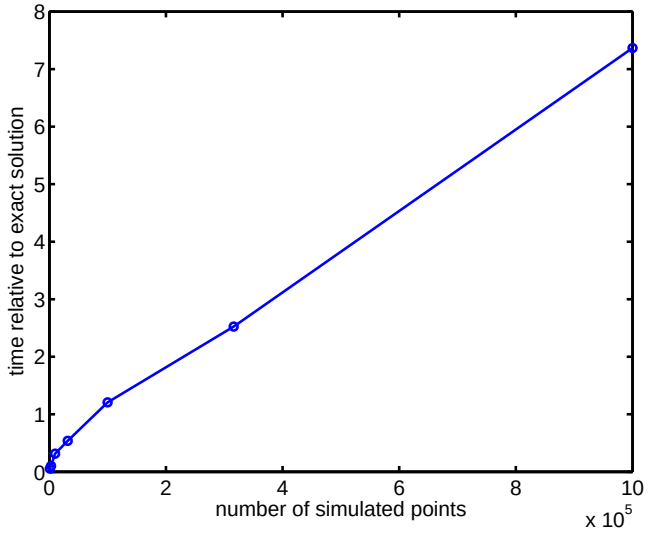
The simulation-based approach is not as accurate as the exact numerical approach, however, it accuracy should be sufficient for most estimation problems, without increasing the computational workload unduly.

[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.



(a) Mean relative error in $-G'(\hat{s})/G(\hat{s})$.



(b) Relative computation time of the simulation-based approach, with respect to the exact approach.

Figure 2: A comparison of simulation-based Laplace transform calculation with the exact approach.

We could further use the above in the estimator as follows: we aim to find the value $\hat{s}$ such that $-G'(\hat{s})/G(\hat{s}) = \bar{d}$, so

$$\begin{aligned} -G'(\hat{s}) &= \bar{d}G(\hat{s}) \\ \frac{1}{n}\sum_{i=1}^{N} t_i e^{-st_i} &= \bar{d}\frac{1}{n}\sum_{i=1}^{N} e^{-st_i} \\ \sum_{i=1}^{N}(t_i - \bar{d})e^{-st_i} &= 0 \end{aligned}$$

We could feed this new function directly into our search algorithm to find the estimate, avoiding the need to calculate both $-G'(\hat{s})$ and $G(\hat{s})$ for multiple values of s. This approach has other advantages as well:

- The gradient of this function is immediately obvious, and not problem specific as it is for closed form solutions, leading to generic estimation functions;
- We could approach the estimation iterative, for intance, by generating M point, performing an estimate, and then refining it with the next M points, continuing until the results converge on a solution with some desired tolerance.

6 Programs

6.1 A Rough Guide

6.2 Numerical Issues

7 Correlations

Correlations between distances [7]

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

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

[3] 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.

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

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

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

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