# Intrinsic Dimension of Point Sets

Mathieu Chabaud, Javier Garcia, Silvia Ghinassi, Garrett Mulcahy, Rico Qi, Vlad Radostev, and Linda Yuan

Washington eXperimental Mathematics Lab, University of Washington – Seattle

DEPARTMENT OF MATHEMATICS
UNIVERSITY of WASHINGTON

## Goal of the project

Our main goal this quarter was to continue our work towards a notion of dimension for finite point sets, analogous to familiar notions of dimension in Geometric Measure Theory for "continuous" sets. We aim to (1) Refine the computational methods for notion of dimension proposed last quarter that we called the Frostman dimension, (2) Explore a well-known approach to computing fractal dimension called the correlation integral and propose a local variant, (3) Develop techniques to "stratify" a dataset into its higher and lower dimensional components, (4) Apply these methods to real (i.e. non-artificially generated) datasets

## Background

Classic measure theory tells us the measure of any finite set is $0$. However, we can adapt certain ideas from classical geometric measure theory, such as Hausdorff measure and dimension [1], for the purposes of describing the dimension of finite data sets.

Hausdorff dimension can be thought of as the measure of an objects roughness. "Smooth" objects, which typically posses a small, finite number of corners, such as a point, line, square, and cube, have integer Hausdorff dimension, $0$, $1$, $2$, and $3$ respectively. "Rough" objects, i.e. fractals, typically have non-integer Hausdorff dimension. In the case of self-similar objects their Hausdorff dimension can be easily computed by the given logarithmic ratio,

$$\dim_{\text{Hausdorff}}(X) = \log(N)/\log(1/r),$$

where $N$ is the number of copies at each iteration, and $r \in (0,1)$ is the common scaling factor.

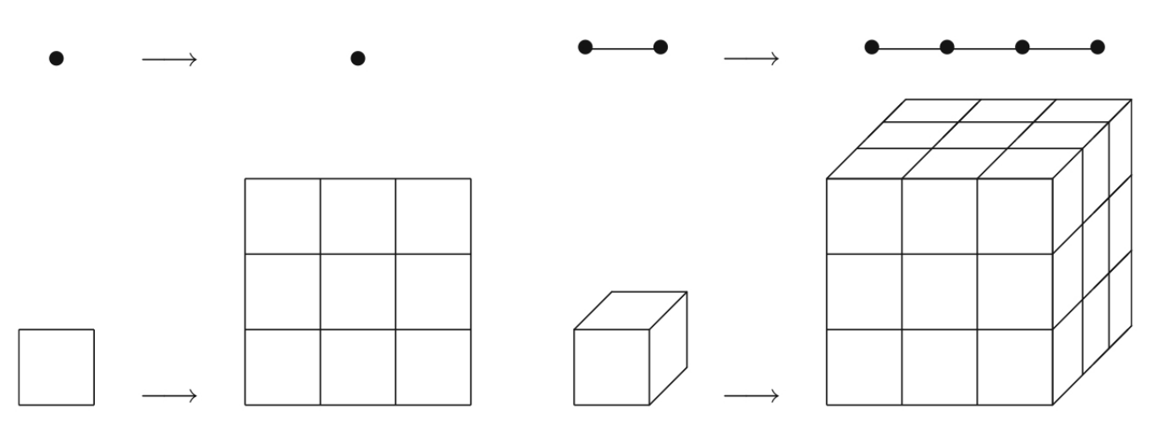

Figure 1. Consider a point, line, square, and cube, scaled by a factor of three. Notice that the scaled version gives $3^d$ copies of itself for each object. Thus we get the power rule for the Hausdorff dimension of self similar objects (without log simplification), $S^d = n$

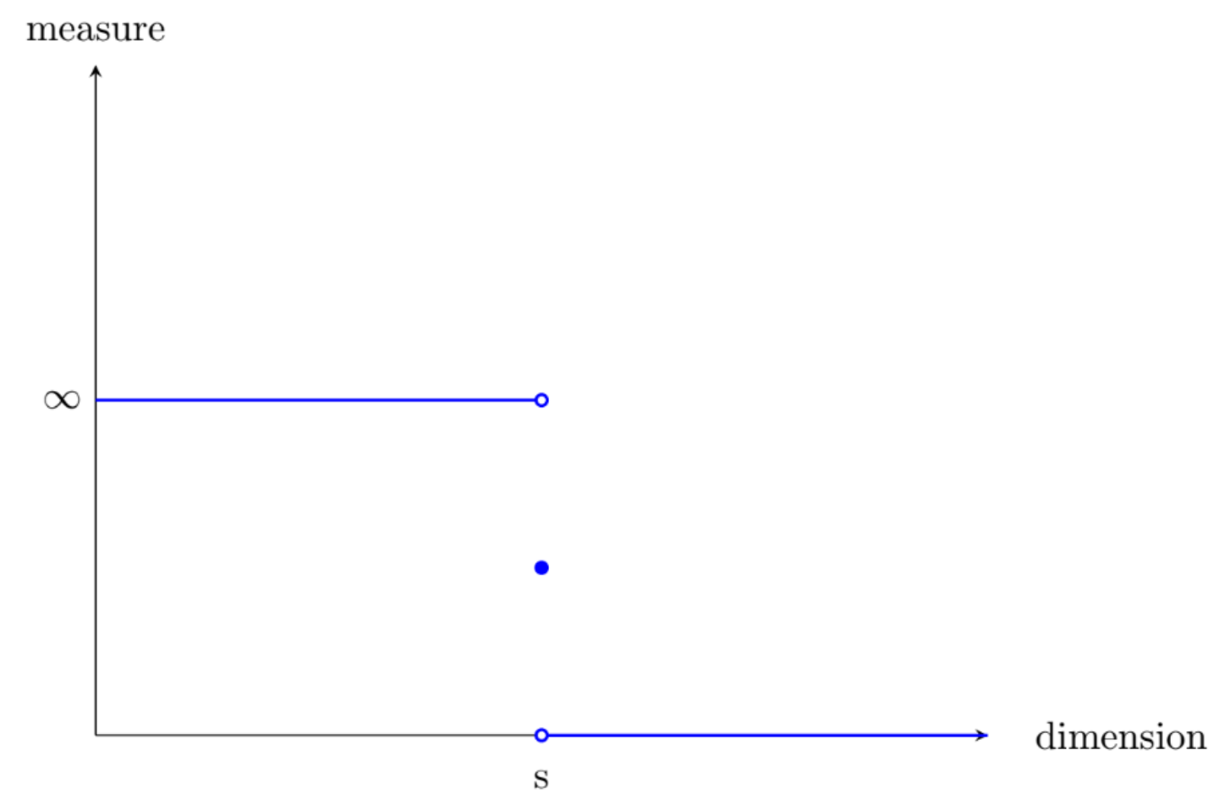

Figure 2. Consider the length and the volume of a square. The change in the plot as the parameter s varies near 2 encapsulates the idea that the square is 2-dimensional.

## Frostman Density

We define a local "density" inspired by Frostman's characterization [2, Theorem 8.8]:

$$F_s(p,n) = |P_n \cap B(p, r(P_n))|/(nr(P_n)^s).$$

and with it, a local dimension:

$$\dim_p(P_n) = \sup\left\{ s \in [0,\infty) \mid \sup_n F_s^1(p,n) < \infty \right\}.$$

where $P_n$ is a uniformly randomly sampled point set of size $n$ chosen from the continuous set $E$, and $r$ is a function that maps finite point sets to real numbers.

## Radius Experiments (Fine-tuning Frostman)

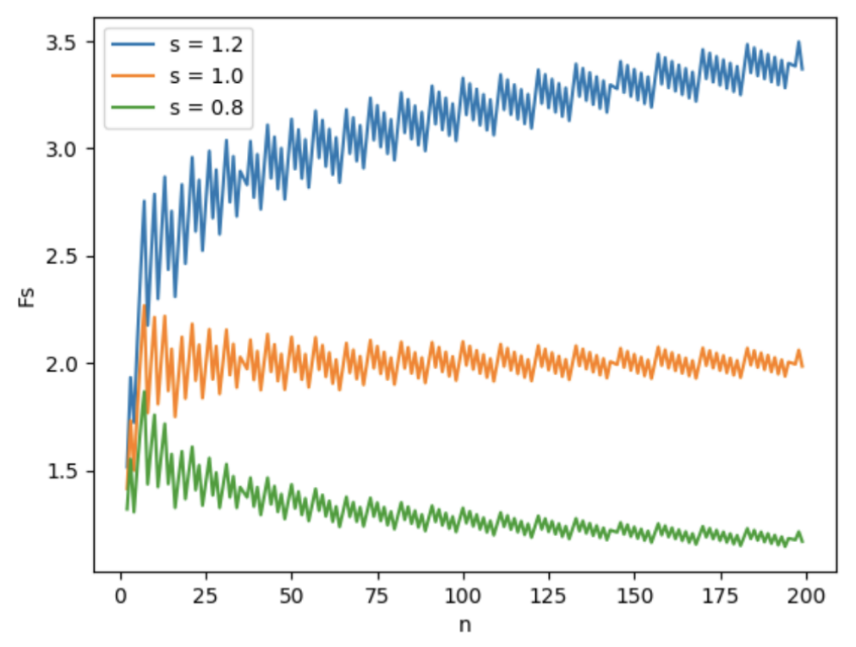

Figure 3. $[0,1]$ with $p = 0.5$, deterministically generated $P_n$, and $r(P_n) = |P_n|^{-0.5}$.

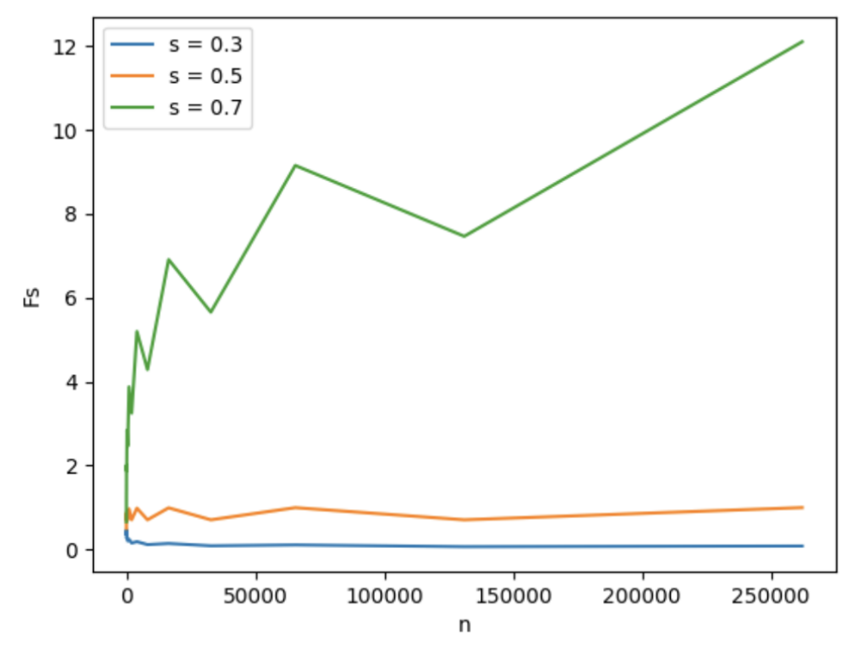

Figure 4. $\frac{1}{2}$-d cantor set with $p = 0$, deterministically generated $P_n$, and $r(P_n) = |P_n|^{-1}$.

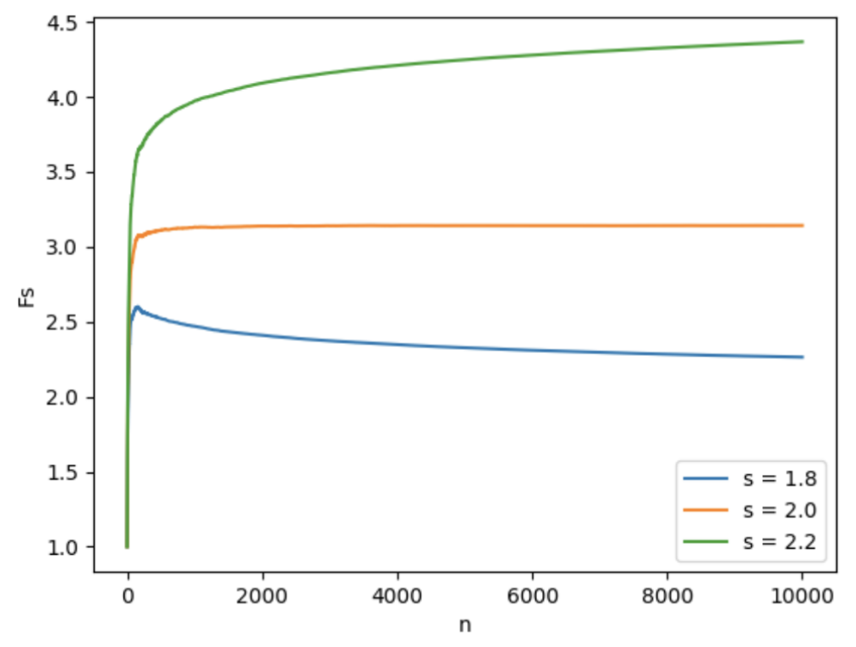

Figure 5. $[0,1] \times [0,1]$ with $p = (0.5, 0.5)$, randomly generated $P_n$, $r(P_n) = |P_n|^{-0.2}$, and smoothing.

- We can explain the plot on the left with $s = 1$.

$$\mathbb{E}(F_s(p,n)) = \mathbb{E}(\frac{|P_n \cap B(p,r)|}{n \cdot r}) = \frac{\mathbb{E}(|P_n \cap B(p,r)|)}{n \cdot r} = \frac{n \cdot 2r}{n \cdot r} = 2.$$

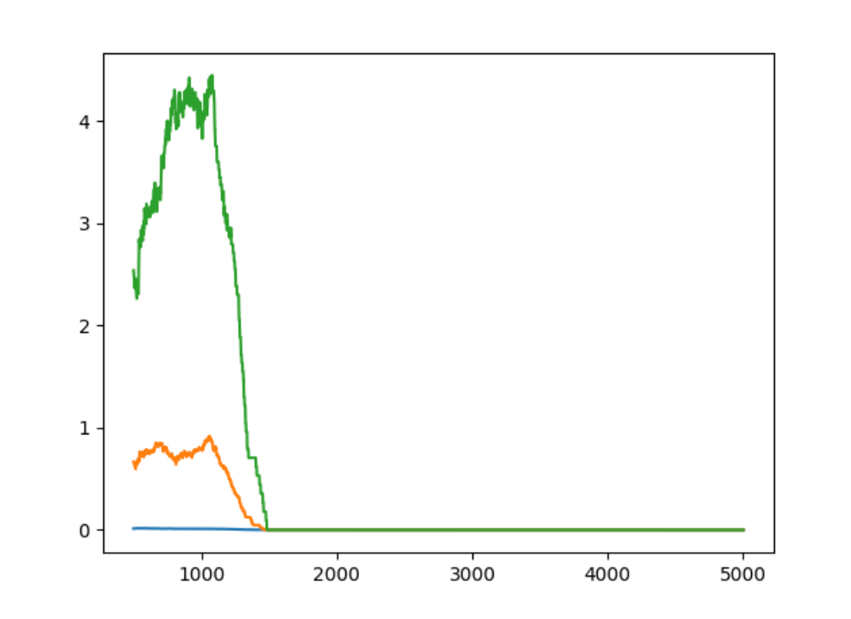

Figure 6. If $r$ is too small, or converges too quickly to $0$, the plot collapses to $0$, due to $|P_n \cap B(p, r(P_n))| = 0$.

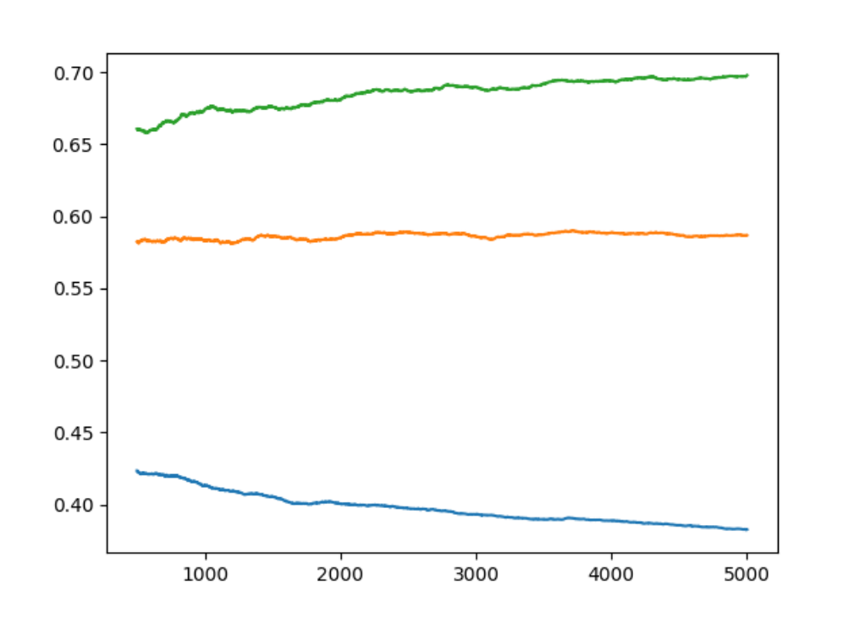

Figure 7. If $r$ is too large, $B(p,r)$ might not encapsulate $E$'s local features around $p$.

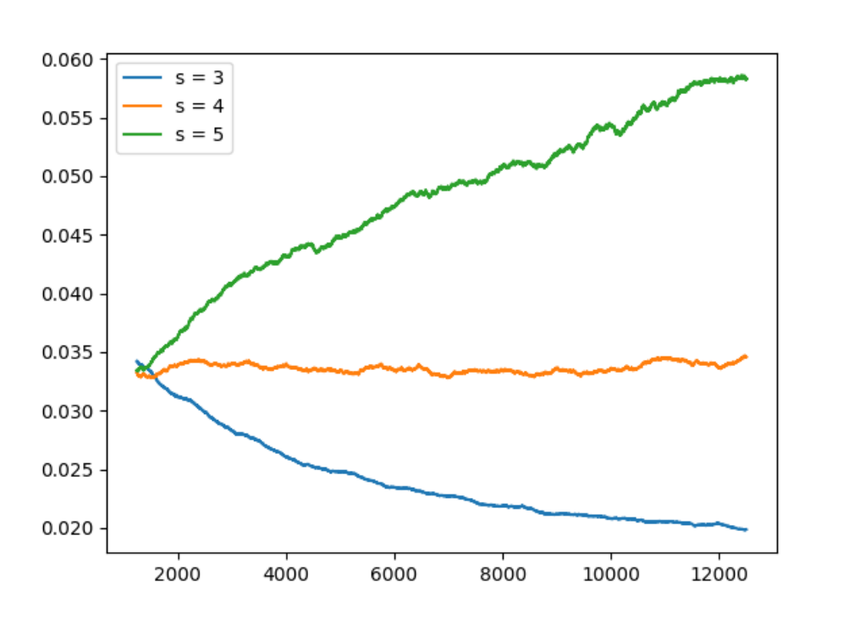

Figure 8. One approach that we have found to be more generalizable is $r = \inf\{r \mid |P_n \cap B(p,r)| = c\}$.

## Correlation Integral

Given a dataset $P = \{x_i : i \in \{1,\dots,n\}\}$, we define for $r > 0$ and $p \in P$

$$C(r) = |\{(p,q) \in P \times P : \|p-q\| \le r\}|$$
$$C(p,r) = |\{q \in P : \|q-p\| \le r\}|$$

We assume that $C(r) \propto r^d$; we call $d$ the **correlation integral dimension**. Taking log on both sides, we see that

$$\log C(r) = d \log r + \kappa,$$

where $\kappa \in \mathbb{R}$ is some constant. Thus, we estimate the slope of the line determined by $(\log r, \log C(r))$. We define the **local correlation integral dimension** analogously, using $C(p,r)$ instead of $C(r)$. That is, we estimate the slope of the line determined by $(\log r, \log C(p,r))$.

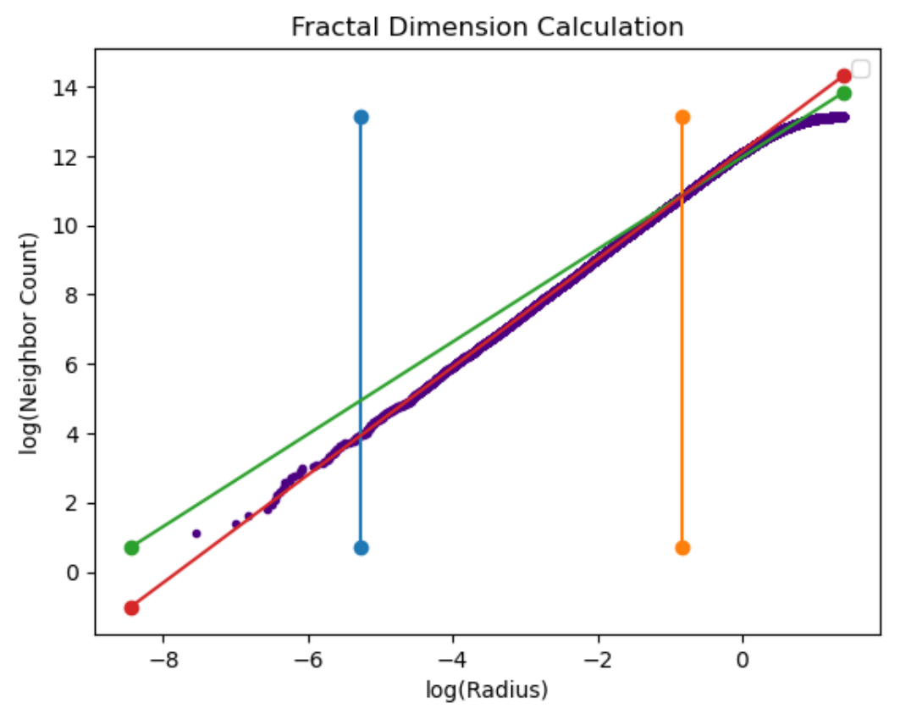

Figure 9. Overall Fractal Dimension of Sierpinski Triangle with manual bounds, slope: 1.5785

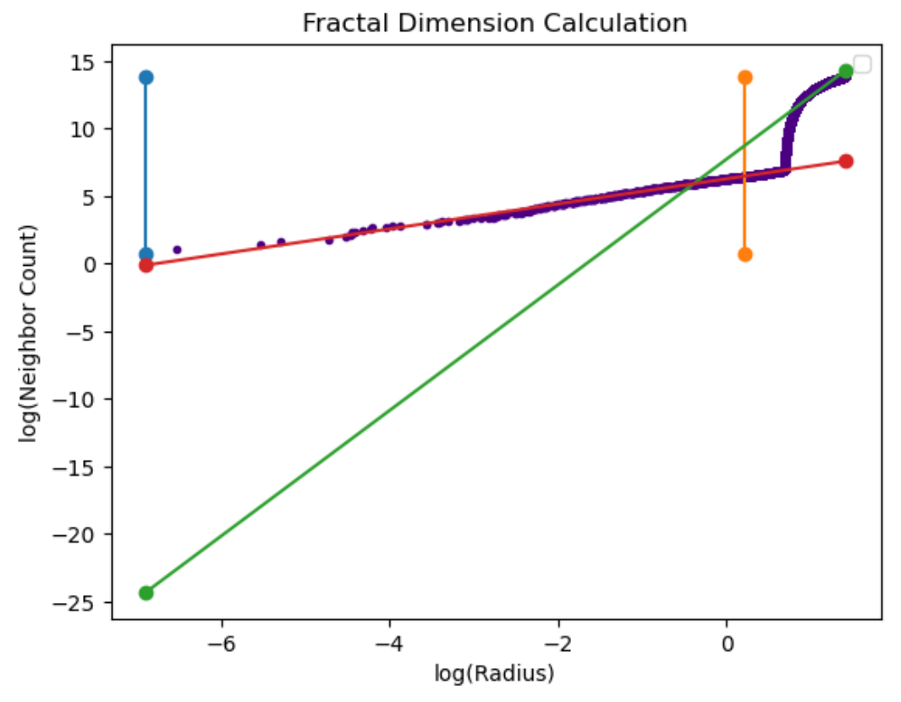

Figure 10. Local Fractal Dimension of far line point for line-box plot, large difference in slope before and after cutoff: 4.5244 before, 0.9763 after

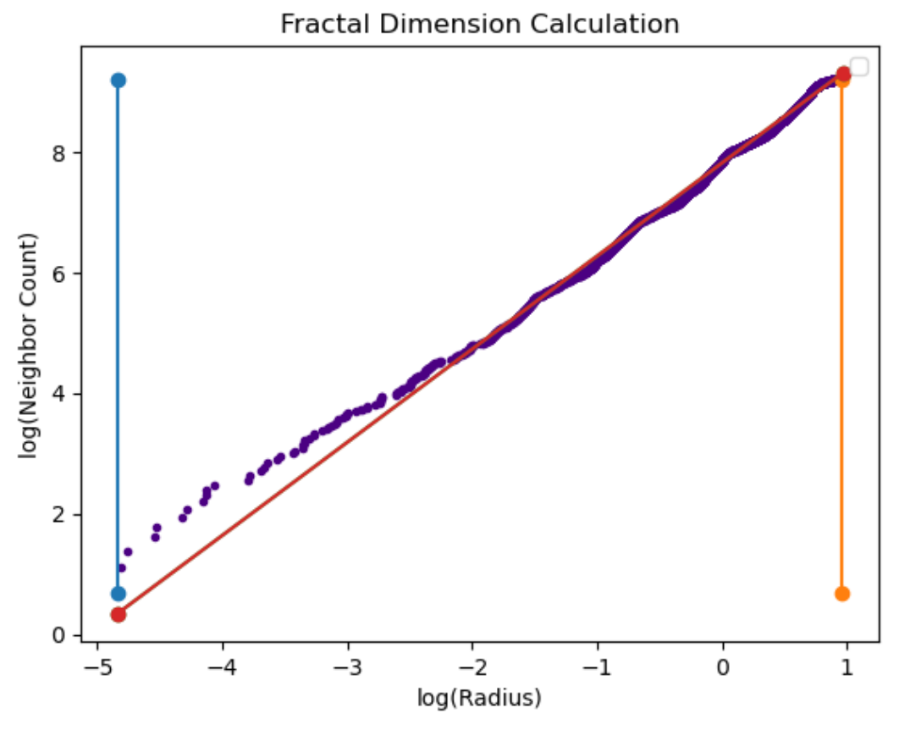

Figure 11. Local Dimension of random point in Sierpinski Triangle, minimal difference in slope before and after cutoff: 1.3281889 before, 1.3284704 after

## Data Set Stratification

Separating chunks of data by its local dimension could be extremely helpful for finding ways of optimizing lossy compression, or stratifying data. Below we have some of our early attempts of calculating, and plotting, the local fractal dimension of points in a dataset.

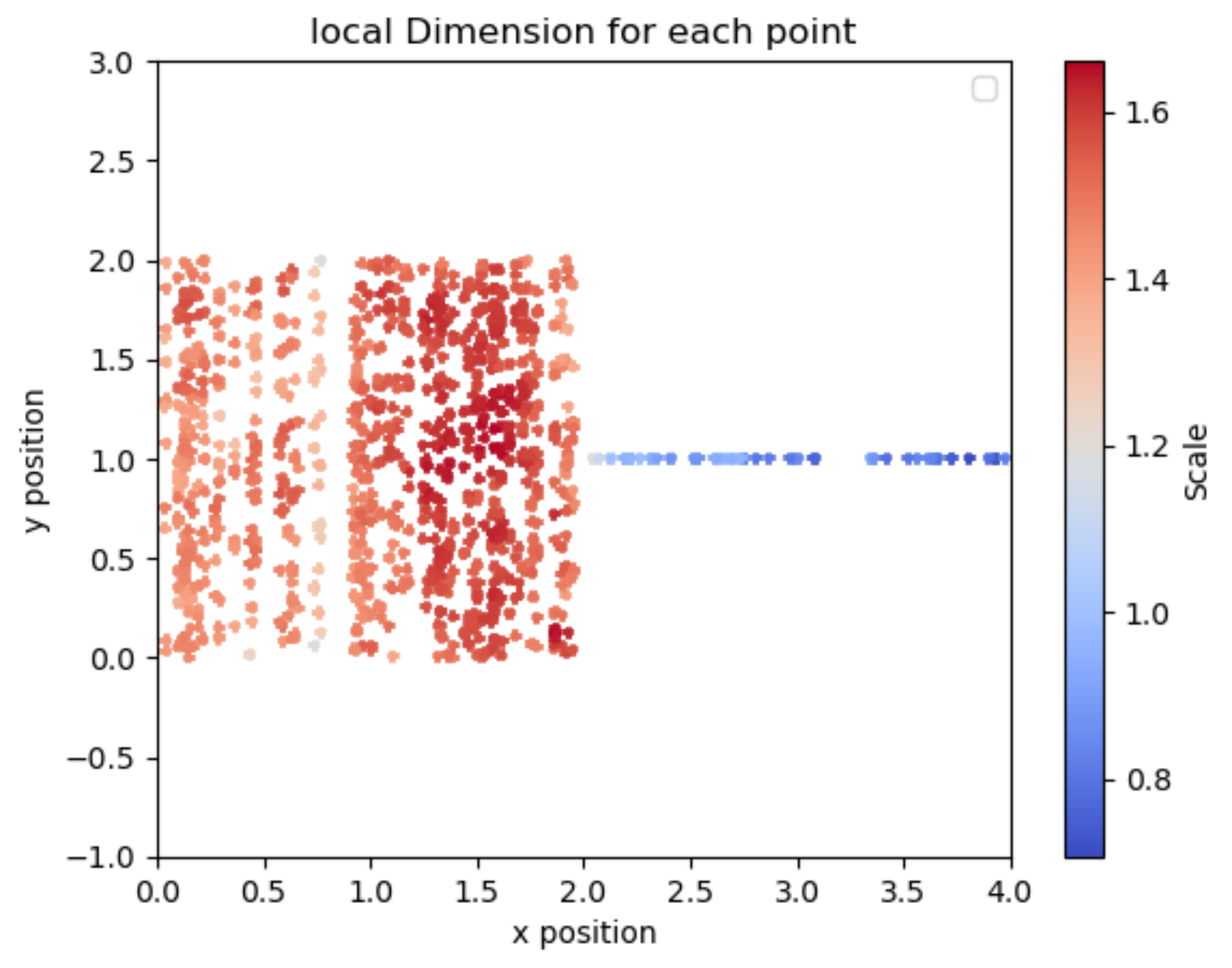

Figure 12. By finding the correlation integral of subsets of the data within some distance of a single point, we can derive a notion of local dimension. By coloring each point according to its local dimension we can see how it could be used to classify sections of data.

We plotted significant points of the square-line graph using Frostman density and found their local dimensions corresponding to the result above with correlation integral.

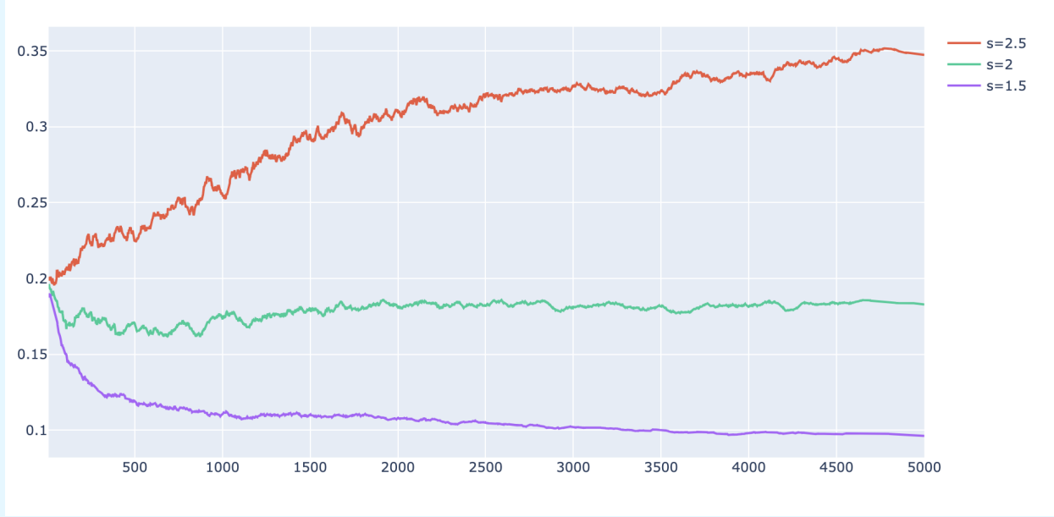

Figure 13. When taking the point (1, 1) at the center of the square, s converges at 2.

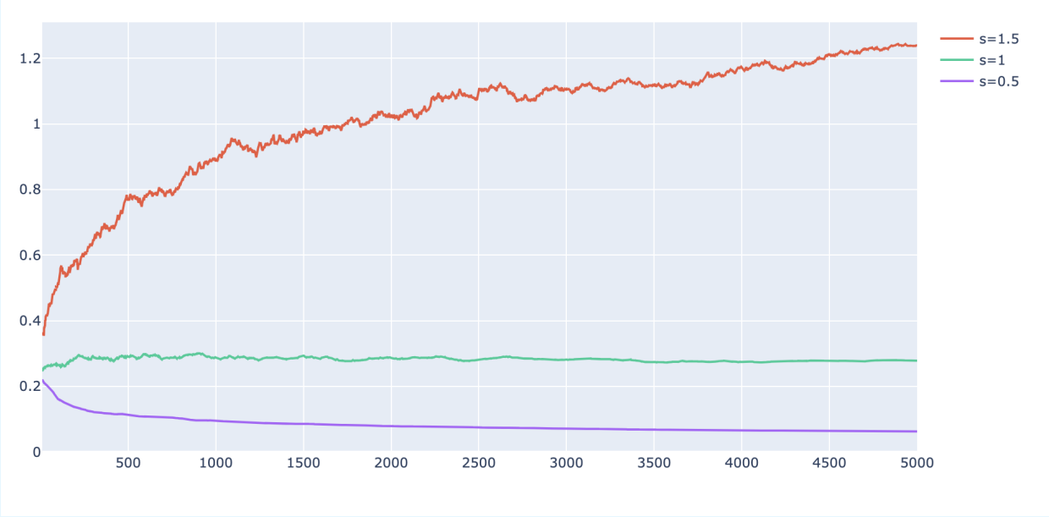

Figure 14. When taking the point (4, 1) on the right end of the line, s converges at 1.

## Future directions

- Optimize our code to handle larger data sets, and high-dimensional data;
- Make sense of the effective dimension in the context of data science; potential applications include public health, where local dimension can be used to identify when the spread of a disease shifts.

## Acknowledgements

We would like to thank Juliette Bergantino for working with us Spring quarter 2023 on learning the mathematical background that enabled us to continue the project. We are also infinitely grateful to Naomi Alterman from the UW eScience institute for her help in navigating data sets!

## References

[1] Kenneth Falconer. *Fractal geometry*. John Wiley & Sons, Ltd., Chichester, 1990. Mathematical foundations and applications.

[2] Pertti Mattila. *Geometry of sets and measures in Euclidean spaces*, volume 44 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1995. Fractals and rectifiability.

[3] Krishna Kumaraswamy skkumar. Fractal dimension for data mining. 2003.