

Off the Beaten Path Tutorial: Stochastic Processes and Simulations – Volume 1

Vincent Granville, vincentg@MLtechniques.com
MachineLearningRecipes.com | MLtechniques.com
Version 5.0, March 2022

Note: External links (in blue) and internal references (in red) are clickable throughout this document. Key-words highlighted in orange are indexed; those in red are both indexed and in the glossary section.

Contents

About this Textbook	2
Target Audience	3
About the Author	4
1 Poisson-binomial or Perturbed Lattice Process	5
1.1 Definitions	6
1.2 Point Count and Interarrival Times	7
1.3 Limiting Distributions, Speed of Convergence	8
1.4 Properties of Stochastic Point Processes	8
1.4.1 Stationarity	8
1.4.2 Ergodicity	9
1.4.3 Independent Increments	9
1.4.4 Homogeneity	9
1.5 Transforming and Combining Multiple Point Processes	10
1.5.1 Marked Point Process	10
1.5.2 Rotation, Stretching, Translation and Standardization	10
1.5.3 Superimposition and Mixing	11
1.5.4 Hexagonal Lattice, Nearest Neighbors	12
2 Applications	12
2.1 Modeling Cluster Systems in Two Dimensions	13
2.1.1 Generalized Logistic Distribution	14
2.1.2 Illustrations	15
2.2 Infinite Random Permutations with Local Perturbations	16
2.3 Probabilistic Number Theory and Experimental Maths	17
2.3.1 Poisson Limit of the Poisson-binomial Distribution, with Applications	18
2.3.2 Perturbed Version of the Riemann Hypothesis	20
2.4 Videos: Fractal Supervised Classification and Riemann Hypothesis	22
2.4.1 Dirichlet Eta Function	22
2.4.2 Fractal Supervised Classification	23
3 Statistical Inference, Machine Learning, and Simulations	24
3.1 Model-free Tests and Confidence Regions	24
3.1.1 Methodology and Example	25
3.1.2 Periodicity and Amplitude of Point Counts	28
3.1.3 A New Test of Independence	30
3.2 Estimation of Core Parameters	32
3.2.1 Intensity and Scaling Factor	32
3.2.2 Model Selection to Identify F	33
3.2.3 Theoretical Values Obtained by Simulations	33
3.3 Hard-to-Detect Patterns and Model Identifiability	34
3.4 Spatial Statistics, Nearest Neighbors, Clustering	36
3.4.1 Stochastic Residues	36
3.4.2 Inference for Two-dimensional Processes	36
3.4.3 Clustering Using GPU-based Image Filtering	40
3.4.4 Black-box Elbow Rule to Detect Outliers and Number of Clusters	41

3.5	Boundary Effect	44
3.5.1	Quantifying some Biases	44
3.5.2	Extreme Values	46
3.6	Poor Random Numbers and Other Glitches	48
3.6.1	A New Type of Pseudo-random Number Generator	48
4	Theorems	49
4.1	Notations	49
4.2	Link between Interarrival Times and Point Count	49
4.3	Point Count Arithmetic	50
4.4	Link between Intensity and Scaling Factor	50
4.5	Expectation and Limit Distribution of Interarrival Times	51
4.6	Convergence to the Poisson Process	51
4.7	The Inverse or Hidden Model	52
4.8	Special Cases with Exact Formula	53
4.9	Fundamental Theorem of Statistics	54
5	Exercises, with Solutions	55
5.1	Full List	55
5.2	Probability Distributions, Limits and Convergence	55
5.3	Features of Poisson-binomial Processes	59
5.4	Lattice Networks, Covering Problems, and Nearest Neighbors	61
5.5	Miscellaneous	65
6	Source Code, Data, Videos, and Excel Spreadsheets	69
6.1	Interactive Spreadsheets and Videos	70
6.2	Source Code: Point Count, Interarrival Times	71
6.2.1	Compute $E[N(B)]$, $\text{Var}[N(B)]$ and $P[N(B) = 0]$	72
6.2.2	Compute $E[T]$, $\text{Var}[T]$ and $E[T^r]$	73
6.2.3	Produce random deviates for various F 's	74
6.2.4	Compute $F(x)$ for Various F	74
6.3	Source Code: Radial Cluster Simulation	74
6.4	Source Code: Nearest Neighbor Distances	75
6.5	Source Code: Detection of Connected Components	79
6.6	Source Code: Visualizations, Density Maps	80
6.6.1	Visualizing the Nearest Neighbor Graph	80
6.6.2	Clustering and Density Estimation via Image Filtering	81
6.7	Source Code: Production of the Videos	85
6.7.1	Dirichlet Eta Function	85
6.7.2	Fractal Supervised Clustering	86
	Glossary	89
	List of Figures	90
	References	91
	Index	94

About this Textbook

This scratch course on stochastic processes covers significantly more material than usually found in traditional books or classes. The approach is original: I introduce a new yet intuitive type of random structure called perturbed lattice or Poisson-binomial process, as the gateway to all the stochastic processes. Such models have started to gain considerable momentum recently, especially in sensor data, cellular networks, chemistry, physics and engineering applications. I present state-of-the-art material in simple words, in a compact style, including new research developments and open problems. I focus on the methodology and principles, providing the reader with solid foundations and numerous resources: theory, applications, illustrations, statistical inference, references, glossary, educational spreadsheet, source code, stochastic simulations, original exercises, videos and more.

Below is a short selection highlighting some of the topics featured in the textbook. Some are research re-

sults published here for the first time.

GPU clustering	Fractal supervised clustering in GPU (graphics processing unit) using image filtering techniques akin to neural networks, automated black-box detection of the number of clusters, unsupervised clustering in GPU using density (gray levels) equalizer
Inference	New test of independence, spatial processes, model fitting, dual confidence regions, minimum contrast estimation, oscillating estimators, mixture and surperimposed models, radial cluster processes, exponential-binomial distribution with infinitely many parameters, generalized logistic distribution
Nearest neighbors	Statistical distribution of distances and Rayleigh test, Weibull distribution, properties of nearest neighbor graphs, size distribution of connected components, geometric features, hexagonal lattices, coverage problems, simulations, model-free inference
Cool stuff	Random functions, random graphs, random permutations, chaotic convergence, perturbed Riemann Hypothesis (experimental number theory), attractor distributions in extreme value theory, central limit theorem for stochastic processes, numerical stability, optimum color palettes, cluster processes on the sphere
Resources	28 exercises with solution expanding the theory and methods presented in the textbook, well documented source code and formulas to generate various deviates and simulations, simple recipes (with source code) to design your own data animations as MP4 videos – see ours on YouTube

This first volume deals with point processes in one and two dimensions, including spatial processes and clustering. The next volume in this series will cover other types of stochastic processes, such as Brownian-related and random, chaotic dynamical systems. The point process which is at the core of this textbook is called the Poisson-binomial process (not to be confused with a binomial nor a Poisson process) for reasons that will soon become apparent to the reader. Two extreme cases are the standard Poisson process, and fixed (non-random) points on a lattice. Everything in between is the most exciting part.

Target Audience

College-educated professionals with an analytical background (physics, economics, finance, machine learning, statistics, computer science, quant, mathematics, operations research, engineering, business intelligence), students enrolled in a quantitative curriculum, decision makers or managers working with data scientists, graduate students, researchers and college professors, will benefit the most from this textbook. The textbook is also intended to professionals interested in automated machine learning and artificial intelligence.

It includes many original exercises requiring out-of-the-box thinking, and offered with solution. Both students and college professors will find them very valuable. Most of these exercises are an extension of the core material. Also, a large number of internal and external references are immediately accessible with one click, throughout the textbook: they are highlighted respectively in red and blue in the text. The material is organized to facilitate the reading in random order as much as possible and to make navigation easy. It is written for busy readers.

The textbook includes full source code, in particular for simulations, image processing, and video generation. You don't need to be a programmer to understand the code. It is well documented and easy to read, even for people with little or no programming experience. Emphasis is on good coding practices. The goal is to help you quickly develop and implement your own machine learning applications from scratch, or use the ones offered in the textbook. The material also features professional-looking spreadsheets allowing you to perform interactive statistical tests and simulations in Excel alone, without statistical tables or any coding. The code, data sets, videos and spreadsheets are available on my GitHub repository.

The content in this textbook is frequently of graduate or post-graduate level and thus of interest to researchers. Yet the unusual style of the presentation makes it accessible to a large audience, including students and professionals with a modest analytic background (a standard course in statistics). It is my hope that it will entice beginners and practitioners faced with data challenges, to explore and discover the beautiful and useful aspects of the theory, traditionally inaccessible to them due to jargon.

About the Author

Vincent Granville, PhD is a pioneering data scientist and machine learning expert, co-founder of Data Science Central (acquired by a publicly traded company in 2020), former VC-funded executive, author and patent owner. Vincent's past corporate experience includes Visa, Wells Fargo, eBay, NBC, Microsoft, CNET, InfoSpace and other Internet startup companies (one acquired by Google). Vincent is also a former post-doctoral fellow from Cambridge University, and the National Institute of Statistical Sciences (NISS). He is currently publisher at DataShaping.com. He makes a living as an independent researcher working on stochastic processes, dynamical systems, experimental math and probabilistic number theory.

Vincent published in Journal of Number Theory, Journal of the Royal Statistical Society (Series B), and IEEE Transactions on Pattern Analysis and Machine Intelligence, among others. He is also the author of multiple books, including "Statistics: New Foundations, Toolbox, and Machine Learning Recipes", "Applied Stochastic Processes, Chaos Modeling, and Probabilistic Properties of Numeration Systems" with a combined reach of over 250,000, as well as "Becoming a Data Scientist" published by Wiley. For details, see my Google Scholar profile, [here](#).