A non-technical research statement

Xiaochuan Yang

November 10, 2020

My research interest lies in Probability Theory and its connection with geometric and topological data analysis, statistical physics, high dimensional statistics, multifractality, and number theory.

1 From ecology to geometric and topological data analysis

Suppose n bushes are randomly scattered in the unit square, and a disease (or fire) then appears at one of them. Once sick, a bush never recovers, and passes on the disease to every other bush within a distance r. Eventually, all the bushes become sick, except for those which are insulated by a zone of radius r containing no bushes that ever become sick. After a long period of time (relative to the time scale of the spread of the disease), all the sick bushes die, leaving behind any insulated bushes. If a sufficient number of such bushes remain, there will be a chance for the forest to regrow. We are here interested in the question: for which values of r is there likely to be one or more such insulated bushes?

The answer to the previous question is related to the maximal edge length of the minimal tree spanning the n bushes. The minimal spanning tree is a fundamental example of combinatorial optimisation which has applications in statistics, physics and computer science.

This ecology example has a clear interpretation in geometric and topological data analysis. The n bushes are data points in a possibly high dimensional space sampled from an unknown probability distribution. The question of utmost importance in geometric and topological data analysis is whether one can infer topological and geometric information of the underlying probability distribution from the observed data. A popular way of doing so is to form combinatorial objects such as graphs from the data points by putting an edge between points at distance r. One then examines the effect of varying r to, say, the connectivity of the graph. Hopefully the information extrapolated from the combinatorial objects approximates some truth of the continuum unknown distribution.

One of the technical tools for analysis is the continuum percolation. Percolation is a popular model in statistical physics and becomes a very active research field on its own over the last sixty years. Our goal is to develop new methods in percolation suitable for our study of the connectivity and other relevant thresholds of random geometric graphs for general underlying distributions and in general dimension, in order to promote their applicability in data analysis.

2 From rare event to exceptional behavior

Large deviation and multifractal analysis bear many similarities. Large deviation theory is the study of rare events, while multifractal analysis is the study of exceptional behaviour. Incidentally, exceptional behaviour is quantified by computing the probability of rare events.

Let us be more concrete in the setting of stochastic processes. A process is the time evolution of a random phenomenon. The so-called typical behaviour is what we can observe *almost all the time*. Multifractal analysis is concerned with quantifying non-typical or exceptional behaviour by measuring the thinness (fractal dimensions) of the exceptional time sets.

The key towards carrying out multifractal analysis lies in interpreting the exceptional behaviour of interest in terms of a sequence of rare events along different scales, all the way to zero if the behavior of interest is local, or to infinity if the behaviour of interest is macroscopic.

A famous example is given by the fast points of Brownian motion. Here, by law of iterated logarithm, we know that typical local modulus of continuity of Brownian paths is $\sqrt{2h\log\log(1/h)}$, while Lévy's uniform modulus of continuity states that with probability one, there exists at least one point t at which the local growth is equivalent to $\sqrt{2h\log(1/h)}$. In this example, a point is called γ-fast if the Brownian path grows as $\gamma\sqrt{2h\log(1/h)}$ around the point, for some $\gamma \in (0,1)$. Determining the fractal dimensions of the γ-fast points is the goal of multifractal analysis.

Notice that in the previous example exceptionality is a kind of local geometric constraints on the Brownian paths. There could be a plethora of choices on the constraints that one might think of depending on the context. Moreover, the setting can be generalized to multifractal analysis of random fields, random measures, random metric spaces etc.

My goal in this area is to develop probabilistic and fractal geometric tools for identifying the typical and quantifying the exceptional behaviour of multifractal random structures.

3 From probability to number theory

How many $n \in \{1, ..., x\}$ are there such that the number of distinct prime divisors of n surpasses

$$\log\log(x) + 2\sqrt{\log\log(x)}\ ?$$

This seemingly difficult combinatorial question can actually be answered satisfactorily with probability theory. Indeed, Erdös and Kac showed that the number of distinct prime divisors of a uniformly chosen random integer behaves like a normal distribution with average and variance both equal to $\log\log(x)$. As a result, the answer to the previous question is approximately $0.02x$ as x gets larger and larger.

The Erdös-Kac theorem is one of many success stories of probabilistic reasoning in number theory. To have an idea of why the theorem holds, we let J_x denote a uniformly chosen sample from $\{1, ..., x\}$, then we realize that the number of prime divisors $\omega(J_x)$ can be written as the sum of indicators

$$\mathbb{1}(p \text{ divides } J_n) \sim \text{Bernoulli law of success probability} \approx \frac{1}{p}$$

with prime $p \leq x$. Provided that the correlations between distinct Bernoulli's are small, the central limit theorem heuristically leads to the conclusion, up to computing the mean and variance of $\omega(J_x)$.

There are lots of exciting central limit theorems in number theory, many of which proved by analytic tools such as Fourier analysis or complex analysis. Analytic methods are beautiful mathematics and often lead to strong statement by sophisticated manoeuvres. One possible drawback of analytic methods is that one might be easily buried in lengthy calculations and estimates without having an intuitive idea of why a certain result holds.

My main interest in this field is to develop intuitive probabilistic tools that lead to statements as strong as, if not stronger, the ones obtained via analytic methods. In particular, I am interested in obtaining optimal rates of Gaussian, Poisson and other distributional approximation in terms of various distances between probability measures.

4 Generalist central limit theorems by Malliavin's stochastic calculus of variations, and Stein's method

A common thread in the aforementioned topics is the probabilistic approximation of complicated random variables by simpler ones such as Gaussian or Poisson distributions. For the sake of concreteness, we present a simple case where

$$N_1, ..., N_n$$

are i.i.d standard normal random variables. We are interested in the fluctuation of the non-linear functional $F = F(N_1, ..., N_n)$ for a smooth $F : \mathbb{R}^n \to \mathbb{R}$. The stochastic calculus of variations, aka Malliavin calculus, is precisely the right tool for this problem. Indeed, if one is able to obtain moment estimates of the functional $\nabla F = \nabla F(N_1, ..., N_n) \in \mathbb{R}^n$, then one has an upper bound for the variance of F in view of the famous Poincaré inequality

$$\mathbb{V}\mathrm{ar}[F] \leq \mathbb{E}[|\nabla F|^2].$$

What's more, if one is able to obtain moment estimates for the Hessian matrix of F evaluated on the Gaussians, then one also has an upper bound for the total variation distance between F and a standard Gaussian, therefore leads to quantified central limit theorems. The latter result is known as the second order Poincaré inequality which is the output of a fruitful interaction between calculus of variation techniques and Stein's method for distributional approximations.

A fair amount of my recent works is devoted to proving quantified central limit theorems by the use of general principles which is alternative to the second order Poincaré techniques. The underlying randomness can be a general Gaussian random field, a Poisson point process, or more generally the (possibly infinite dimensional) stationary distribution of a Markov operator.

The obtained generalist CLT often has a wide range of applicability because of its high level abstraction of general principles for distributional approximation. Also, in contrast to the second order Poincaré estimates, only one derivative is needed in our approach which further widens the applicability. Examples include but not limited to fluctuation results for stochastic geometric models mentioned in Section 1, or number theoretical results in Section 3.