Les mathématiques de l'apprentissage machine
« Atelier sur les statistiques combinatoires »
30 avril au 4 mai 2018

Mathematics of Machine Learning
"Workshop on Combinatorial Statistics"
April 30 - May 4, 2018

Recovering a hidden hamiltonian cycle via linear programming

Yihong Wu*

yihong.wu@yale.edu

One of the most pressing challenges in genomics is to reconstruct a long and contiguous DNA sequence from short DNA subsequences (contigs). Enabled by very recent developments in sequencing technology one can now obtain linkage information that is statistically correlated with the true contig ordering, so that many links are observed between neighboring pairs of contigs, while relatively few links are observed for non-neighboring pairs.

Representing contigs as vertices and observed numbers of links between contigs as edge weights, the problem of contig assembly reduces to a Travelling Salesman Problem (TSP) in a weighted complete graph of size n with a hidden Hamiltonian cycle corresponding to the true ordering. We assume a statistical model where the edge weights on the hidden Hamiltonian cycle are drawn independently from a distribution P_n, while the remaining edge weights are drawn independently from Q_n. Despite the worst-case intractability of solving the TSP, we show that a simple linear programming (LP) relaxation, namely the fractional 2-factor (F2F) LP, recovers the hidden Hamiltonian cycle with probability tending to one as *ntoinfty* provided that $alpha_n - logn \to \infty$, where $\alpha_n \triangleq -2 \log \int \sqrt{dP_n dQ_n}$. This condition is information-theoretically optimal in the sense that, under mild distributional assumptions, $\alpha_n \geq (1+o(1) \log n$ is necessary for any algorithm to succeed regardless of the computational cost.

The analysis relies on the combinatorial characterization (in particular, the half-integrality) of vertices of the F2F polytope and the representation of extremal solutions as bicolored balanced multi-graphs, which can be further decomposed into simpler "blossom-type" structures whose statistical deviation can be controlled.

This is joint work with Vivek Bagaria (Stanford), Jian Ding (Penn), David Tse (Stanford), and Jiaming Xu (Purdue).

*Department of Statistics and Data Science, Yale University, 24 Hillhouse Ave, New Haven, CT 06511, USA