Projects for Stat375: A few suggestions

Each project will include a survey of the existing literature and a part of research work. This can in turn be either mathematical analysis or implementation experimentation. Roughly speaking some of the proposals below are ‘tools-oriented’ (implement/improve/analyze a method), while others are ‘application oriented’.

We are available for more information for each of these proposals that can be further specified/splitted. Finally, these are just suggestions, any alternative (motivated) idea is welcome!

Inference on phylogenies. Phylogenies describe the evolution and relationships among genomic sequences of different species. They can in turn be modeled as graphical models. The project aim is to study the application of approximate inference algorithms in this context.

Group testing. In group testing, a set of i.i.d. Bernoulli random variables $X_1, \ldots, X_n \in \{0, 1\}$ with $X_i \sim \text{Bernoulli}(p)$ is probed by measuring $Y_a = \bigvee_{i \in \partial a} X_i$ where $a \in [m]$ and $\partial a$ is a subset of $n$. It is in general NP hard to infer the $X_i$’s given the observed values of $\{Y_a\}$. The aim will be to study the tradeoff between inference accuracy, bias parameter $p$, and number of observations per variables $m/n$.

Both the behavior under exact (NP-hard) inference and for low complexity algorithms are of interest.

Message passing for constraint satisfaction problems. A instance of a binary constraint satisfaction problem involves $n$ Boolean variables $(x_1, \ldots, x_n)$ and $m$ logical constraints on them, each involving $k$ variables. The uniform measure over solutions of such a problem can be regarded as a graphical model, and belief propagation can be applied to such a model.

The information thus gathered can be exploited to find a solution. The objective of this project is to explore efficient ways of exploiting such an information, and ways to improve existing schemes. An alternative research direction consists in understanding the performances of other message passing algorithms such as survey propagation.

Loop corrections. Belief propagation is exact on trees. Several alternative algorithms have been put forward to deal with loops. Among the others, A.M. and T. Rizzo proposed a ‘loop corrected cavity method’, and M. Chertkov and V. Chernyak proposed a ‘loop series’ or ‘loop calculus’ method. This project aims at investigating these approaches, in particular: (i) What is their relationship? (ii) How to implement them efficiently? (iii) Can one provide any performance guarantees?

Linear programming and Max-Product algorithm. In a few combinatorial optimization problems, such as max-weight matching, $b$-factor, max-weight independent set, it has been shown that Max-Product is roughly as powerful as linear-programming relaxations. On the other hand, Max-Product has a few advantages: (i) It naturally exploits sparsity; (ii) It is distributed; (iii) It is extremely easy to implement.

Results of this type have been proved (among the others) by M. Bayati, D. Shah and collaborators. The objective of this project is to understand whether such results can be generalized.

Variational gaussian approximation (aka log-det relaxations). The Gibbs variational principle can be used to derive approximate variational algorithms. The basic idea is to use simple treatable families of distributions as ‘test-distributions’. Naive mean-field theory amounts
to using factorized distribution. An interesting alternative is to use gaussian distributions (studied, among others, by M. Wainwright and M. Jordan, and by L. El Ghaoui).

The aim of this project is develop suitable low complexity methods to optimize this variational approximation, as well as possible improvements of it.

**Continuous variables.** The focus of the course is on graphical models for discrete random variables $x_i$. This project would investigate extension to continuous variables. In particular, what can be good choices for the messages and the update rules in message passing? What can be said about convergence/correctness?

As a prototype problem one can consider computing the volume of a convex set. For instance, given an $n \times m$ sparse matrix $H$ with non-negative entries, and a vector $b \in \mathbb{R}_+^m$ one can consider the set

$$S = \{ x \in [0, 1]^n : H x \leq b \}. \quad (1)$$