Open problems SPOC and IDEPHICS 2023

Lenka Zdeborova

February 2023

What is this about?

List of open problems that will be looked at in the framework of the doctoral course in spring 2023. All of these problems can be readily addressed with the techniques covered in the lecture (or the master lecture Statistical Physics of Computation) and most of them can easily lead to a publication if sufficient work is put in.

Problem A: Distance L vertex cover

Vertex cover is a combinatorial problem where nodes of a graph are either full or empty. The rule is such that every edge needs to have at least one of its nodes full. A typical question is the size of the smallest vertex cover, i.e. how many nodes have to be full? The problem is often motivated by the guard in the museum that sits in the crossroads and guards the corridors. This problem has been studied quite extensively in the statistical physics literature on random graphs, see e.g. [3] and references therein.

Consider a distance-2 version of the vertex cover problem where an edge is considered covered if one of its nodes or one of their neighbors are full. The question is the same, what is the minimal number of nodes one needs to make full? Considering the problems on d-regular random graphs is the simplest case.

A paper where a related (corresponding to distance L vertex cover) problem is treated is [7].

Problem B: Distance L independent set

An Independent set is a combinatorial problem where nodes of a graph are either full or empty in such a way that two full nodes cannot be neighbors. As such the problem of independent sets and vertex cover are dual to each other, it is enough to switch between full and empty nodes.

Here we consider a distance-2 version of the independent set where if a node is full then its neighbors and also second neighbors have to be empty. What is the maximum number of nodes one can fill in a random d-regular graph? This problem can be solved in a way related to the model of [4, 12].

A distance L version is then related to the problem is treated in [7].

Problem C: Mixture of linear regression models

This is a variant of the classical linear regression where there are k possible ground truth functions (teachers) producing the output labels. Let us denote the corresponding teachers weights w^*_{il}, where $l = 1, \ldots, k$, and $i = 1, \ldots, d$ is the dimensionality. Input data $X_{\mu i}$ with $\mu = 1, \ldots, n$, $i = 1 \ldots, d$ are assumed to be iid Gaussian to enable analysis. For each sample μ we have a one-hot encoded vector $v^*_{\mu l} = 1$ if for sample μ the teacher vector l is chosen, and $v^*_{\mu l} = 0$ otherwise. The labels are then generated as

$$y_\mu = \sum_{l=1}^{k} v_{\mu l} \sum_{i=1}^{d} X_{\mu i} w_{il} + \xi_\mu$$

, where ξ_μ is Gaussian additive noise of variance Δ. The prior on the teacher vectors $P(w_{il})$ can be Gaussian or sparse Gauss-Bernoulli. The prior on the v_μ is uniform.

The goal is to estimate w^*, and v^* from the observations on n pairs $X_\mu \in \mathbb{R}^d, y_\mu \in \mathbb{R}$.

This is a generalization of the generalized linear model [2] that structurally resembles the committee machine [1]. The calibration in compressed sensing [6] is also mathematically related.

Problem D: Randomly sparse linear regression

This is another variant of the classical sparse linear regression where the sparsity pattern is different for every sample.

Let us denote the corresponding teachers' weights w^*_i, where $i = 1, \ldots, d$ is the dimensionality. Let us denote $s_{\mu i} \in \{0, 1\}$ the sparsity pattern for sample μ, assume that these numbers are drawn iid from the Bernoulli distribution of density ρ. We assume the input data $X_{\mu i}$ with $\mu = 1, \ldots, n$, $i = 1 \ldots, d$ to be iid Gaussian to enable analysis. The labels are then generated as

$$y_\mu = \sum_{i=1}^{d} s_{\mu i} X_{\mu i} w_i + \xi_\mu,$$

where ξ_μ is Gaussian additive noise of variance Δ. The prior on the teacher vectors $P(w_{il})$ is Gaussian.

The goal is to estimate w^*, from the observations on n pairs $X_\mu \in \mathbb{R}^d, y_\mu \in \mathbb{R}$. The matrix $s_{\mu i}$ is not given.

This is a generalization of the generalized linear model [2]. The sparsity pattern $s_{\mu i}$ can be seen as a type of intrinsic noise, thus already the version with $\Delta = 0$ should be non-trivial. One interesting question is whether there is enough information in the labels to estimate something about $s_{\mu i}$ for the training set to improve the estimation of the teacher vector beyond the accuracy obtained if the sparsity is replaced by an effective noise.

Problem E: Random satisfiability modulo theory

The random K-satisfiability problem is one of the classical problems studied with the tools of statistical physics [10]. Consider the classical random sat formula with a twist where the Boolean satisfying assignment y must be in the range of a simple generative neural network $y = \text{sign}(Xw)$ where $X \in \mathbb{R}^{n \times p}$ is a random Gaussian matrix and $w \in \mathbb{R}^p$ is a vector or real-values latent variables. The SAT formula is then constructed in the standard manner over y.

In theoretical computer science, this variant of the K-satisfiability problem where variables are in fact inequalities (half-spaces) is known as the satisfiability modulo theories.

Problem F: Graph-fused LASSO

The standard LASSO (aka compressed sensing) that was studied using tools of statistical physics considered random Gaussian matrix X and observations $y = XW$ where w is sparse.

A graph-fused LASSO is a variant where the sparsity is structured and there is a graph constructed over the element of w with edges being more likely between two elements that have similar values.

One simple case is when the element of w are binary and the graph is then an instance of the stochastic block model (sparse or dense) with one group much smaller than the other.

Problem G: Sampling with Stochastic Localization

Stochastic Localization is an amazing tool (very close to the diffusion model in machine learning) when coupled with message passing. In fact, it has the potential to be an alternative to MCMC (Monte-Carlo sampling).

Say that our goal is to sample from the Boltzmann mesure $\mu(\mathbf{x}) = \frac{1}{Z} e^{-\beta H(\mathbf{x})}$. The idea is to consider instead the tilted measure

$$\mu_t(\mathbf{x}) = \frac{1}{Z}(t, y) e^{-\beta H(\mathbf{x}) + \mathbf{x}.\mathbf{y} - \frac{t}{2}\|\mathbf{x}\|_2^2} \tag{1}$$

where the vector y just follow a Brownian motion $\mathbf{y}_{t+1} = \mathbf{y}_t + \mathbf{m}(t)\delta_t + \xi\delta_t$ and the vector m is the marginal of the tilded measure $\mathbf{m}(t) = \int d\mathbf{x}\mathbf{x}\mu_t(\mathbf{x})$. Stochastic localisation (ELDAN 2013) insure that as $t \to \infty$ the vector μ_t converge to a perfect random sample of the measure $\mu(\mathbf{x})$.

The project is to implement this sampling method! Can one use it to sample from sample uniformly in simple models such as Curie-Weiss, the RFIM, or even diluted systems such as ferromagnet in a random field? In these cases the computation of $\mathbf{m}(t) = \int d\mathbf{x}\mathbf{x}\mu_t(\mathbf{x})$ can be done with BP or mean-field techniques.

Problem H: How many samples to recognize two Gaussians

Consider two d-dimensional Gaussian clowds $\mathcal{N}(\pm\mu/\sqrt{d}, 1)$ with random centroid μ. It is well-known how large the norm of the centroids must be for the two clouds to be distinguishable from a single one $\mathcal{N}(0, 1)$ [8]. Now consider that we want to use a perceptron classifier to classify points generated from the mixture of two Gaussians $y = 1$, from the one generated by a single Gaussian $y = -1$. How many samples are needed for that? This paper may be useful [11]. Compare to the Bayes optimal performance.

Problem I: Prove the finite temperature replica equation for linear models in machine learning

The paper [9] proves a very generic replica equation at zero temperature for linear models. However, the finite temperature case is open. This should be doable by the rigorous cavity method (adding one variable at a time) and is an interesting and direct application of our methods to a simple Bayesian machine learning problem. The problem is presented in [5].

References

[1] Benjamin Aubin, Antoine Maillard, Jean Barbier, Florent Krzakala, Nicolas Macris, and Lenka Zdeborová. The committee machine: Computational to statistical gaps in learning a two-layers neural network. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124023, 2019.

[2] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

[3] Jean Barbier, Florent Krzakala, Lenka Zdeborová, and Pan Zhang. The hard-core model on random graphs revisited. In *Journal of Physics: Conference Series*, volume 473, page 012021. IOP Publishing, 2013.

[4] Giulio Biroli and Marc Mézard. Lattice glass models. *Physical review letters*, 88(2):025501, 2001.

[5] Lucas Clarté, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. A study of uncertainty quantification in overparametrized high-dimensional models. *arXiv preprint arXiv:2210.12760*, 2022.

[6] Marylou Gabrié, Jean Barbier, Florent Krzakala, and Lenka Zdeborová. Blind calibration for compressed sensing: state evolution and an online algorithm. *Journal of Physics A: Mathematical and Theoretical*, 53(33):334004, 2020.

[7] Alberto Guggiola and Guilhem Semerjian. Minimal contagious sets in random regular graphs. *Journal of Statistical Physics*, 158(2):300–358, 2015.

[8] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(7):073403, 2017.

[9] Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151, 2021.

[10] Marc Mézard, Giorgio Parisi, and Riccardo Zecchina. Analytic and algorithmic solution of random satisfiability problems. *Science*, 297(5582):812–815, 2002.

[11] Francesca Mignacco, Florent Krzakala, Yue Lu, Pierfrancesco Urbani, and Lenka Zdeborova. The role of regularization in classification of high-dimensional noisy gaussian mixture. In *International conference on machine learning*, pages 6874–6883. PMLR, 2020.

[12] Olivier Rivoire, Giulio Biroli, Olivier C Martin, and Marc Mézard. Glass models on bethe lattices. *The European Physical Journal B-Condensed Matter and Complex Systems*, 37(1):55–78, 2004.