Boltzmann machines

organized by
Asja Fischer, Guido Montufar, and Jason Morton

Workshop Summary

Introduction

Boltzmann machines (BMs) are a type of probability models that belong to the class of undirected probabilistic graphical models also known as Markov random fields. They have been extremely important in machine learning (ML), playing a central role in the come-back of neural networks and the rise of deep learning, and have continued to play a central role in theory. They provide a window to numerous generalizations of simpler mathematical models and approaches that have been treated in information geometry and algebraic statistics. An example is the implicitization problem, where a set of probability distributions is expressed in terms of algebraic conditions that need to be satisfied among individual probabilities. Therefore, BMs offer a rare opportunity to combine several fields of mathematics.

Lectures

We had a small number of presentations by workshop participants, summarized in the following.

- **Introduction to Restricted Boltzmann Machines (RBMs)**, by Guido Montúfar: Overview of recent results on the geometry of the sets of probability distributions representable by RBMs, with a theoretical focus.
- **RBMs: applications in Machine Learning and Quantum Computers**, by Jason Rolfe: Overview on applications of RBMs and the role they play in quantum computers.
- **Algebraic statistics of RBMs**, by Anna Seigal: Introduction to the algebraic implicitization problem and discussion of recent advances and consequences for RBMs.
- **Information geometry of Boltzmann Machines**, by Ryo Karakida: Presentation of results on the critical points of contrastive divergence in binary Gaussian RBMs and their interpretations in terms of component analysis. Contrastive divergence is the most common learning algorithm for RBMs.
- **EM algorithm critical points**, by Kaie Kubjas: Presentation of recent results on stationary points of iterative optimization methods (EM algorithm) for a type of graphical models with hidden variables.
- **Optimal transport and information geometry**, by Wuchen Li: Introduction to optimal transport and the geometric structures that derive from there in connection with information geometry and estimation in probability models.
- **Hopfield networks**, by Christopher Hillar: Overview about recent results on the dynamics and robust storage capacity of Hopfield networks and description of a series of open problems. Hopfield networks are deterministic variants of Boltzmann machines.

Problems

In preparation for the workshop, we collected interesting open problems, as is the standard practice at AIM, and prepared overview and introductory materials in connection to the areas represented among participants (Restricted Boltzmann Machines: Introduction and Review https://arxiv.org/abs/1806.07066, which describes connections between RBMs and various areas of mathematics, and includes a list of 15 open problems).

In order to ensure the most productive use of time during the workshop, we contacted several of the participants ahead of time, to start discussing research questions of joint interest and possible collaborations. Before and during the workshop we received many contributed problems (a summary can be found at the open problems site http://aimpl.org/boltzmann/, edited by Thomas Merkh) and groups quickly organized to work focused and collaboratively on some of these, as outlined in the following section.

Working groups

The main groups and topics of discussion were the following.

- **Representational Capacity**. What is the smallest number of hidden variables for which an RBM is a universal approximator?

  The working group considered the question of whether RBM$(4, 3)$ i.e. an RBM with 4 visible and 3 hidden variables, is a universal approximator; that is, whether the set RBM$(4, 3)$ can get arbitrarily close to any point in the 15-dimensional probability simplex. Guido Montufar showed in his thesis that RBM$(4, 3)$ can attain any probability distribution supported on eight disconnected vertices of the 4D hypercube; the group decided to investigate whether RBM$(4, 3)$ could attain similar probability distributions supported on eight disconnected vertices as well as a single additional vertex of the hypercube. We also came up with a proof outline that RBM$(4, 3)$ could achieve any probability distribution on eight vertices of the hypercube. Separate from this, we thought about trying to look more closely at RBM$(3, 3)$, which is known to be the smallest universal approximator for $n = 3$. Since RBM$(4, 3)$ has one more visible unit but the same number of hidden units, RBM$(3, 3)$ arises as a kind of marginalization of RBM$(4, 3)$, and we expect that studying the former could lead to insights about the latter.

  Summary based on a description by Leon Zhang.
- **RBM Optimization and Regularization**. Can we derive a regularization term corresponding to dropout in training RBMs?

  Dropout is an effective technique for regularizing deep neural networks. In this working group, we explored a mechanistic extension of dropout to RBMs, derived an analytic regularizer that is equivalent to dropout for RBMs, and considerd the impact of this regularization term on the learned representations. We also derived a lower-variance approximation to dropout for RBMs, which we plan to investigate experimentally in future. Finally, we discussed the existence of extensions that render dropout-RBMs more similar to deep networks of rectified linear units, and could allow to analyze the behavior of deterministic networks through the lens of the dropout-RBM regularizer.

  Summary description by Asja Fisher.

- **Topics related to Wasserstein distance and Optimal Transport Theory**. Here, the discussions focused on how to implement natural gradients for Boltzmann machines.

  The Wasserstein gradient offers an alternative to the Fisher natural gradient, which has been investigated for Boltzmann machines in information geometry. This group studied fast algorithms for Wasserstein natural gradient. To reduce the computational cost, the group investigated approaches to computing the metric by taking an average over empirical samples (and its neighboring states) instead of the model average. Following similar ideas from works on the Fisher-Rao natural gradient, the group studied low rank approximations for inverting the Wasserstein metric tensor. Several computational complexity issues are under investigation.

  Summary based on a description by Wuchen Li and Ryo Karakida.
- **Quantum vs classic Boltzmann Machines**. What are the relations between the classic BMs and the quantum version?

  This group worked on a comparison of the expressive power of a classic RBM and the corresponding complex RBM associated to quantum states. The first goal was to compare the probability distributions inside the probability simplex that can be obtained by these two interpretations of the RBM.

  Let us consider an RBM with N visible units (with $n_1, \dots, n_N$ states, respectively) and M hidden units (with $r_1, \dots, r_M$ states, respectively). The set of probability distributions on these variables are tensors of size $n_1 \times \cdots \times n_N$ with non-negative entries summing to one. A tensor lies in the classic RBM if it can be written as the Hadamard product (i.e., entry-wise product) of M tensors, whose non-negative ranks are at most $r_1, \dots, r_M$. The complex RBM consists of complex tensors of size $n_1 \times \cdots \times n_N$ which are Hadamard product of M tensors of complex rank at most $r_1, \dots, r_M$. In this way, the classic RBM can be viewed as a restriction of the complex RBM. However, we interpret the complex RBM as representing a space of pure quantum states.

  For each quantum state, i.e. complex tensor of size $n_1 \times \cdots \times n_N$, we can construct its density matrix by taking the outer-product of the state with its complex conjugate. Such a density matrix is Hermitian and positive semi-definite, with trace equal to one. If we restrict to the diagonal of the density matrix, we obtain a probability distribution inside the (classical) probability simplex. This subspace of the probability simplex is a linear projection of the space of density matrices of quantum states given by the complex RBM - it is different than the subset of the simplex obtained by restricting the complex RBM to the simplex.

  As a first result, this group obtained that every probability distribution in the classic RBM is the diagonal of a density matrix of a quantum state in a corresponding complex RBM.

  Summary description by Anna Seigal, Alessandro Oneto, and Dimitri Marinelli.
- **Infinite Boltzmann machines**. What kind of objects do we obtain in the continuous limit of variables?

  This working group discussed exchangeable RBMs with infinitely many nodes, binary RBMs, and learning landscapes. They discussed about generalizing Ryo Karakida's work to sparse binary RBMs with random weights using random matrix theory. Furthermore, the group discussed the loss landscape of the binary RBMs and properties of its global minima. One approach is to expand the likelihood function by taking a perturbation around diagonal weight matrices. We also discussed the solution of RBMs with sparse connectivities and, in particular, challenged to reveal how the sparseness modulates the solution in the analytically tractable case, i.e., a Gaussian-Gaussian RBM.

Summary based on a description by Ngoc Tran and Ryo Karakida.

- **Tropical RBMs**. What are possible characterizations of the tropical version of the RBM?

  This group discussed connections between tropical geometry and RBMs (or neural networks in general). We discussed the tropical degree. In a follow up this leaded to some work on the question of: how to make deep networks with rectified linear units (ReLUs) more efficient (fewer nodes and higher representational capacity of, say, tropical polynomials), using operations other than convolutions. This work is built on connections between tropical rational functions of Lek-Heng Lim et al. and Ngoc Tran's work on factorization of tropical rational functions.

  Summary based on a description by Ngoc Tran.

We had brief reports on the advances of the working groups each day. Before each working session, we had a brief discussion among all participants to decide on the problems to be worked on during the session. Mostly, the working groups persisted throughout the week.

Final comments and future plans

The workshop gave us a unique opportunity to explore connections and synergies between various areas of mathematics, and to make advances on a number of concrete problems in relation to the representational power, optimization, and regularization of Boltzmann machines. Some of the areas represented in the workshop were machine learning, applied algebra, algebraic statistics, tropical geometry, information geometry, optimal transport, optimization, sampling, inference.

As a result of the interactions and focused discussion during the workshop, a number of new collaborations emerged. Several of the working groups plan to continue working on the problems that they started to work on during the workshop. Several participants have organized for mutual

- Exploration of Caratheodory numbers of exponential families.
- Exploration of the limitations in representational power of Boltzmann machines emerging from sparse connectivity in the graph of the model.
- Exploration of a technique called "dropout" as a regularizer in training restricted Boltzmann machines.
- Exploration of the maximal divergences from data distributions to Boltzmann machines.
- Exploration of the EM algorithm and its critical points for the binary Boltzmann machine.