ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

Understanding and improving unsupervised training of Boltzmann machines

Gorka Muñoz-Gil[1], Alejandro Pozas-Kerstjens[1], Miguel Angel Garcia-March[1], Maciej Lewenstein[1,2] and Przemysław R. Grzybowski[3] *

1- ICFO-Institut de Ciencies Fotoniques,
The Barcelona Institute of Science and Technology,
08860 Castelldefels (Barcelona), Spain

2- ICREA, Passeig Lluis Companys 23, 08010 Barcelona, Spain
3- Faculty of Physics, Adam Mickiewicz University,
Umultowska 85, 61-614 Poznań, Poland

Abstract. We have analyzed the training of Boltzmann machines under the perspective of statistical physics. We argue that training models in spin-glass regime is highly inefficient and unnecessary. To that end, previously we have presented RAPID, a method to control the frustration of spin models and to train them without the need of expensive sampling methods. In this contribution we study effects of initialising Boltzmann machines in easily sampling regime and training with standard methods.

1 Introduction

One of the most important models in unsupervised learning are Boltzmann machines (BMs) [1]. Training of Boltzmann machines requires calculating a set of averages with respect to data and model for every learning step. In general, the these averages cannot be computed exactly for large models due to their large dimensionality, but they can be estimated by sampling through Markov chain Monte Carlo (MCMC) methods. Unfortunately, the sampling of BMs is hard and MCMC algorithms are numerically costly. At present, the most popular BMs are restricted Boltzmann machines (RBMs) with only one layer of hidden neurons and no intra-layer connections. This architecture usually allows for acceptable learning by means of simple MCMC-based algorithms such as Contrastive Divergence (CD). Improving sampling offers huge learning benefits and more numerically demanding methods, like Persistent Contrastive Divergence (PCD) [2] and Population Annealing [3], have been developed to this end.

From the point of statistical mechanics *any BMs initialised in a standard way* is equivalent to Sherrington-Kirkpatrick spin-glass (SKSG) model [4]. We identify such initialisation in the SK spin-glass regime as unncecesary bottleneck in training of Boltzmann machines. In recent paper [5] we taken a radically different approach: we regularized the couplings in the Boltzmann machine in order to avoid a spin-glass behavior at any point of training. We also used a new sampling technique, related to that regulariation. We shown that the proposed method allows for efficient learning and generalization. Furthermore, we shown an improvement in training speed of orders of magnitude.In present

*Corresponding authors: gorka.munoz@icfo.eu, alejandro.pozas@icfo.eu, grzyb@amu.edu.pl

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

contribution we study initialisation of RBMs in non-SKSG regime in conjunction with standard CD and PCD learning algorithms.

2 RAPID: Regularized Associations and Pattern-InDuced correlations

In this section we briefly recapitulate basics of our method. We start by recalling the standard Boltzmann machine, which consists of N binary neurons $\boldsymbol{\sigma}$ (here we use values $\sigma_j = \pm 1$ which are standard in physics of spin systems), which can be separated into disjoint sets of visible and hidden neurons, $\boldsymbol{\sigma} = (\boldsymbol{v}, \boldsymbol{h})$. An energy is associated to every configuration of neurons $\boldsymbol{\sigma}$ via an energy function, $E_\theta(\boldsymbol{\sigma}) = -\sum_{ij}^{N} W_{ij}\sigma_i\sigma_j - \sum_{i}^{N} b_i\sigma_i$, where the weights W_{ij} describe neuron-neuron connections, or *associations*, and b_i are local biases. The probability of having a visible configuration $\boldsymbol{v}$ is given by a Boltzmann distribution $P_\theta(\boldsymbol{v}) = \sum_{\boldsymbol{h}} e^{-E_\theta(\boldsymbol{\sigma})} / \sum_{\boldsymbol{\sigma}} e^{-E_\theta(\boldsymbol{\sigma})}$. Since the main problems we discuss are related to the distribution of weights W_{ij}, in the following we will neglect the biases b_i. The goal of training is to determine the parameters θ such that P_θ represents as close as possible the distribution P^{data} underlying some dataset $\mathcal{T}$. For that minimizing the negative log-likelihood, $\mathcal{L}_\theta = -\sum_{\boldsymbol{v}\in\mathcal{T}} P^{\text{data}}(\boldsymbol{v}) \log P_\theta(\boldsymbol{v})$, with respect to the parameters of the model is usually employed. As P^{data} is independent of these variables, the minimization is only performed to $\log P_\theta(\boldsymbol{v})$. The derivative of this terms takes the form $\partial_{W_{ij}} \log P_\theta(\boldsymbol{v}) = \langle \sigma_i\sigma_j \rangle_{\text{data}} - \langle \sigma_i\sigma_j \rangle_{\text{model}}$, where the bracket $\langle\cdot\rangle$ denotes the expectation value with respect to either the training data in $\mathcal{T}$ or the model given by P_θ.

2.1 Regularized Associations

The SG phase is related to the so-called spin frustration, which occurs when there is no configuration that minimizes the energy of all interactions at the same time. With increasing frustration, the number of low energy minima grows exponentially [6]. Mattis solved the frustration problem in a very simple model [7]: choose one configuration (or *pattern*) $\boldsymbol{\xi} \in \{-1, 1\}^N$ at random, and define $W_{ij} = \xi_i\xi_j$. The unique ground state of the spin model defined by such couplings is $\boldsymbol{\xi}$. Here we employ a generalization of Mattis' approach, where the weights are constructed from an arbitrary number K of patterns $\boldsymbol{\xi}^{(k)}$:

$$W_{ij} = \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \xi_i^{(k)} \xi_j^{(k)}. \tag{1}$$

The form of the weights in Eq. (1) is well known in machine learning from the Hopfield model of associative memory [8, 9]. Although a main focus in the Hopfield model is on retrieval of fixed patterns from dynamics of neural networks instead of generalising data distributions (the case of BMs), both models are closely related [10]. The number of independent patterns that can be faithfully retrieved from a Hopfield model was thoroughly studied in [11]. For very low

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

temperatures, the Hopfield model is in "retrieval phase" if $K/N < 0.12$ and in the spin-glass phase where patterns cannot be faithfully retrieved for $K/N > 0.12$. Following those results, for a given training data problem, we experimentally choose K high enough to faithfully represent the data probability distribution, but keeping the ratio K/N below the threshold that will lead to a spin-glass behavior. In the case of RBMs, this ratio can be lowered arbitrarily for a desired number of patterns K by increasing the number of hidden neurons accordingly.

2.2 Training via Pattern-InDuced correlations

As mentioned above, it is easy to characterize spin models with weights of the form of Eq. (1) whenever one has a small number of patterns K: in fact, when $K \ll N$ the patterns $\boldsymbol{\xi}^{(k)}$ are low-energy configurations themselves. A consequence of this is that the model averages required for training, $\langle \sigma_i \sigma_j \rangle_{\text{model}}$, can be well approximated by

$$\langle \sigma_i \sigma_j \rangle_{\text{model}} \approx \frac{1}{K} \sum_{k=1}^{K} \xi_i^{(k)} \xi_j^{(k)}. \tag{2}$$

This suggests a natural procedure for minimizing $\mathcal{L}_\theta$: first, choose $K \ll N$ random patterns and compute the model's weights via Eq. (1); second, compute the derivatives of $\mathcal{L}_\theta$ with respect to each individual pattern component $\xi_i^{(k)}$ and replace all averages over the model by averages over the patterns, as in Eq. (2); third, update $\xi_i^{(k)}$ according to such gradients, and from them recompute new valid patterns.

We note that such procedure does not need any MCMC sampling, and this already gives good results in learning simple datasets. However, we have observed that when learning more complex datasets the best results were obtained when training with PID was complemented with a few steps of Gibbs sampling of the patterns $\boldsymbol{\xi}^{(k)}$.

3 Results

3.1 Training of RA-initialised machines

Previously we demonstrated superiority of RAPID over standard RBMs trained with CD or PCD showing importance of spin-glass control in BMs. However this begs a question - what will be preformance of RBM initialised out of SKSG and trained with PC or PCD. To this end we studied RA-initialised CD or PCD trained RBMs i.e. weights are initialised with RA recepie but after that patterns are not used at all and weight are changed with CD or PCD procedure.

We are studying of learning a small dataset, consisting of 4×4 images with full vertical stripes. The dataset contains a total of 16 inequivalent images. While this number of visible neurons in this data set may seem small, it allows to compute exactly the ground state for RBM with any number of hidden neurons (and here we use 1000 hidden). This in turn allows to compute the Gibbs

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

sampling accessibility: For given machine we perform 10 Gibbs sampling steps starting from random configuration and record lowest energy of encountered configurations. The ratio of such energy to the ground state energy is a measure of accessibility of low energy configurations during sampling.

In this dataset we contrast RAPID with training RA-initialised and SKSG-initialised RBMs trained with standard methods, namely CD and PCD with 10 steps of Gibbs sampling and (in the case of PCD) 2048 fantasy particles.

(a) (b)

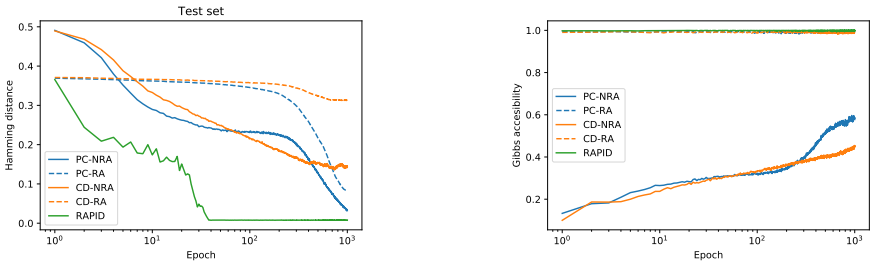

Fig. 1: Training of RA initialised RBMs. (a) Hamming distance between reconstructions of partial images and expected results in the test sets of the 4×4 Bars dataset. NRA denotes standard initialisation in spin-glass regime. (b) Gibbs sampling accessibility. In all cases, the models tested have $H = 1000$ hidden neurons, and are trained in the 4×4 stripes dataset. For the case training with RAPID (in blue), we employ $K = 8$ patterns).

To perform an accuracy test on the machines, we reconstruct corrupted images from the dataset. Then, we compute the Hamming distance (HD) between the reconstructed and the original image. In Figs. 1a we present results for the HD when reconstructing image from the training and test sets, respectively. As reported previously RAPID allows to learn the dataset in a surprisingly small number of epochs ($\sim$20 epochs) compared to the SKSG-initialised RBMs trained with CD or PCD methods ($\mathcal{O}(10^3)$ epochs). However unexpectedly RA-initialised RBMs trained with CD or PCD methods perform worse than SKSG-initialised ones although sampling in RA-initialised RBMs is greatly improved. To understand this we checked Gibbs sampling accesibility, presented on Figs. 1b. The SKSG-initialised machines start with very poor sampling and gradually go out of SKSG regime during learning. However RA-initialised RBMs enjoy perfect sampling, similar to RAPID. Their poor performance is associated with what we called "mirroring effect".

3.2 Mirroring effect

The mirroring effect occurs when *any* visible configuration after one-step Gibbs sampling remains in large part unchanged. It is closely related to, well known in

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

statistical physics, Onsager correction field in cavity method. Onsager realised that in a spin systems with long-range interactions, in the paramagnetic phase, the given spin orientation will slightly polarise other spins interacting with it. If, due to long-range interactions or large dimensionality, the number of such spins will be large the effective local field acting on polarising spin will substantially favour starting orientation. In terms of learning such fields can be problematic, hampering any training method based on Gibbs sampling. The problem is most important in the very begining phase of training. Once the network learns something the mirroring of random patterns is much weaker than reconstruction of learned features.

We observed mirroring effect for SKSG-initialised RBMs when number of hidden units was around 5000. It occurs that for RA-initialised RBMs the strong mirroing effect happens already for 1000 hidden. This is probably related to the lower frustration, as suggested by analytical results which we obtained for unfrustrated Mattis system.

In Fig. 1 we present results of mirroring effect (panel (a)) - the lower Hamming distance the stronger mirroring effect. On panel (b) we show also the evolution of local fields for RAPID and SKSG-initialised RBMs trained with CD or PCD. Interestingly one can observe a convergence of the latter to the RAPID - strengthening hypothesis formulated in previous paper that RAPID is the model of a *well trained* BM.

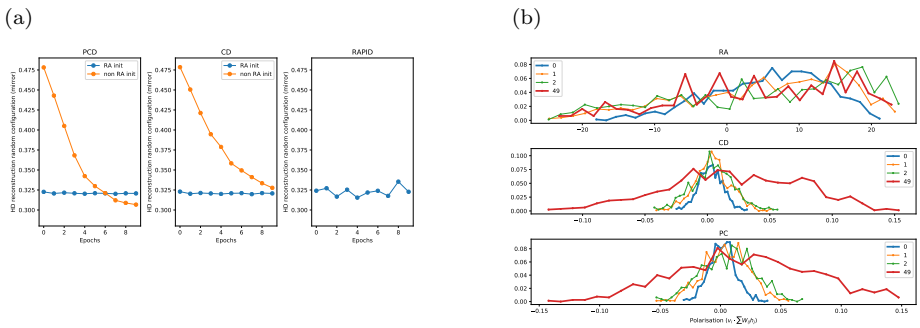

Fig. 2: Mirroring effect: (a) Hamming distance between *random* configuration and it one-step Gibbs sampling follower, for RA-initialised RBMs. In all cases, the models tested have $H = 1000$ hidden neurons, and are trained in the 4×4 stripes dataset. For the case training with RAPID (in blue), we employ $K = 8$ patterns) (b) Distribution of local fields acting on visible neurons obtained after one-step Gibbs sampling of random image, for RAPID and SKSG-initialised RBMs. Sign of field denotes it direction in respect with visible spin. Colors denote different epochs.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

4 Conclusions

We observed that simply initialising Boltzmann machine in non-spin-glass regime brings radical improvement of sampling, strengthening our claim that training models in spin-glass regime is highly inefficient and unnecessary. However reduced frustration may increase mirroring effect hampering training methods based on Gibbs sampling. This problem of course does not concerns RAPID since sampling in this method is based on patterns.

References

[1] G. E. Hinton and T. J. Sejnowski. Optimal perceptual inference. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 448–453, 1983.

[2] T. Tieleman. Training Restricted Boltzmann Machines using approximations to the likelihood gradient. In Proceedings of the 25th International Conference on Machine Learning, pages 1064–1071, 2008.

[3] K. Hukushima and Y. Iba. Population annealing and its application to a spin glass. AIP Conference Proceedings, 690(1):200–206, 2003.

[4] K. Binder and A. P. Young. Spin glasses: Experimental facts, theoretical concepts, and open questions. Rev. Mod. Phys., 58:801–976, 1986.

[5] Alejandro Pozas-Kerstjens, Gorka Muñoz-Gil, Miguel Ángel García-March, Antonio Acín, Maciej Lewenstein, and Przemysław R. Grzybowski. Efficient training of energy-based models via spin-glass control. arXiv e-prints, page arXiv:1910.01592, Oct 2019.

[6] K. Binder and A. P. Young. Spin glasses: Experimental facts, theoretical concepts, and open questions. Rev. Mod. Phys., 58:801–976, 1986.

[7] D. C. Mattis. Solvable spin systems with random interactions. Phys. Lett. A, 56(5):421 – 422, 1976.

[8] W. A. Little. The existence of persistent states in the brain. Math. Biosci., 19(1):101 – 120, 1974.

[9] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. PNAS, 79(8):2554–2558, 1982.

[10] A. Barra, A. Bernacchia, E. Santucci, and P. Contucci. On the equivalence of Hopfield networks and Boltzmann machines. Neural Networks, 34:1 – 9, 2012.

[11] D. J. Amit, H. Gutfreund, and H. Sompolinsky. Storing infinite numbers of patterns in a spin-glass model of neural networks. Phys. Rev. Lett., 55:1530–1533, 1985.