Efficient training of energy-based models via spin-glass control

Alejandro Pozas-Kerstjens[*] **Gorka Muñoz-Gil**[*]
Miguel Ángel García-March
ICFO–Institut de Ciencies Fotoniques
The Barcelona Institute of Science and Technology
08860 Castelldefels (Barcelona), Spain

Antonio Acín, Maciej Lewenstein
ICFO–Institut de Ciencies Fotoniques
The Barcelona Institute of Science and Technology
08860 Castelldefels (Barcelona), Spain
ICREA
Passeig Lluis Companys 23
08010 Barcelona, Spain

Przemysław R. Grzybowski
Faculty of Physics
Adam Mickiewicz University
Umultowska 85, 61-614 Poznań, Poland

Abstract

We present an efficient method for learning probabilistic graphical models. Inspired by statistical physics, we develop a technique that avoids the spin glass behavior of Boltzmann machines, allowing efficient sampling and training. We compare the proposed technique against standard training methods, and report improvements in the speed of training and retrieval of samples, and in generalization power.

1 Introduction

Machine learning has emerged as a disruptive technology transforming industries, society and science. In the long term, unsupervised learning is expected to be much more important than supervised learning, where most of the research and investment is put nowadays [1]. One of the most important models in unsupervised learning is Boltzmann machines, which have especially prospective properties. A Boltzmann machine with latent neurons allows for feature extraction, while its learning algorithm is remarkably simple [2].

Yet training of Boltzmann machines, as a particular instance of probabilistic graphical models [3], is hard due to the need of obtaining samples from the model built. In general, the averages required for training cannot be computed exactly for large models due to their large dimensionality, and even numerical methods like Markov Chain Monte Carlo are costly. This hardness is a consequence of the equivalence of Boltzmann machines with *spin glasses* (SG) [4]. It is widely known in statistical mechanics that determining the lowest-energy state of a SG is an NP-complete problem [5]. Despite this, many methods have been developed that are based in heuristics [6] and statistical

[*] These authors contributed equally to this work.

Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.

physics [7, 8, 9, 10]. In fact, the problem of sampling of Boltzmann machines is so important and difficult that task-specific hardware systems have been developed for it. These include systems operating in the regime of classical physics [11, 12, 13, 14], as well as based on quantum principles or hybrid classical-quantum machines [15].

A key point we raise in this paper is that Boltzmann machines reproducing typical training data probability distributions cannot be in a true spin glass phase. The fact that current methods of training of restricted Boltzmann machines (RBMs) lead to outputs that have a behavior closer to a spin glass than the training data they attempt to learn [16] points actually to a deficiency of such methods. The SK spin glass phase of Boltzmann machines thus constitutes an unnecessary bottleneck in learning.

Therefore, instead of pursuing the challenging problem of efficient sampling in spin glasses, we take a radically different approach: we constrain couplings in our models in such a way to avoid problems with a spin glass phase in the first place. This restriction suggests a new algorithm for estimating the gradient of the log-likelihood function, which we employ to train RBMs in various datasets. In all of the experiments reported, our combination of Restricted Associations (RA) and training by Pattern-InDuced correlations (PID) outperforms standard Contrastive Divergence and its variants both in training time and quality, showing remarkable generalization ability.

2 RAPID: Regularized Associations and Pattern-InDuced correlations

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 $\mathcal{T}$. This is usually done by 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. 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. Sampling from such distributions is the main challenge of Boltzmann machines as discussed previously. Restricted Boltzmann machines were introduced in order to facilitate sampling from $\langle\sigma_i\sigma_j\rangle_{\text{data}}$. However, even for RBMs a good estimation of $\langle\sigma_i\sigma_j\rangle_{\text{model}}$ is still very difficult if the weights are random and thus the system is in the Sherrington-Kirkpatrick (SK) spin glass phase [4].

2.1 Regularized Associations

Perhaps the most profound result stemming from the statistical physics perspective on Boltzmann machines is the understanding of the origin of the sampling hardness of such models. A Boltzmann machine with uniformly-random weights (the typical starting point when training) is equivalent to the Ising model with long-range random couplings, known as the SK spin glass model. Determining the lowest-energy configuration of a general SKSG model defined on a non-planar graph is known to be an NP-complete hard problem [5]. 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 [17]. Mattis solved the frustration problem in a very simple model [18]: 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 $\boldsymbol{\xi}$. Unfortunately, such construction is too constrained, and presents a too strong overfitting to $\boldsymbol{\xi}$. 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 [19, 20] which implements the Hebbian rule that "neurons wire together if they fire together" [21]. 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 [22]. The number of independent patterns that can be faithfully retrieved from a Hopfield model was thoroughly studied in [23]. For very low 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 $\boldsymbol{\xi}^{(k)}$.

The novelty of the combination of Restricted Associations and training via Pattern-InDuced correlations (RAPID) comes from avoiding the SK spin glass phase at any moment of training. That means that Eq. (1) is not novel in itself, but the way of utilizing is: for instance, when training RBMs, we would scale H to keep a low K/N ratio, relax pattern variables for easier learning and exploit regularized patterns for approximating the low-energy space in an efficient way.

3 Results

3.1 Benchmark with exact training: 4x4 Stripes

We first start by applying RAPID for learning a small dataset, consisting of 4×4 images with full vertical stripes. The dataset contains a total of 16 inequivalent images. While this dataset may seem small, it allows to compute exactly the denominator $\sum_{\boldsymbol{\sigma}} e^{-E_\theta(\boldsymbol{\sigma})}$ of the Boltzmann distribution. This enables the exact computation of the loss function $\mathcal{L}_\theta$, and thus to employ the exact gradients during training. Therefore, in this small dataset we can contrast RAPID with training via exact stochastic gradient descent. We additionally compare to standard methods for training RBMs, namely CD and PCD with 10 steps of Gibbs sampling and (in the case of PCD) 2048 fantasy particles.

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 and 1b we present results for the HD when reconstructing image from the training and test sets, respectively. RAPID allows to learn the dataset in a surprisingly small number of epochs ($\sim$20 epochs) compared to the rest of training methods ($\mathcal{O}(10^3)$ epochs). One may consider that, given the construction of the machine by means of the Hebbian rule, the machine would be prone to memorize the images of the dataset. However, seeing that the HD for the test set decreases in similar fashion to

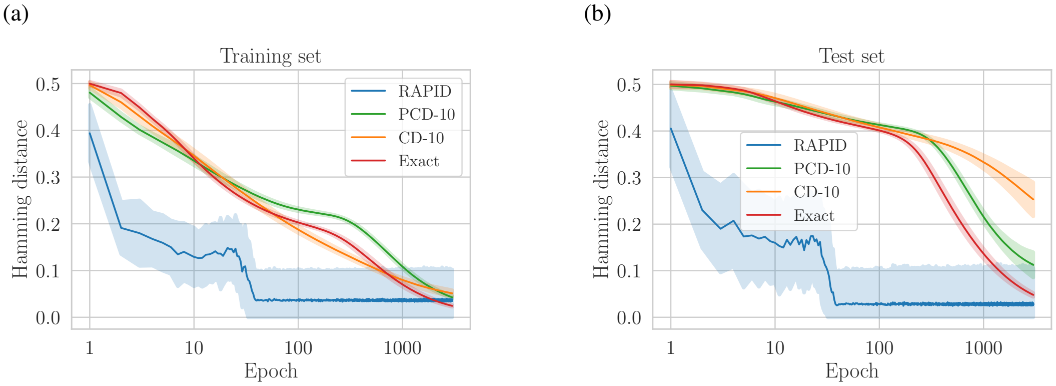

Figure 1: Training effectiveness of RAPID. Hamming distance between reconstructions of partial images and expected results in the (a) training and (b) test sets of the 4×4 Bars dataset. The shaded areas around the lines denote the standard deviation of 100 independent training instances. Images in the test set—a total of four—are never used during the training phase, and they are chosen so as to have no relation (via negation) to any images of the training set. 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).

the training set shows that RAPID allows to generalize in a faster way that the remaining methods, including training with the exact gradients of the log-likelihood.

3.2 Learning complex datasets

We now focus applying RAPID to learn more complex datasets. First, we consider the 12×12 Bars and Stripes dataset, which consists in 8188 images containing images of only vertical bars or horizontal stripes. As the complexity of the problem to solve increases, one needs to increase the number of auxiliary patterns K and, in order to keep the frustration of the model low, the number of hidden neurons H of the RBM. In Fig. 2a we show the results for the HD of reconstructed images, analogous to the graphs shown in Fig. 1. The HD of the reconstructions and the target images for the training and test sets decrease parallel to each other, proving that the model is indeed learning and not memorizing the images of the training set. Moreover, when generating samples of the model from free dynamics, we observe instances that were not contained in the training set, further demonstrating the generalization ability of RAPID-trained RBMs.

Next, we train our machine to generate the MNIST dataset. We show some samples returned after training in Fig. 2b. Despite the necessary increase in K and H for having enough expressivity without increasing frustration, we keep observing a speedup in training.

4 Conclusions

We have analyzed the training of Boltzmann machines under the perspective of statistical physics. Given that standard datasets do not exhibit exotic phenomena such as spin glass phases, we argue that training models with such phenomena is highly inefficient. To that end, we have presented RAPID, a method to control the frustration of spin models and to train them without the need of expensive sampling methods. Furthermore, we have shown how restricted Boltzmann machines trained with RAPID outperform standard training techniques in both learning speed and generalization ability.

Although the experiments focused on restricted Boltzmann machines, RAPID is applicable to arbitrary Boltzmann machines. We thus expect that RAPID, or variations of it, prove important in the use of deep Boltzmann machines for unsupervised learning tasks.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436, 2015.

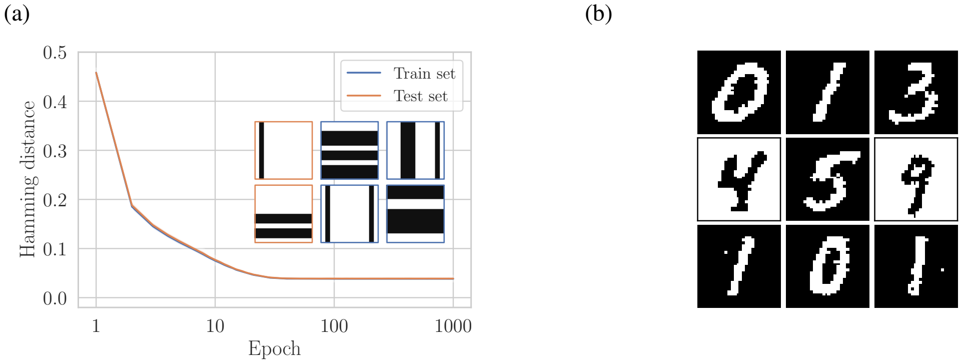

Figure 2: (a) Hamming distance between reconstructed images and expected results for the 12×12 Bars and Stripes dataset. The model employed is an RBM with 1000 hidden neurons, trained with 30-pattern RAPID. The shaded regions denote the standard deviations in 15 independent training instances. The inset shows instances sampled from the model. The leftmost samples, surrounded in red, were not part of the training set. (b) Samples drawn from an RBM trained with RAPID in the binarized-MNIST dataset. The RBM has 3000 hidden neurons, and the weights are constructed from $K = 200$ patterns.

[2] 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.

[3] C. M. Bishop. *Pattern recognition and machine learning*. Springer-Verlag New York, 2006.

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

[5] F. Barahona. On the computational complexity of Ising spin glass models. *J. Phys. A: Math. Gen.*, 15(10):3241–3253, 1982.

[6] 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.

[7] E. Marinari and G. Parisi. Simulated tempering: A new Monte Carlo scheme. *EPL*, 19(6):451–458, 1992.

[8] G. Desjardins, A. Courville, Y. Bengio, P. Vincent, and O. Delalleau. Tempered Markov Chain Monte Carlo for training of Restricted Boltzmann Machines. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 2010.

[9] F. Wang and D. P. Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.*, 86:2050–2053, 2001.

[10] Y. Du and I. Mordatch. Implicit generation and modeling with energy-based models. *arXiv:1903.08689*.

[11] M. Yamaoka, C. Yoshimura, M. Hayashi, T. Okuyama, H. Aoki, and H. Mizuno. A 20k-spin ising chip to solve combinatorial optimization problems with cmos annealing. *IEEE J. Solid-State Circuits*, 51(1), Jan 2016.

[12] T. Inagaki, Y. Haribara, K. Igarashi, T. Sonobe, S. Tamate, T. Honjo, A. Marandi, P. L. McMahon, T. Umeki, K. Enbutsu, O. Tadanaga, H. Takenouchi, K. Aihara, K. Kawarabayashi, K. Inoue, S. Utsunomiya, and H. Takesue. A coherent Ising machine for 2000-node optimization problems. *Science*, 354(6312):603–606, 2016.

[13] P. L. McMahon, A. Marandi, Y. Haribara, R. Hamerly, C. Langrock, S. Tamate, T. Inagaki, H. Takesue, S. Utsunomiya, K. Aihara, R. L. Byer, M. M. Fejer, H. Mabuchi, and Y. Yamamoto. A fully programmable 100-spin coherent Ising machine with all-to-all connections. *Science*, 354(6312):614–617, 2016.

[14] Sanroku Tsukamoto, Motomu Takatsu, Satoshi Matsubara, and Hirotaka Tamura. An accelerator architecture for combinatorial optimization problems. *FUJITSU Sci. Tech. J.*, 53(5):8–13, 2017.

[15] M. Benedetti, J. Realpe-Gómez, R. Biswas, and A. Perdomo-Ortiz. Quantum-assisted learning of hardware-embedded probabilistic graphical models. *Phys. Rev. X*, 7:041052, 2017.

[16] G. S. Hartnett, E. Parker, and E. Geist. Replica symmetry breaking in bipartite spin glasses and neural networks. *Phys. Rev. E*, 98:022116, 2018.

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

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

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

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

[21] D. Hebb. *The organization of behavior: A neuropsychological theory*. Wiley, 1949.

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

[23] E. Gardner. The space of interactions in neural network models. *Journal of Physics A: Mathematical and General*, 21(1):257–270, 1988.