Restricted Boltzmann Machines and Mean Field Approximations

A tutorial for VDSP-ESI Winter School 2020: Machine Learning in Physics.

Marylou Gabrié, Alia Abbara[1]
NYU, Flatiron Institute, LPENS

(Dated: February 9, 2020)

Corresponding jupyter notebook to be downloaded or forked from:
https://github.com/marylou-gabrie/tutorial-winter-school-ml-physics-rbm .

Restricted Boltzmann Machines (RBMs) are a type of generative model that is inspired from the famous Ising model in physics. For a binary data point $\mathbf{x} \in \{0,1\}^N$, a RBM defines the probability distribution

$$p(\mathbf{x}) = \frac{1}{\mathcal{Z}} \sum_{\mathbf{h}\in\{0,1\}^M} e^{-\beta E(\mathbf{x},\mathbf{h})} \tag{1}$$

$$E(\mathbf{x},\mathbf{h}) = -\sum_{i=1}^{N} a_i x_i - \sum_{\alpha=1}^{M} b_\alpha h_\alpha - \sum_{i=1}^{N}\sum_{\alpha=1}^{M} W_{i\alpha} x_i h_\alpha. \tag{2}$$

The weights $W_{i\alpha}$ (a.k.a. couplings) and the biases a_i and h_α (a.k.a. local magnetic fields) are the trainable parameters of the RBM. Differently from the traditional Ising model in physics the RBM has two different types of units (a.k.a. spins), the inputs x_i and the hidden (or latent) variables h_α. The benefice of the hidden layer is to mediate arbitrary interactions between all the input units whereas the traditional Ising model only includes pairwise-interactions. The representation power of RBMs is therefore much greater and we will see that they can for example model datasets of images.

RBMs are also interpreted as *undirected* neural networks with two layers. Given an input $\mathbf{x}$ the conditional probability of the hidden layer is

$$p(h_\alpha = 1|\mathbf{x}) = \sigma(\sum_{i=1}^{N} W_{i\alpha} x_i + b_\alpha) \tag{3}$$

with $\sigma(x) = (1+e^x)^{-1}$, and similarly the conditional probability of the input layer given the state of the hidden layer $\mathbf{h}$ is

$$p(x_i = 1|\mathbf{h}) = \sigma(\sum_{\alpha=1}^{M} W_{i\alpha} h_\alpha + a_i). \tag{4}$$

A. Maximum likelihood training

a. Training objective function

Given a training datasets of binary datapoints $\mathcal{D} = \{\mathbf{x}^{(k)}\}_{k=1}^{P}$, the parameters of the RBM $\{W, \mathbf{a}, \mathbf{b}\}$ can be adjusted by maximizing the probability that the model assigns to the training data, the likelihood, or equivalently its logarithm

$$\ell(W, \mathbf{a}, \mathbf{b}) = \ln \prod_{k=1}^{P} p(\mathbf{x}^{(k)}) = \sum_{k=1}^{P} \ln p(\mathbf{x}^{(k)}). \tag{5}$$

b. Gradient ascent

As often in machine learning the optimization is done through a stochastic gradient ascent. After a random initialization the parameters are updated in the direction of increasing gradients. For a minibatch of size 1 with the

training sample $\mathbf{x}^{(k)}$, we have

$$W_{i\alpha}^{t+1} = W_{i\alpha}^{t} + \eta \left.\frac{\partial \ell}{\partial W_{i\alpha}}\right|_{\mathbf{x}^{(k)}} \tag{6}$$

$$a_{i}^{t+1} = a_{i}^{t} + \eta \left.\frac{\partial \ell}{\partial a_{i}}\right|_{\mathbf{x}^{(k)}} \tag{7}$$

$$b_{\alpha}^{t+1} = b_{\alpha}^{t} + \eta \left.\frac{\partial \ell}{\partial b_{\alpha}}\right|_{\mathbf{x}^{(k)}} . \tag{8}$$

One can show that

$$\frac{\partial \ell}{\partial W_{i\alpha}} = \langle x_i h_\alpha \rangle_{\mathbf{x}^{(k)}} - \langle x_i h_\alpha \rangle \tag{9}$$

$$\frac{\partial \ell}{\partial a_{i}} = \langle x_i \rangle_{\mathbf{x}^{(k)}} - \langle x_i \rangle \tag{10}$$

$$\frac{\partial \ell}{\partial b_{\alpha}} = \langle h_\alpha \rangle_{\mathbf{x}^{(k)}} - \langle h_\alpha \rangle. \tag{11}$$

For a minibatch with more than one sample, gradients are averaged over the minibatch.

c. Intractable objective and gradients

Note that the partition function $\mathcal{Z}$ of the RBM does not have a simplified analytical expression. The summation over all the states of the input and hidden units is not realistically doable for RBMs with more than a few units. As a result, to learn the MNIST dataset we will need to resort to approximations.

B. Monte Carlo approximation

RBM became popular with an approximate Monte-Carlo approximation of the gradients introduced by Hinton[2]. The idea is to run a Markov chain from the training data points in the minibatch and average the final points to estimate the means and correlations appearing in the gradients (9)-(11).

> Launch jupyter notebook in the directory of the tutorial.

> We start by defining a python object, a class, to store and manipulate information about RBMs. Do you agree with the sampling functions using the conditional probabilities (3)-(4)?

> We give an example of implementation of the training by contrastive divergence of an RBM. Launch the fit for 50 epochs. How do we suggest to evaluate the quality of the training?

C. Naive mean-field approximation

1. Principle

The naive mean-field method consists in approximating the Boltzmann distribution by a fully factorized distribution over the degrees of freedom. Therefore, it ignores correlations between random variables. Among multiple methods of derivation, we present here the variational method. For the purpose of demonstration we consider an Ising model with spin variables $\mathbf{x} = (x_1, \cdots, x_N) \in \mathcal{X} = \{0,1\}^N$, and energy function

$$E(\mathbf{x}) = -\sum_{i=1}^{N} b_i x_i - \sum_{(ij)} W_{ij} x_i x_j = -\mathbf{b}^\top \mathbf{x} - \frac{1}{2}\mathbf{x}^\top \mathbf{W} \mathbf{x}\,, \quad \mathbf{b} \in \mathbb{R}^N\,, \quad \mathbf{W} \in \mathbb{R}^{N\times N}\,, \tag{12}$$

where the notation (ij) stands for pairs of connected spin-variables, and the weight matrix $\mathbf{W}$ is symmetric. The choices of $\{0,1\}$ rather than $\{-1,+1\}$ for the variable values, the notations $\mathbf{W}$ for weights (instead of couplings), $\mathbf{b}$ for biases (instead of local fields) are leaning towards the machine learning conventions and the interpretation of the

model as a Boltzmann machine (without hidden units). We denote by $q_{\mathbf{m}}$ a fully factorized distribution on $\{0,1\}^N$, which is a multivariate Bernoulli distribution parametrized by the mean values $\mathbf{m} = (m_1, \cdots, m_N) \in [0,1]^N$ of the different units (denoted by q_{m_i}):

$$q_{\mathbf{m}}(\mathbf{x}) = \prod_{i=1}^{N} q_{m_i}(x_i) = \prod_{i=1}^{N} m_i \delta(x_i - 1) + (1 - m_i)\delta(x_i). \tag{13}$$

We look for the optimal $q_{\mathbf{m}}$ distribution to approximate the Boltzmann distribution $p(\mathbf{x}) = e^{-\beta E(\mathbf{x})}/\mathcal{Z}$ by minimizing the KL-divergence

$$\min_{\mathbf{m}} \mathrm{KL}(q_{\mathbf{m}}||p) = \min_{\mathbf{m}} \sum_{\mathbf{x}\in\{0,1\}^N} q_{\mathbf{m}}(\mathbf{x}) \log \frac{q_{\mathbf{m}}(\mathbf{x})}{p(\mathbf{x})} \tag{14}$$

$$= \min_{\mathbf{m}} \sum_{\mathbf{x}\in\{0,1\}^N} q_{\mathbf{m}}(\mathbf{x}) \log q_{\mathbf{m}}(\mathbf{x}) + \beta \sum_{\mathbf{x}\in\{0,1\}^N} q_{\mathbf{m}}(\mathbf{x})(E(\mathbf{x})) + \log \mathcal{Z} \tag{15}$$

$$= \min_{\mathbf{m}} \ \beta G(q_{\mathbf{m}}) - \beta F \geq 0, \tag{16}$$

where the last inequality comes from the positivity of the KL-divergence. For a generic distribution q, $G(q)$ is the *Gibbs free energy* for the energy $E(\mathbf{x})$,

$$G(q) = \sum_{\mathbf{x}\in\{0,1\}^N} q(\mathbf{x})(E(\mathbf{x})) + \frac{1}{\beta} \sum_{\mathbf{x}\in\{0,1\}^N} q(\mathbf{x}) \log q(\mathbf{x}) = U(q) - H(q)/\beta \geq F, \tag{17}$$

involving the average energy $U(q)$ and the entropy $H(q)$. It is larger than the true free energy F except when $q = p$, in which case they are equal. Note that this fact also means that the Boltzmann distribution minimizes the Gibbs free energy. Restricting to factorized $q_{\mathbf{m}}$ distributions, we obtain the naive mean-field approximations for the mean value of the variables (or *magnetizations*) and the free energy:

$$\mathbf{m}^* = \arg\min_{\mathbf{m}} G(\mathbf{m}) = \langle \mathbf{x} \rangle_{q_{\mathbf{m}^*}} \tag{18}$$

$$F_{\mathrm{NMF}} = G(\mathbf{m}^*) \geq F. \tag{19}$$

The choice of a very simple family of distributions $q_{\mathbf{m}}$ limits the quality of the approximation but allows for tractable computations of other observables, for instance the two-spin correlations $\langle x_i x_j \rangle_{q_{\mathbf{m}^*}} = m_i^* m_j^*$.

> In our example of the Boltzmann machine show that the Gibbs free energy for the factorized ansatz is

$$U_{\mathrm{NMF}}(m) = \langle E(\mathbf{x}) \rangle_{q_{\mathbf{m}}} = -\mathbf{b}^\top \mathbf{m} - \frac{1}{2} \mathbf{m}^\top \mathbf{W} \mathbf{m}\,, \tag{20}$$

$$H_{\mathrm{NMF}}(m) = -\langle \log q(\mathbf{x}) \rangle_{q_{\mathbf{m}}} = -\sum_{i=1}^{N} m_i \log m_i + (1 - m_i) \log(1 - m_i)\,, \tag{21}$$

$$G_{\mathrm{NMF}}(m) = U_{\mathrm{NMF}}(m) - H_{\mathrm{NMF}}(m)/\beta. \tag{22}$$

> Looking for the stationary points of G verify that the m_i^* obey a closed set of non linear equations,

$$\left.\frac{\partial G}{\partial m_i}\right|_{\mathbf{m}^*} = 0 \quad \Rightarrow \quad m_i^* = \sigma(\beta b_i + \sum_{j\in\partial i} \beta W_{ij} m_j^*) \quad \forall i\,, \tag{23}$$

where $\sigma(x) = (1 + e^{-x})^{-1}$. The solutions can be computed by iterating these relations until a fixed point is reached.

2. Application to RBMs

We will now leverage the mean-field approximation for the RBM training.

> Adapt the expression of the mean-field free energy (22) and the fixed point equations of the fully connected Boltzmann machines (23) to the case of the RBM with hidden units.

> Write a function to compute the mean-field approximation of the log-likelihood following the

> Uncomment the lines from the MCMC training function and relaunch the fit to store the evolution of the mean-field approximation of the likelihood during the training.

> By computing the derivatives of the mean-field likelihood, obtain the mean-field approximation of the gradients. Taking inspiration from the MCMC functions implement the RBM training with the mean-field gradients.

D. The Thouless-Anderson-Palmer mean field correction

It is possible to further refine the mean-field approximation by morally adding one second term in the square of the weights W_{ij}. The TAP mean-field equations[4] were originally derived as an exact mean-field theory for the Sherrington-Kirkpatrick (SK) model[3]. This emblematic *spin glass* model corresponds to a fully connected Ising model with energy (12) and disordered couplings W_{ij} drawn independently from a Gaussian distribution with zero mean and variance W_0/N. Here we will not take the time to go trough the more involved derivation, but we directly give for the Boltzmann machine (12), the TAP equations and TAP free energy (truncated at second order) which read[4],

$$m_i = \sigma\left(\beta b_i + \sum_{j\in\partial i} \beta W_{ij} m_j - \beta^2 W_{ij}^2 (m_i - \frac{1}{2})(m_j - {m_j}^2)\right) \ \forall i \tag{24}$$

$$\beta G(\mathbf{m}) = F_{\mathrm{NMF}} - \frac{\beta^2}{2}\sum_{(ij)} W_{ij}^2 (m_i - {m_i}^2)(m_j - {m_j}^2)\,,$$

where naive mean-field entropy was defined in (21).

> Compare the TAP and the mean-field equations.

> Check that you agree with the proposed implementation of the TAP approximation for training[1].

E. Conclusion

> Retrieve the history of the three types of training and compare them.

> Compare also the MCMC samples that the three types of RBM can generate.

References

[1] Marylou Gabrié, Eric W. Tramel, and Florent Krzakala. Training Restricted Boltzmann Machines via the Thouless-Anderson-Palmer Free Energy. *Advances in Neural Information Processing Systems 28*, pages 640—-648, jun 2015.
[2] Geoffrey E. Hinton. Training products of experts by minimizing Contrastive divergence. *Neural computation*, 14:1771–1800, 2002.
[3] David Sherrington and Scott Kirkpatrick. Solvable Model of a Spin-Glass. *Physical Review Letters*, 35(26):1792–1796, dec 1975.
[4] D. J. Thouless, P. W. Anderson, and R. G. Palmer. Solution of 'Solvable model of a spin glass'. *Philosophical Magazine*, 35(3):593–601, 1977.

Note on mean-field

To understand the implication of the restriction to factorized distributions, it is instructive to compare this naive mean-field equation with the exact identity

$$\langle x_i \rangle_p = \langle \sigma(\beta b_i + \sum_{j\in\partial i} \beta W_{ij} x_j) \rangle_p \,. \tag{25}$$

Under the Boltzmann distribution $p(\mathbf{x}) = e^{-\beta E(\mathbf{x})}/\mathcal{Z}$, these averages are difficult to compute. The naive mean-field method is neglecting the fluctuations of the effective field felt by the variable x_i: $\sum_{j\in\partial i} W_{ij} x_j$, keeping only its mean $\sum_{j\in\partial i} W_{ij} m_j$. This incidentally justifies the name of mean-field methods.