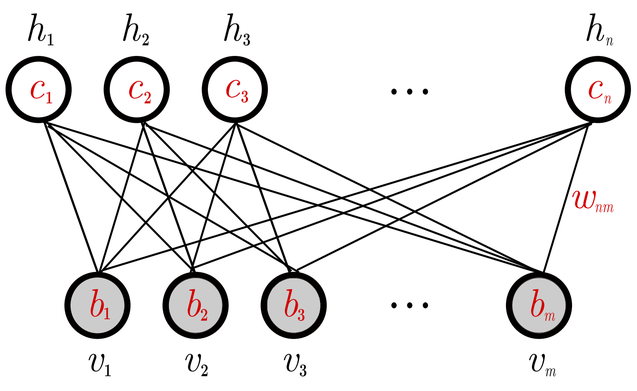
[1]*Institute for Neural Computation – Ruhr-University Bochum*
[2]*Department for Computer Science – University of Copenhagen*

How to center restricted Boltzmann machines

Asja Fischer[1,2], Jan Melchior[1], Nan Wang[1], Laurenz Wiskott[1]

Binary restricted Boltzmann machines

An Restricted Boltzmann Machine (RBM) is an undirected graphical model



modeling the joint probability distribution

$$p(\mathbf{v},\mathbf{h}) = \frac{1}{Z}\mathrm{e}^{-E(\mathbf{v},\mathbf{h})}$$

with normalization constant $Z = \sum_{\mathbf{x}}\sum_{\mathbf{h}}\mathrm{e}^{-E(\mathbf{x},\mathbf{h})}$ and energy

$$E(\mathbf{v},\mathbf{h}) = -\mathbf{v}^T\mathbf{b} - \mathbf{c}^T\mathbf{h} - \mathbf{v}^T\mathbf{W}\mathbf{h}$$

where $\mathbf{W}$ is the matrix collecting all weights w_{ij} and $\mathbf{b}^T = (b_1,\ldots,b_m)$ and $\mathbf{c}^T = (c_1,\ldots,c_n)$ are the hidden and visible bias vector, respectively.

Training corresponds to maximizing the likelihood of the training samples under the model based on gradient ascent, where the log likelihood gradient can be written as

$$\begin{aligned}\nabla\mathbf{W} &= \langle\mathbf{v}\mathbf{h}^T\rangle_d - \langle\mathbf{v}\mathbf{h}^T\rangle_m\\ \nabla\mathbf{b} &= \langle\mathbf{v}\rangle_d - \langle\mathbf{v}\rangle_m\\ \nabla\mathbf{c} &= \langle\mathbf{h}\rangle_d - \langle\mathbf{h}\rangle_m\end{aligned}$$

Common RBM training methods (as Contrastive Divergence (CD), Persistent Contrastive Divergence (PCD) or Parallel Tempering (PT)) approximate the expectation $\langle\cdot\rangle_m$ under the RBM distribution by samples gained from different Markov chain Monte Carlo methods.

Centered restricted Boltzmann machines

Different parametrization of the energy using offset parameters $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$

$$\begin{aligned}E(\mathbf{v},\mathbf{h}) = &-(\mathbf{v}-\boldsymbol{\mu})^T\mathbf{b} - \mathbf{c}^T(\mathbf{h}-\boldsymbol{\lambda})\\ &-(\mathbf{v}-\boldsymbol{\mu})^T\mathbf{W}(\mathbf{h}-\boldsymbol{\lambda})\end{aligned}$$

Offset parameters suggested before

- $\boldsymbol{\mu} = \langle\mathbf{v}\rangle_d$ and $\boldsymbol{\lambda} = \mathbf{0}$ lead to invariance of RBM to flips of the data set [4].
- $\boldsymbol{\mu} = \langle\mathbf{v}\rangle_d$ and $\boldsymbol{\lambda} = \langle\mathbf{h}\rangle_d$ lead to better generative performance of locally connected deep Boltzmann machines [3].

The log likelihood gradient now gets

$$\begin{aligned}\nabla\mathbf{W} &= \langle(\mathbf{v}-\boldsymbol{\mu})(\mathbf{h}-\boldsymbol{\lambda})^T\rangle_d - \langle(\mathbf{v}-\boldsymbol{\mu})(\mathbf{h}-\boldsymbol{\lambda})^T\rangle_m\\ \nabla\mathbf{b} &= \langle\mathbf{v}-\boldsymbol{\mu}\rangle_d - \langle\mathbf{v}-\boldsymbol{\mu}\rangle_m = \langle\mathbf{v}\rangle_d - \langle\mathbf{v}\rangle_m\\ \nabla\mathbf{c} &= \langle\mathbf{h}-\boldsymbol{\lambda}\rangle_d - \langle\mathbf{h}-\boldsymbol{\lambda}\rangle_m = \langle\mathbf{h}\rangle_d - \langle\mathbf{h}\rangle_m\end{aligned}$$

Offsets may change during training (e.g. $\boldsymbol{\lambda} = \langle\mathbf{h}\rangle_d$) and thus may need to be updated. An RBM with offsets $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ can be transformed to an RBM with offsets $\boldsymbol{\mu}'$ and $\boldsymbol{\lambda}'$ by

$$\begin{aligned}\mathbf{W}' &= \mathbf{W}\\ \mathbf{b}' &= \mathbf{b} + \mathbf{W}(\boldsymbol{\lambda}' - \boldsymbol{\lambda})\\ \mathbf{c}' &= \mathbf{c} + \mathbf{W}^T(\boldsymbol{\mu}' - \boldsymbol{\mu})\end{aligned}$$

such that $E(\mathbf{v},\mathbf{h}|\boldsymbol{\theta},\boldsymbol{\mu},\boldsymbol{\lambda}) = E(\mathbf{v},\mathbf{h}|\boldsymbol{\theta}',\boldsymbol{\mu}',\boldsymbol{\lambda}') + const$ is guaranteed.

Objectives

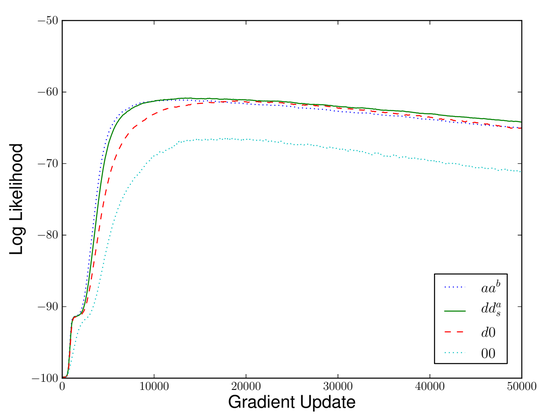
Analyze the properties of centered RBMs and the performance for different choices of the offset parameters.

Results [2]

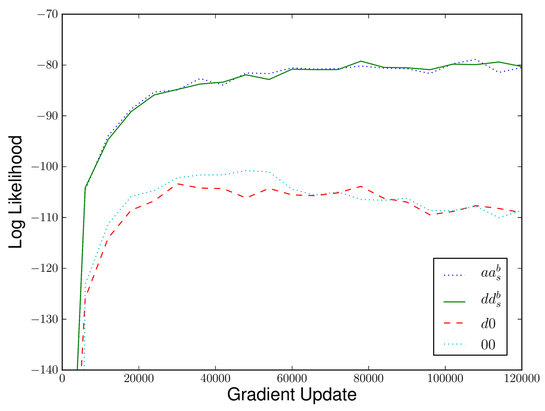
- Centering **can be interpreted as a different update direction** for gradient ascent for normal RBMs. Transforming normal binary to centered RBM, performing gradient update and transforming back yields

$$\begin{aligned}\nabla_c\mathbf{W} &= \langle(\mathbf{v}-\boldsymbol{\mu})(\mathbf{h}-\boldsymbol{\lambda})^T\rangle_d - \langle(\mathbf{v}-\boldsymbol{\mu})(\mathbf{h}-\boldsymbol{\lambda})^T\rangle_m\\ \nabla_c\mathbf{b} &= \langle\mathbf{v}\rangle_d - \langle\mathbf{v}\rangle_m - \nabla_c\mathbf{W}\boldsymbol{\lambda}\\ \nabla_c\mathbf{c} &= \langle\mathbf{h}\rangle_d - \langle\mathbf{h}\rangle_m - \nabla_c\mathbf{W}^T\boldsymbol{\mu}\end{aligned}$$

- By setting $\boldsymbol{\mu} = \frac{1}{2}(\langle\mathbf{v}\rangle_d + \langle\mathbf{v}\rangle_m)$ and $\boldsymbol{\lambda} = \frac{1}{2}(\langle\mathbf{h}\rangle_d + \langle\mathbf{h}\rangle_m)$ this becomes **equal to the 'enhanced gradient'** [1]
- It can be shown analytically and empirically that the performance of a RBM using this update direction is **invariant to flips of the data set** for certain choices of offset parameters (e.g. the expectation of the variables under any distribution).
- Experiments on *Bars-and-Stripes* and *MNIST* data set show: centering of visible and hidden neurons leads to **significantly higher log likelihood** values (aa corresponds to the enhanced gradient, dd to $\boldsymbol{\mu} = \langle\mathbf{v}\rangle_d$ and $\boldsymbol{\lambda} = \langle\mathbf{h}\rangle_d$, $d0$ to $\boldsymbol{\mu} = \langle\mathbf{v}\rangle_d$ and $\boldsymbol{\lambda} = 0$ and 00 to the usual binary RBM).

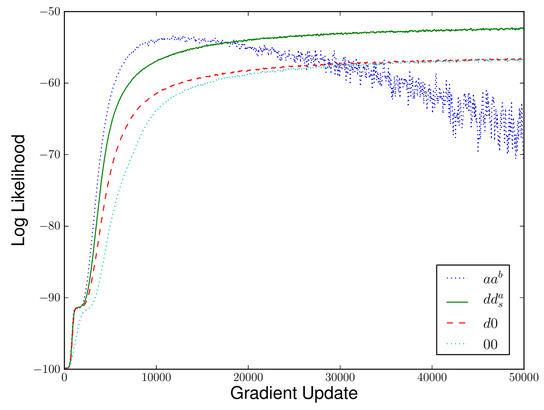


Log likelihood during training with CD-1 on the *Bars-and-Stripes* data set



Estimated log likelihood (based on AIS) during training with PCD-1 on the *MNIST* data set

- The **enhanced gradinet leads to a severe divergence** problem when used together with PT.



Log likelihood during training with PT on the *Bars-and-Stripes* data set

- The divergence **can be prevented by using a sliding average** for the offset estimation.

Conclusion

Centering of visible and hidden variables solves the invariance problem and leads to better generative models with low additional computational cost. So it should always be used!

References

[1] K. Cho, T. Raiko, and A. Ilin. Enhanced gradient and adaptive learning rate for training restricted boltzmann machines. In L. Getoor and T. Scheffer, editors, *International Conference on Machine Learning (ICML)*, pages 105–112. ACM, 2011.

[2] J. Melchior, A. Fischer, N. Wang, and L. Wiskott. How to center binary restricted boltzmann machines. arXiv.org e-Print archive, 2013.

[3] G. Montavon and K. Müller. Deep boltzmann machines and the centering trick. *Lecture Notes in Computer Science (LNCS)*, 7700:621–637, 2012.

[4] Y. Tang and I. Sutskever. Data normalization in the learning of restricted Boltzmann machines. Technical report, Department of Computer Science, University of Toronto, 2011.