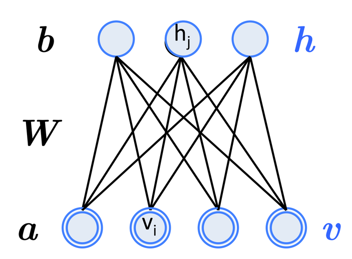
Training Restricted Boltzmann Machines via the Thouless-Anderson-Palmer Free Energy

Marylou Gabrié[LPS-ENS], Eric W. Tramel[LPS-ENS], Florent Krzakala[LPS-ENS]

Training RBMs

Restricted Boltzmann Machines : bipartite energy based graphical model with *visible* neurons representing data and *hidden* latent neurons. Typically,



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

implying joint pdf

$$p(\mathbf{v},\mathbf{h};\mathbf{a},\mathbf{b},\mathbf{W}) = e^{-E(\mathbf{v},\mathbf{h})}/Z\,.$$

Unsupervised learning : maximization of the log-likelihood,

$$\ell(\mathbf{a},\mathbf{b},\mathbf{W}) = \ln \sum_{\mathbf{h}} p = -F^c(\mathbf{v}) + F$$

interpreted as difference between full model free energy F and data-clamped free energy F^c.

$$F = -\ln Z \quad ; \quad F^c(\mathbf{v}) = \mathbf{a}^T\,\mathbf{v} - \sum_{j=1}^{H} \ln\left(1 + e^{-(b_j + (\mathbf{v}^T\,\mathbf{W})_j)}\right)\,.$$

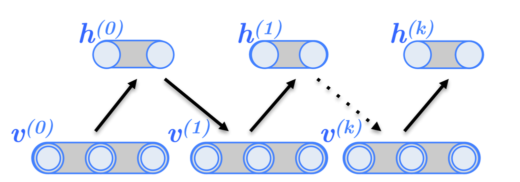
Iterative parameter updates in the direction of likelihood gradients, for instance

$$\frac{\partial \ell}{\partial W_{ij}} = \mathbb{E}[v_i h_j|\,\mathbf{v}] - \mathbb{E}[v_i h_j] = -\frac{\partial F^c(\mathbf{v})}{\partial W_{ij}} + \frac{\partial F}{\partial W_{ij}}\,,$$

Yet, exact computation of full model expectation is intractable.

Gradients Evaluation Techniques

Contrastive divergence : Few steps of Gibbs sampling proved satisfying,



$$p(h_j^{(k)}|\,\mathbf{v}^{(k)}) = \sigma\left(b_j + (\mathbf{v}^{(k)T}\,\mathbf{W})_j\right)$$

$$\mathbb{E}[v_i h_j] = v_i^{(k)} h_j^{(k)} \qquad \text{(CD-k)}$$

Mean-field : Replacing stochastic binary variables by determintic real valued units

$$p(m_j^{h(k)}|\,\mathbf{m}^{\mathrm{v}(k)}) = \sigma\left(b_j + (\mathbf{m}^{\mathrm{v}(k)T}\,\mathbf{W})_j\right)$$

$$\mathbb{E}[v_i h_j] = m_i^{v(k)} m_j^{h(k)} \qquad \text{(MF-k)}$$

References

[1] A. Georges and J. S. Yedidia. How to expand around mean-field theory using high-temperature expansions. *Journal of Physics A: Mathematical and General*, 24(9):2173–2192, 1999.

[2] T. Plefka. Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model. *Journal of Physics A: Mathematical and General*, 15(6):1971–1978, 1982.

[3] 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.

Links

Paper **Source Code** **SPHINX Team**

An Expansion for the Ising Model

The Ising Model: Set of binary spins interacting according to the Hamiltonian $H(\mathbf{s}) = -\,\mathbf{a}^T\,\mathbf{s} - \mathbf{s}^T\,\mathbf{W}\,\mathbf{s}$. Probability of a configuration $\mathbf{s}$ at inverse temperature $\beta = 1/k_BT$ is

$$p(\mathbf{s}) = e^{-\beta H(\mathbf{s})}/Z\,,$$

and the associated free enerygy is $-\beta F = \ln Z$.

Legendre transforms: Using a newly introduced auxiliary external field $\mathbf{q}$, we define $-\beta\tilde{F}[\mathbf{q}] = \ln\sum_{\mathbf{s}} \mathrm{e}^{-\beta E(\mathbf{s})+\beta\sum_i q_i s_i}$, and compute its Legendre transform as a function of $\mathbf{m} = -\frac{dF}{d\mathbf{q}}$

$$-\beta\Gamma[\mathbf{m}] = -\beta\max_{\mathbf{q}}[\tilde{F}[\mathbf{q}] + \sum_i q_i m_i]\,.$$

Inverse transform finally yields an expression of F in terms $\mathbf{m}$.

$$-\beta F = -\beta\tilde{F}[\mathbf{q}=0] = -\beta\min_{\mathbf{m}}[\Gamma[\mathbf{m}]]\,.$$

High temperature expansion: One can expand $-\beta\Gamma[\mathbf{m}]$ around $\beta = 0$ at fixed $\mathbf{m}$ [1, 2, 3] which yields

$$-\beta F^{\mathrm{EMF}} = -\left[\mathbf{m}^T\ln\mathbf{m} + (1-\mathbf{m})^T\ln(1-\mathbf{m})\right] + \beta(\mathbf{a}^T\,\mathbf{m} + \mathbf{m}^T\,\mathbf{W}\,\mathbf{m}) + \frac{\beta^2}{2}\sum_{(i,j)} W_{ij}^2\,v_i\,v_j + \cdots$$

where $v_i = m_i - m_i^2$ and $\mathbf{m}$ is given by order-dependent self consistency relations

$$m_i = \sigma\left[a_i + \sum_j W_{ij} m_j - W_{ij}^2\left(m_i - \frac{1}{2}\right)v_j + \cdots\right].$$

Extended mean field for RBMs

Evaluation of likelihood : Given a set of RBM parameters $\mathbf{a}$, $\mathbf{b}$, $\mathbf{W}$, self consistency relations are iterated to get magnetizations

$$m_j^h[t+1] \leftarrow \sigma\left[b_j + \sum_i W_{ij} m_i^v[t] \quad - \sum_j W_{ij}^2\left(m_j^h[t] - \frac{1}{2}\right)v_i^v[t] \quad + \sum_j W_{ij}^3\left(\frac{1}{3} - 2\,v_j^h[t]\right)m_i^v[t]\,v_i^v[t]\right],$$

$$m_i^v[t+1] \leftarrow \sigma\left[a_i + \sum_j W_{ij} m_j^h[t+1] \quad - \sum_j W_{ij}^2\left(m_i^v[t] - \frac{1}{2}\right)v_j^h[t+1] \quad + \sum_j W_{ij}^3\left(\frac{1}{3} - 2\,v_i^v[t]\right)m_j^h[t+1]\,v_j^h[t+1]\right].$$

From which the EMF approximate free energy yields a straigthforward log-likelihood estimate $\ell(\mathbf{a},\mathbf{b},\mathbf{W}) = -F^c(\mathbf{v}) + F^{\mathrm{EMF}}$

$$F^{\mathrm{EMF}} = -\,S(\mathbf{m}^v,\mathbf{m}^h) - \mathbf{a}^T\,\mathbf{m}^{\mathrm{v}} - \mathbf{b}^T\,\mathbf{m}^{\mathrm{h}} - \mathbf{m}^{\mathrm{v}T}\,\mathbf{W}\,\mathbf{m}^{\mathrm{h}} \quad + \sum_{i,j}\frac{W_{ij}^2}{2}\,v_i\,v_j \quad - \frac{2}{3}W_{ij}^3\,v_i\,v_j\left(\frac{1}{2} - m_i^v\right)\left(\frac{1}{2} - m_j^h\right).$$

Learning : Gradients are computed using $\mathbf{m}^{\mathrm{h}}$, $\mathbf{m}^{\mathrm{v}}$ and F^{EMF} : $\frac{\partial F^{\mathrm{EMF}}}{\partial W_{ij}} = -m_i^v m_j^h + W_{ij}\,v_i^v\,v_j^h - 2W_{ij}^2\,v_i^v\,v_j^h\left(\frac{1}{2} - m_i^v\right)\left(\frac{1}{2} - m_j^h\right)$.

Experimental framework

Algorithm : The resulting training algorithm is similar to the contrastive divergence. Sampling steps are replaced with fixed points iterations.

Algorithm 1 EMF TRAINING

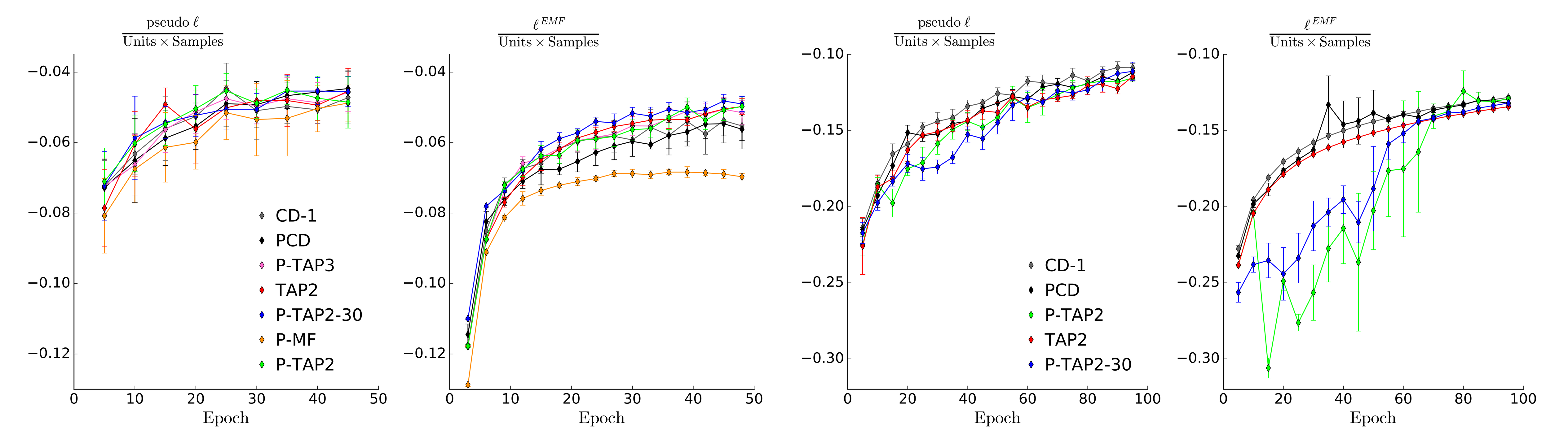
```
Input: {v^(k)}, lr, numepochs, order, numiter,
Initialize {W, a, b}, {m^v, m^h}
for epoch = 1 to numepochs do
  for k = 1 to numcases do
    for t = 1 to numiter do
      m^h[t+1] ← update_mh^(order)(m^v[t], m^h[t])
      m^v[t+1] ← update_mv^(order)(m^v[t], m^h[t+1])
    end for
    Δa = lr (−∇_a F^c(v^(k)) + ∇_a F^EMF(m^v, m^h))
    Δb = lr (−∇_b F^c(v^(k)) + ∇_b F^EMF(m^v, m^h))
    ΔW = lr (−∇_W F^c(v^(k)) + ∇_W F^EMF(m^v, m^h))
    a ← a + Δa
    b ← b + Δb
    W ← W + ΔW
  end for
end for
```

Parameters of interest : Experiments test training quality according to

- EMF *order*
- number of $\mathbf{m}^{\mathrm{h}}$, $\mathbf{m}^{\mathrm{v}}$ iterations
- persistency of iterations

Result 1

Estimates of the per-sample log-likelihood over MNIST test set (left) and over Caltech 101 Silhouette test set (right), normalized by the total number of units, as a function of the number of training epochs.



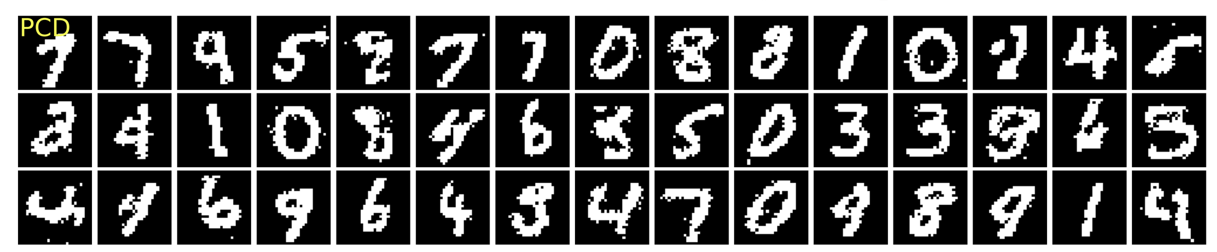
Funding

This work has been supported in part by the ERC under the European Union's 7th Framework Programme Grant Agreement 307087-SPARCS, by the Grant DySpaN of "Triangle de la Physique," and by FAPESP under grant 13/01213-8.

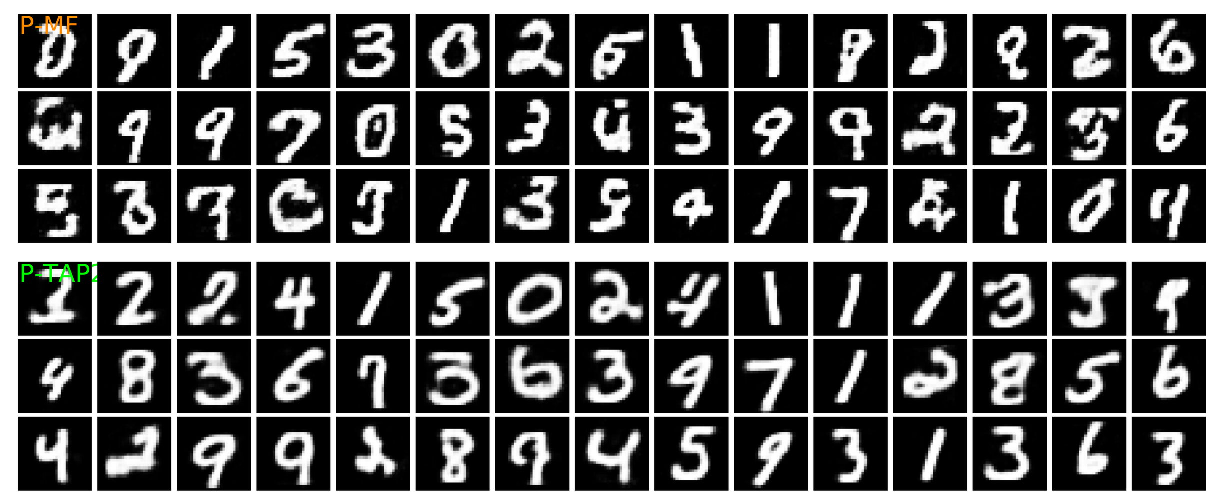
Result 2

Fantasy particles generated by a 500 hidden unit RBM after 50 epochs of training on the MNIST dataset

For PCD chains are binary samples.



For EMF methods, chains are real-valued magnetizations.



Result 3

Test set classification accuracy using logistic regression on the hidden-layer marginal probabilities. As a baseline comparison, the classification accuracy of logistic regression performed directly on the data is given as a black dashed line.