Contrastive Divergence

Giannopoulou Ourania

10 July, 2018

Outline

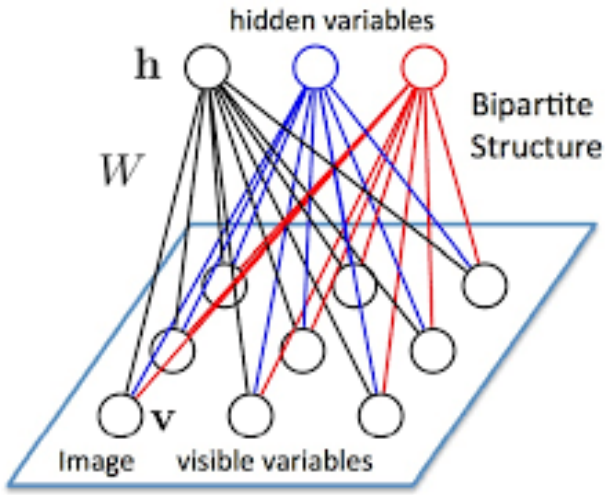
Introduction

- Contrastive divergence is a method to train/learn Restricted Boltzmann Machines

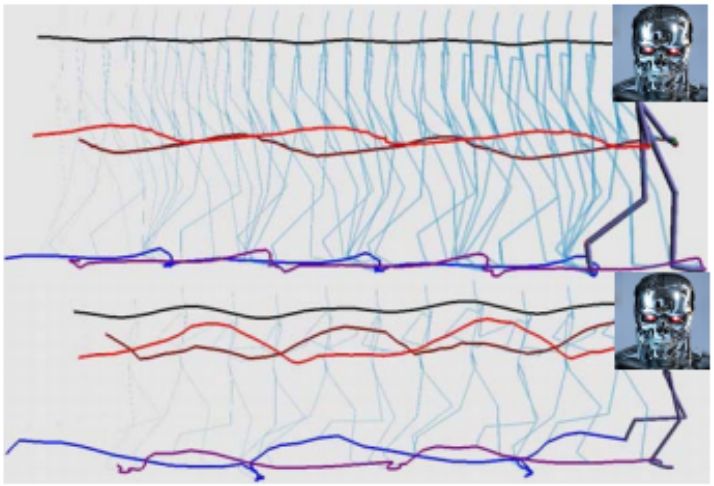
Introduction

- Contrastive divergence is a method to train/learn Restricted Boltzmann Machines

- An RBM is a parametrized model representing a probability distribution
- Learning an RBM means adjusting its parameters such that the its probability distr. fits the training data
- After successful learning they can be used to generate data



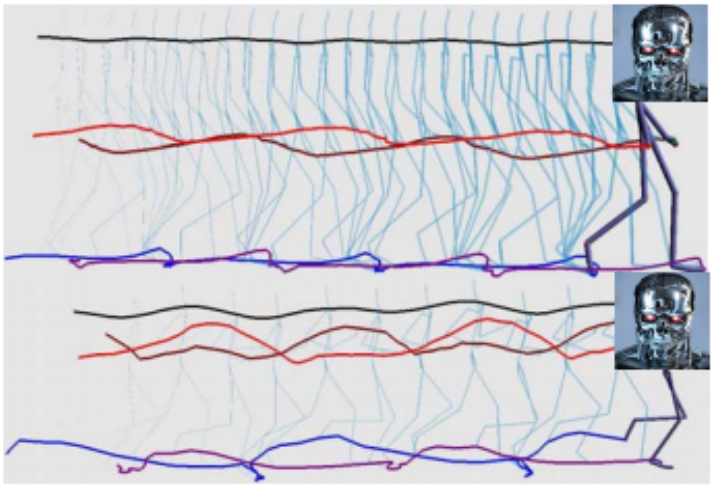
Restricted Boltzmann Machines - Applications



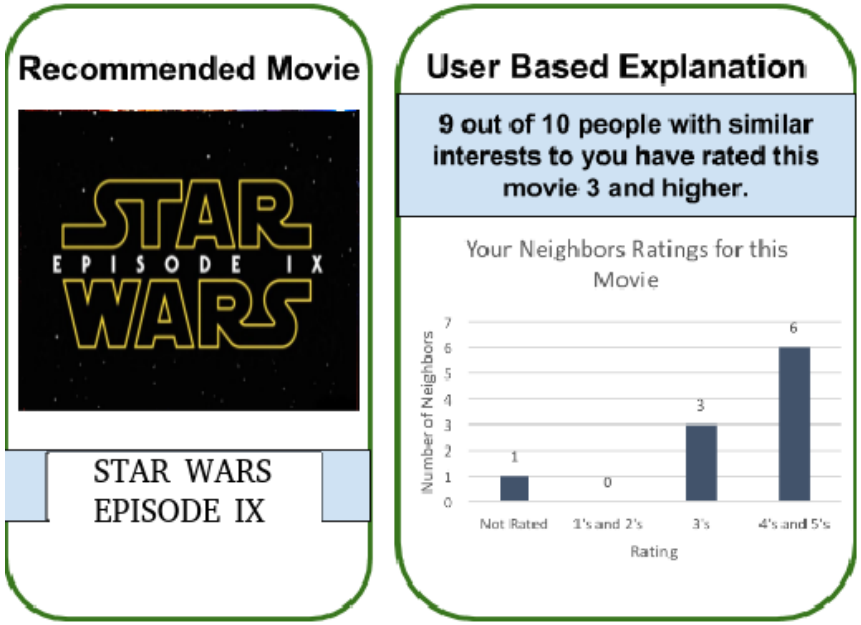
G. W. Taylor et al. (2007), Modeling Human Motion Using Binary Latent Variables

- learning movement patterns

Restricted Boltzmann Machines - Applications

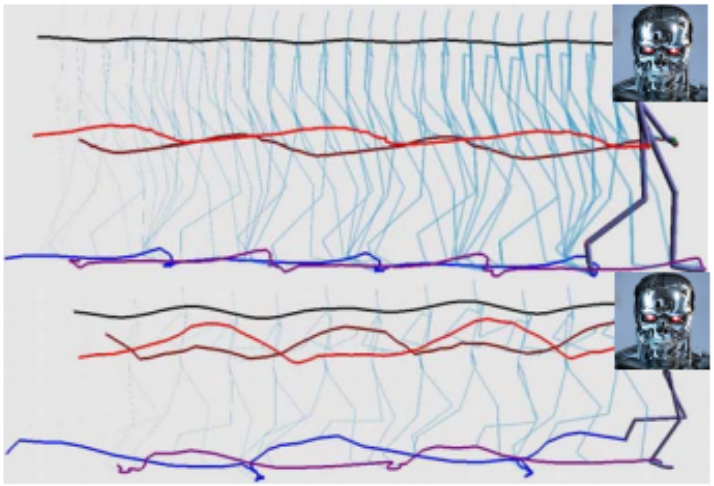


G. W. Taylor et al. (2007), Modeling Human Motion Using Binary Latent Variables

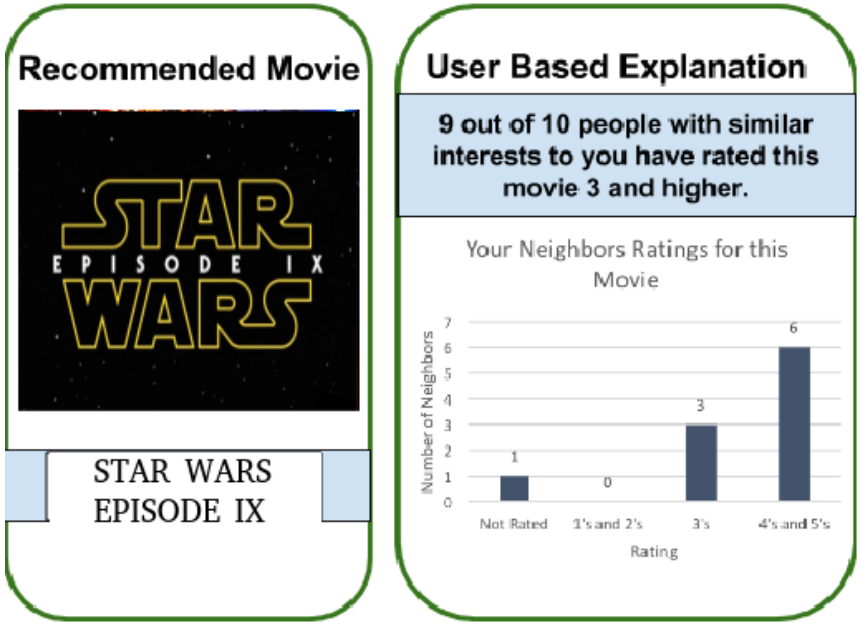


- learning movement patterns
- recommender systems (e.g. movies)

Restricted Boltzmann Machines - Applications



G. W. Taylor et al. (2007), Modeling Human Motion Using Binary Latent Variables



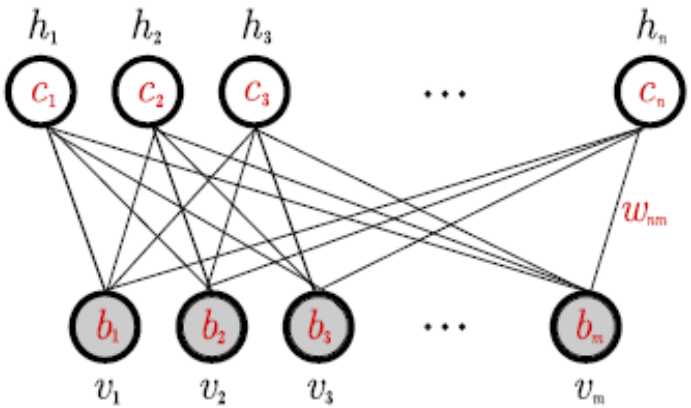
- learning movement patterns
- recommender systems (e.g. movies)
  - Image classification, processing & generation
  - acoustic modelling

Restricted Boltzmann Machines

Invented under the name **Harmonium** by P. Smolensky in 1986

$$p(\mathbf{v},\mathbf{h}) = \frac{1}{Z} e^{-E(\mathbf{v},\mathbf{h})}, \quad (\mathbf{v},\mathbf{h}) \in \{0,1\}^{m+n} \tag{1}$$

$$E(\mathbf{v},\mathbf{h}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij} h_i v_j - \sum_{j=1}^{m} b_j u_j - \sum_{i=1}^{n} c_i h_i, \tag{2}$$



No intralayer connections:

$$p(\mathbf{h}|\mathbf{v}) = \prod_{i=1}^{n} p(h_i|\mathbf{v}), \tag{3}$$

$$p(\mathbf{v}|\mathbf{h}) = \prod_{j=1}^{m} p(v_i|\mathbf{h}), \tag{4}$$

$$p(H_i = 1|\mathbf{v}) = sig\left(\sum_{j=1}^{m} w_{ij} v_j + c_i\right).$$

Restricted Boltzmann Machines - Training

Maximizing the loglikelihood

The gradient of loglikelihood given a single training example $\mathbf{v}$ is:

$$\frac{\partial lnL(\boldsymbol{\theta}|\mathbf{v})}{\partial \boldsymbol{\theta}} = -\sum_{\mathbf{h}} p(\mathbf{h}|\mathbf{v}) \frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \boldsymbol{\theta}} + \sum_{\mathbf{v},\mathbf{h}} p(\mathbf{v},\mathbf{h}) \frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \boldsymbol{\theta}}. \tag{5}$$

Let's take for example the weights w_{ij}:

$$\frac{\partial lnL(\boldsymbol{\theta}|\mathbf{v})}{\partial w_{ij}} = p(H_i = 1|\mathbf{v}) v_j - \sum_{\mathbf{v}} p(\mathbf{v}) p(H_i = 1|\mathbf{v}) v_j, \tag{6}$$

$$p(\mathbf{v}) = \frac{1}{Z} \prod_{j=1}^{m} e^{b_j v_j} \prod_{i=1}^{n} \left(1 + e^{c_i + \sum_{j=1}^{m} w_{ij} v_j}\right) \tag{7}$$

For a whole training set $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_l\}$

$$\frac{1}{l} \sum_{\mathbf{v} \in S} \frac{\partial lnL(\boldsymbol{\theta}|\mathbf{v})}{\partial w_{ij}} = \frac{1}{l} \sum_{\mathbf{v} \in S} \left(-\mathbb{E}_{p(\mathbf{h}|\mathbf{v})} \left(\frac{\partial E(\mathbf{v},\mathbf{h})}{\partial w_{i,j}} \right) + \mathbb{E}_{p(\mathbf{h},\mathbf{v})} \left(\frac{\partial E(\mathbf{v},\mathbf{h})}{\partial w_{ij}} \right) \right)$$

Contrastive Divergence

How to Avoid exponential complexity of 2^m ?

Contrastive Divergence

How to Avoid exponential complexity of 2^m ?

$$\Downarrow$$

Approximate the expectation by samples from the model distribution $p(\mathbf{v})$ (Gibbs sampling)

Contrastive Divergence

How to Avoid exponential complexity of 2^m ?

$$\Downarrow$$

Approximate the expectation by samples from the model distribution $p(\mathbf{v})$ (Gibbs sampling)

$$\Downarrow$$

Run the Markov Chain for infinite time to ensure stationarity : (

Contrastive Divergence

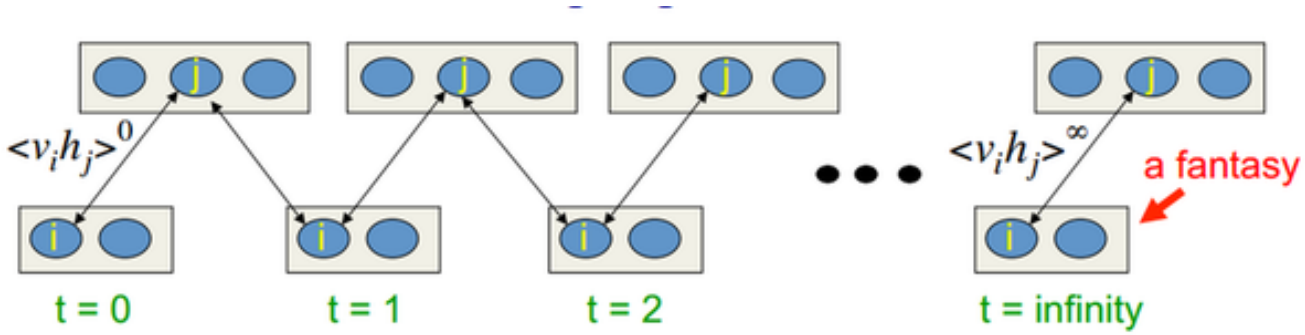
How to Avoid exponential complexity of 2^m ?

$\Downarrow$

Approximate the expectation by samples from the model distribution $p(\mathbf{v})$ (Gibbs sampling)

$\Downarrow$

Run the Markov Chain for infinite time to ensure stationarity : (
IDEA OF CD-k: Instead of sampling from the RBM distribution, run a Gibbs chain for only k steps



Contrastive Divergence

IDEA OF CD-k: Instead of sampling from the RBM distribution, run a Gibbs chain for only k steps

- Initialise the Gibbs chain with a training example $\mathbf{v}^{(0)}$
- at each step, sample $\mathbf{h}^{(t)}$ from $p(\mathbf{h}|\mathbf{v}^{(t)})$ and subsequently $\mathbf{v}^{(t+1)}$ from $p(\mathbf{v}|\mathbf{h}^{(t)})$
- this yields the sample $\mathbf{v}^{(k)}$ after k-steps

Hinton (2002) "Training Products of Experts by Minimizing Contrastive Divergence"

Contrastive Divergence

IDEA OF CD-k: Instead of sampling from the RBM distribution, run a Gibbs chain for only k steps

- Initialise the Gibbs chain with a training example $\mathbf{v}^{(0)}$
- at each step, sample $\mathbf{h}^{(t)}$ from $p(\mathbf{h}|\mathbf{v}^{(t)})$ and subsequently $\mathbf{v}^{(t+1)}$ from $p(\mathbf{v}|\mathbf{h}^{(t)})$
- this yields the sample $\mathbf{v}^{(k)}$ after k-steps

$$\frac{\partial lnL(\boldsymbol{\theta}|\mathbf{v})}{\partial \boldsymbol{\theta}} = -\sum_{\mathbf{h}} p(\mathbf{h}|\mathbf{v}) \frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \boldsymbol{\theta}} + \sum_{\mathbf{v},\mathbf{h}} p(\mathbf{v},\mathbf{h}) \frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \boldsymbol{\theta}}.$$

$$CD_k(\boldsymbol{\theta}, \mathbf{v}^{(0)}) = -\sum_{h} p(\mathbf{h}|\mathbf{v}^{(0)}) \frac{\partial E(\mathbf{v}^{(0)},\mathbf{h})}{\partial \boldsymbol{\theta}} + \sum_{\mathbf{h}} p(\mathbf{h}|\mathbf{v}^{(k)}) \frac{\partial E(\mathbf{v}^{(k)},\mathbf{h})}{\partial \boldsymbol{\theta}}. \quad (8)$$

Hinton (2002) "Training Products of Experts by Minimizing Contrastive Divergence"

Contrastive Divergence

IDEA OF CD-k: Instead of sampling from the RBM distribution, run a Gibbs chain for only k steps

- Initialise the Gibbs chain with a training example $\mathbf{v}^{(0)}$
- at each step, sample $\mathbf{h}^{(t)}$ from $p(\mathbf{h}|\mathbf{v}^{(t)})$ and subsequently $\mathbf{v}^{(t+1)}$ from $p(\mathbf{v}|\mathbf{h}^{(t)})$
- this yields the sample $\mathbf{v}^{(k)}$ after k-steps

$$\frac{\partial lnL(\boldsymbol{\theta}|\mathbf{v})}{\partial \boldsymbol{\theta}} = -\sum_{\mathbf{h}} p(\mathbf{h}|\mathbf{v}) \frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \boldsymbol{\theta}} + \sum_{\mathbf{v},\mathbf{h}} p(\mathbf{v},\mathbf{h}) \frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \boldsymbol{\theta}}.$$

$$CD_k(\boldsymbol{\theta}, \mathbf{v}^{(0)}) = -\sum_{h} p(\mathbf{h}|\mathbf{v}^{(0)}) \frac{\partial E(\mathbf{v}^{(0)},\mathbf{h})}{\partial \boldsymbol{\theta}} + \sum_{\mathbf{h}} p(\mathbf{h}|\mathbf{v}^{(k)}) \frac{\partial E(\mathbf{v}^{(k)},\mathbf{h})}{\partial \boldsymbol{\theta}}. \quad (8)$$

⚠ The approximation is biased

Hinton (2002) "Training Products of Experts by Minimizing Contrastive Divergence"

Some Convergence results

For a converging Gibbs chain

$$\mathbf{v}^{(0)} \Rightarrow \mathbf{h}^{(0)} \Rightarrow \mathbf{v}^{(1)} \Rightarrow \mathbf{h}^{(1)} \ldots, \tag{9}$$

starting from the data point $\mathbf{v}^{(0)}$, the log likelihood gradient can be written as a sum of terms containing the k-th sample

$$\frac{\partial}{\partial \boldsymbol{\theta}} ln p(\mathbf{v}^{(0)}) = -\sum_{\mathbf{h}} p(\mathbf{h}|\mathbf{v}^{(0)}) \frac{\partial E(\mathbf{v}^{(0)}, \mathbf{h})}{\partial \boldsymbol{\theta}} + E_{p(\mathbf{v}^{(k)}|\mathbf{v}^{(0)})} \left(\sum_{\mathbf{h}} p(\mathbf{h}|\mathbf{v}^{(k)}) \frac{\partial E(\mathbf{v}^{(k)}, \mathbf{h})}{\partial \boldsymbol{\theta}} \right) + E_{p(\mathbf{v}^{(k)}|\mathbf{v}^{(0)})} \left(\frac{\partial ln p(\mathbf{v}^{(k)})}{\partial \boldsymbol{\theta}} \right), \tag{10}$$

where the bias converges to zero as $k \to \infty$.
Upper bound for the bias term $= 2^m (1 - 2^m 2^n sig(-\alpha)^m sig(-\beta)^n)^k$

Bengio & Dellaleau (2009) "Justifying and Generalizing Contrastive Divergence"

Some Convergence results

The upper bound on the expectation of the error for the CD-k approximation of the loglikelihood derivative with respect to some RBM parameter θ is

$$\left| E_{q(\mathbf{v}^{(0)})} \left(E_{p(\mathbf{v}^{(k)}|\mathbf{v}^{(0)})} \left(\frac{\partial ln p(\mathbf{v}^{(k)})}{\partial \theta} \right) \right) \right| \leq \frac{1}{2} |q - p| \left(1 - e^{-(m+n)\Delta} \right)^{k}, \qquad (11)$$

where p is the marginal distribution of the visible units , q is the empirical distribution defined by a set of samples $\mathbf{v}_1, \ldots, \mathbf{v}_l$,

$$\Delta = \max \left(\max_{l \in \{1,\ldots,m\}} \theta_l, \max_{l \in \{1,\ldots,n\}} \xi_l \right),$$

$$\theta_l = \max \left(\left| \sum_{i=1}^{n} I_{\{w_{il}>0\}} w_{il} + b_l \right|, \left| \sum_{i=1}^{n} I_{\{w_{il}<0\}} w_{il} + b_l \right| \right),$$

$$\xi_l = \max \left(\left| \sum_{j=1}^{m} I_{\{w_{lj}>0\}} w_{lj} + c_l \right|, \left| \sum_{j=1}^{m} I_{\{w_{lj}<0\}} w_{lj} + c_l \right| \right),$$

Fisher & Igel (2011) "Bounding the Bias of Contrastive Divergence Learning"

Drawbacks

- Due to the approximation error CD learning does not necessarily lead to a maximum likelihood estimate of the model parameters. Specific conditions under which this happens are given by Yuille (2005)
- The bias can lead to distortion of the learning process, increases with the magnitude of the RBM parameters

Alternatives to CD - Weight decay

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \eta \frac{\partial}{\partial \boldsymbol{\theta}^{(t)}} \left(ln\ L(\boldsymbol{\theta}^{(t)}|S) \right) - \lambda \boldsymbol{\theta}^{(t)} \tag{12}$$

where $S = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l\}$ training data set, $\eta \in \mathbb{R}^+$ learning rate , $\lambda \in \mathbb{R}^+_{\nu}$ weight decay parameter.
Penalizes weights with large magnitudes
Optimize the objective function :

$$\theta^* = \arg\min_{\theta} \left(ln\ L(\theta|S) \right) - \frac{1}{2}||\theta||^2. \tag{13}$$

✓ Smaller weights $\rightarrow$ faster mixing rate $\rightarrow$ less biased approximation
✗ Parameter λ is difficult to tune

Alternatives to CD - Persistent CD / Fast Persistent CD

Persistent CD

"Persistent" chains which are run for k Gibbs steps after each parameter update (the initial state of the current Gibbs chain is equal to $v^{(k)}$ from the previous update).

Fast Persistent CD

Introduce fast parameters w_{ij}^f, b_j^f, c_i^f. Then Gibbs sampling is based on

$$\tilde{p}(H_i = 1|\mathbf{v}) = sig\left(\sum_{j=1}^{m}(w_{ij} + w_{ij}^f)v_j + (c_i + c_i^f)\right) \tag{14}$$

$$\tilde{p}(V_i = 1|\mathbf{h}) = sig\left(\sum_{i=1}^{n}(w_{ij} + w_{ij}^f)h_i + (b_j + b_j^f)\right), \tag{15}$$

- update the regular parameters with a small learning rate to keep close to the model distr.
- update the fast parameters with a fast learning rate to achieve faster mixing and with large weight decay parameter to keep them close to the regular

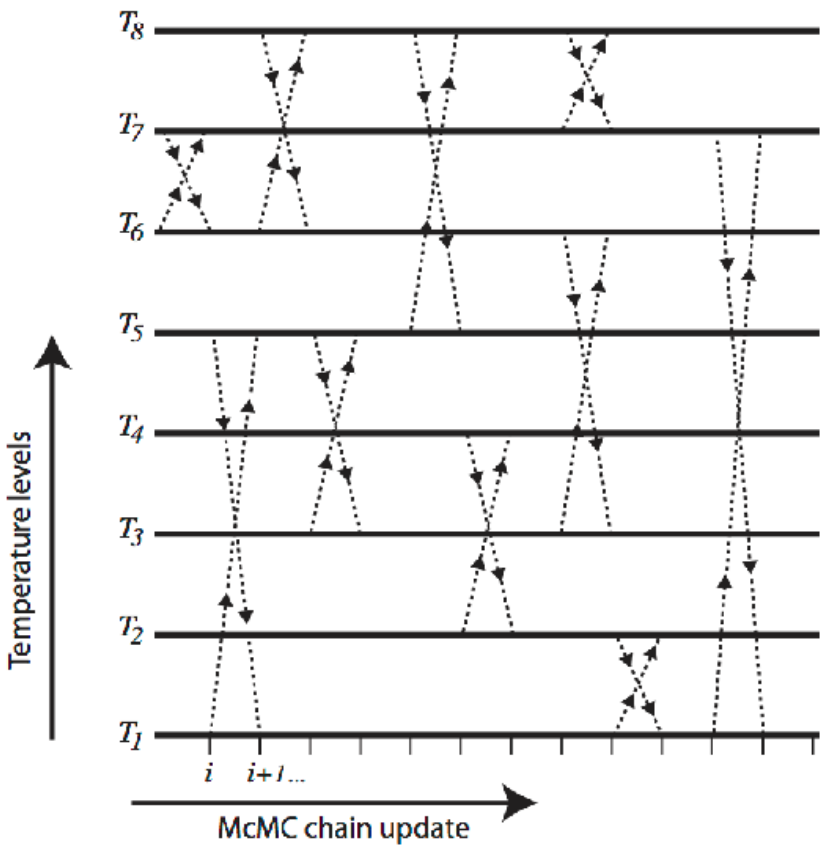
Alternatives to CD - Parallel Tempering

M temperatures $1 = T_1 < T_2 < \cdots < T_M$ & a set of Markov chains with stationary distributions:

$$p_r(\mathbf{v}, \mathbf{h}) = \frac{1}{Z_r} e^{-\frac{1}{T_r} E(\mathbf{v}, \mathbf{h})}$$

p_1 is the model distribution

- run several replicas in parallel for k Gibbs sampling steps
- exchange particles based on the ratio $min\left(1, e^{\left(\frac{1}{T_r} - \frac{1}{T_{r-1}}\right)(E_r - E_{r-1}}\right)$
- take the sample $\mathbf{v}_1$ as a sample from the RBM distribution
- Repeat L times to get samples $\mathbf{v}_{1,1}, \ldots, \mathbf{v}_{1,L}$



✓ PT : faster mixing Markov chain / less biased approximation.

Conclusions

- Problem: The CD approximation produces a bias approximation of the loglikelihood
- Solution: Dynamically control k number of steps and use weight decay to control the magnitudes of the parameters.
- Another solution: use PCD to keep the Markov chains close to equilibrium distribution. In the case where it is practically inefficient due to for example large data dimensionality the use of FPCD offers fast mixing with low computational overhead.
- The State-of-the-art solution: PT has the fastest mixing rate possible.
- The curse of exponential complexity: Also for PT, the bound of the convergence rate is still exponentially dependent on the size of the smallest layer and the absolute values of the RBM parameters, (Fischer & Igel (2015)).

[1] Y. Bengio and O. Delalleau. Justifying and generalizing contrastive divergence. *Neural Computation*, 21(6):1601–1621, 2009. PMID: 19018704.

[2] A. Fischer and C. Igel. Bounding the bias of contrastive divergence learning, 12 2011.

[3] A. Fischer and C. Igel. Training restricted boltzmann machines: An introduction. *Pattern Recognition*, 47(1):25 – 39, 2014.

[4] A. Fischer and C. Igel. A bound for the convergence rate of parallel tempering for sampling restricted boltzmann machines, 05 2015.

[5] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[6] A. L. Yuille. The convergence of contrastive divergences. In *Advances in neural information processing systems*, pages 1593–1600, 2005.

Thank you!