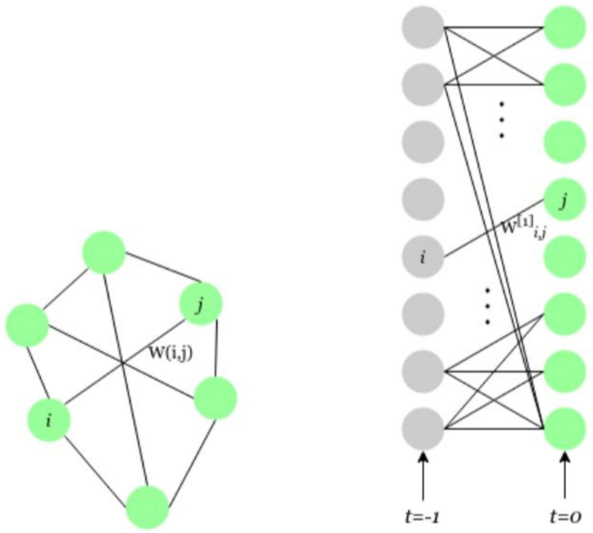
Dynamic Boltzmann Machine

Rajat Patel and Suyash Bagad

Indian Institute of Technology, Bombay

Boltzmann Machine



(a) Boltzmann machine (b) Restricted Boltzmann machine

$$P_\theta(\mathbf{x}) = \frac{e^{-\tau E_\theta(\mathbf{x})}}{\sum_{\tilde{\mathbf{x}}} e^{-\tau E_\theta(\tilde{\mathbf{x}})}}$$

$$E_\theta(\mathbf{x}) = -\mathbf{b}^T\mathbf{x} - \frac{1}{2}\mathbf{x}^T\mathbf{W}\mathbf{x}$$

$$W_{i,j} \leftarrow W_{i,j} + \eta(x_i x_j - \langle X_i, X_j \rangle_\theta)$$

Restricted Boltzmann Machine

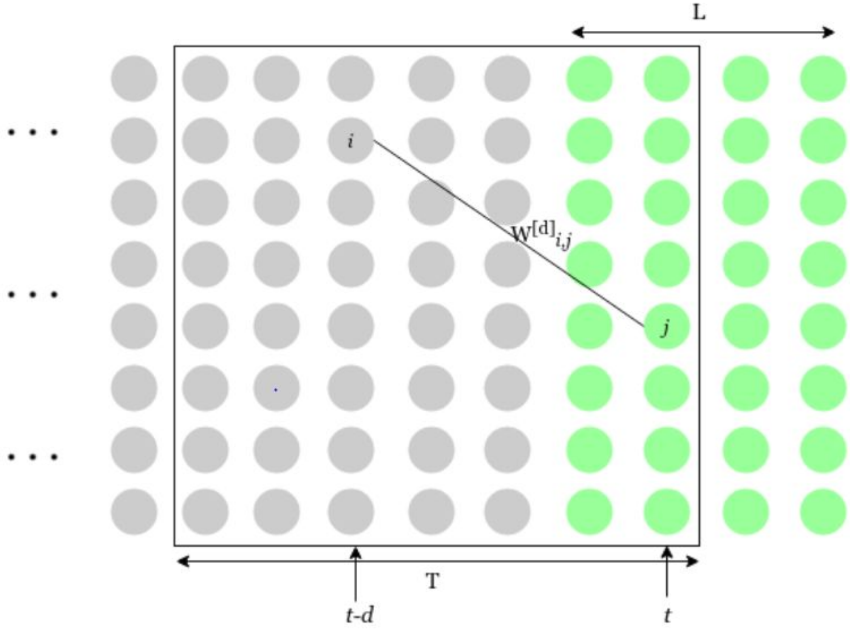
$$P_\theta(\mathbf{x}^{[2]}|\mathbf{x}^{[1]}) = \prod_{j\in[1,N]} P_{\theta,j}(x_j^{[2]}|\mathbf{x}^{[1]})$$

$$= \prod_{j\in[1,N]} \frac{e^{-\tau E_\theta(x_j^{[2]}|\mathbf{x}^{[1]})}}{\sum_{x_j^{[2]}\in\{0,1\}} e^{-\tau E_\theta(x_j^{[2]}|\mathbf{x}^{[1]})}}$$

$$E_\theta(x_j^{[2]}|\mathbf{x}^{[1]}) = -b_j x_j^{[2]} - (\mathbf{x}^{[1]})^T W_{:,j}^{[1]} x_j^{[2]}$$

Train using Contrastive Divergence to get the expectation term. Time intensive!!

Time series Boltzmann Machine



$$p(\mathbf{x}) = \prod_{t\in[-L+1,-1]} P_\theta(\mathbf{x}^{[t]}|\mathbf{x}^{(t-T,t-1]})$$

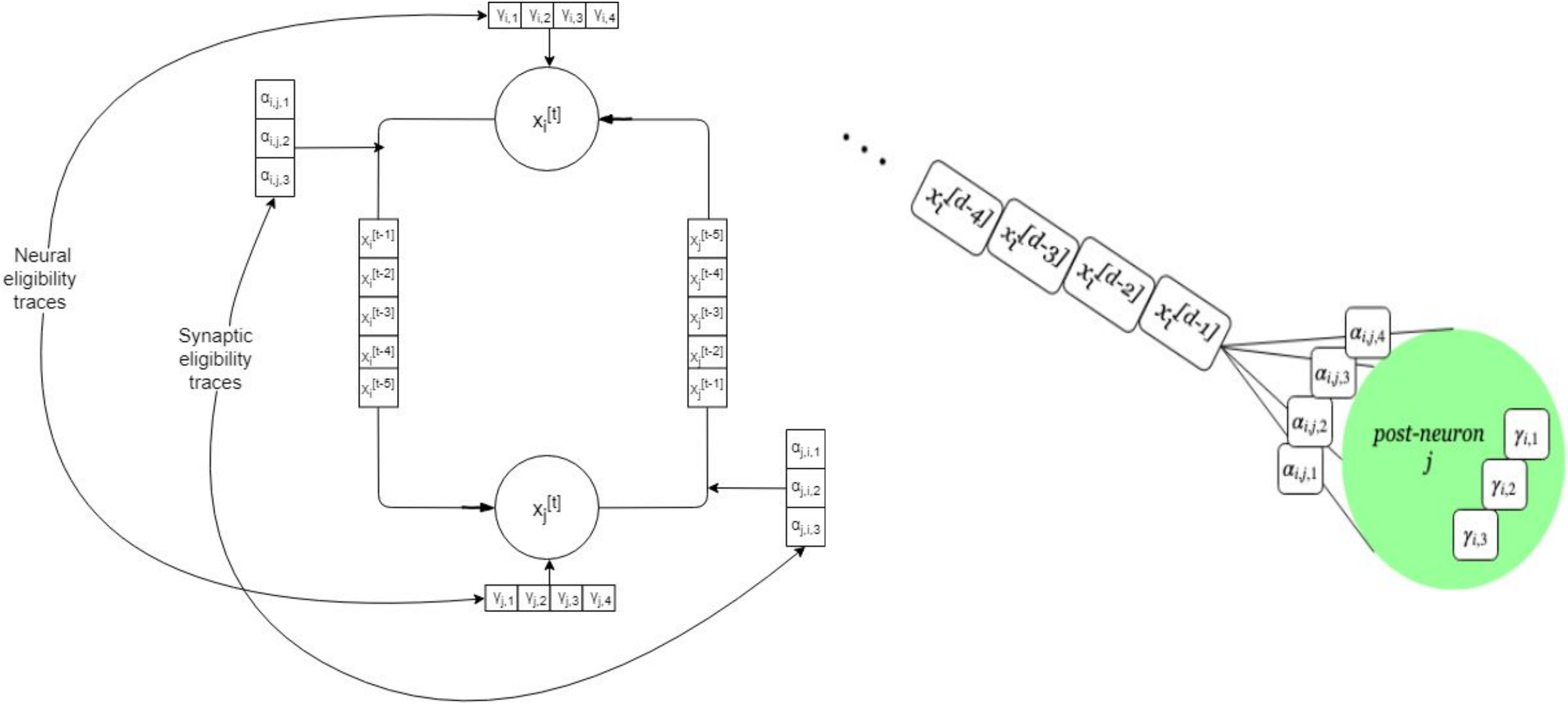
$$\nabla_\theta log\ p(\mathbf{x}) = \sum_{t=-L+1}^{0} \nabla_\theta log\ P_\theta(\mathbf{x}^{[t]}|\mathbf{x}^{(t-T,t-1]})$$

Time series Boltzmann Machine

$$\nabla_\theta log\ P_\theta(\mathbf{x}^{[t]}|\mathbf{x}^{(t-T,t-1]})$$
$$= -\tau^{-1}\sum_{j=1}^{N}\left(\nabla_\theta E_{\theta,j}(x_j^{[t]}|\mathbf{x}^{[:t-1]}) - \sum_{\tilde{x}_j^{[t]}\in\{0,1\}} P_{\theta,j}(\tilde{x}_j^{[t]}|\mathbf{x}^{(t-T,t-1]})\nabla_\theta E_{\theta,j}(\tilde{x}_j^{[t]}|\mathbf{x}^{[:t-1]})\right)$$

What if we want T $\rightarrow \infty$? Infinite no. of parameters, training becomes increasingly expensive

Dynamic Boltzmann Machine Structure



DyBM- Restrictions on Weights

$$W_{i,j}^{[\delta]} = \hat{W}_{i,j}^{[\delta]} + \hat{W}_{j,i}^{[-\delta]}$$

$$\hat{W}_{i,j}^{[\delta]} = \begin{cases} 0 & \delta = 0 \\ \sum_{k \in \mathcal{K}} u_{i,j,k} \lambda_k^{\delta - d_{i,j}} & \delta \geq d_{i,j} \\ \sum_{l \in \mathcal{L}} -v_{i,j,k} \mu_l^{-\delta} & otherwise \end{cases}$$

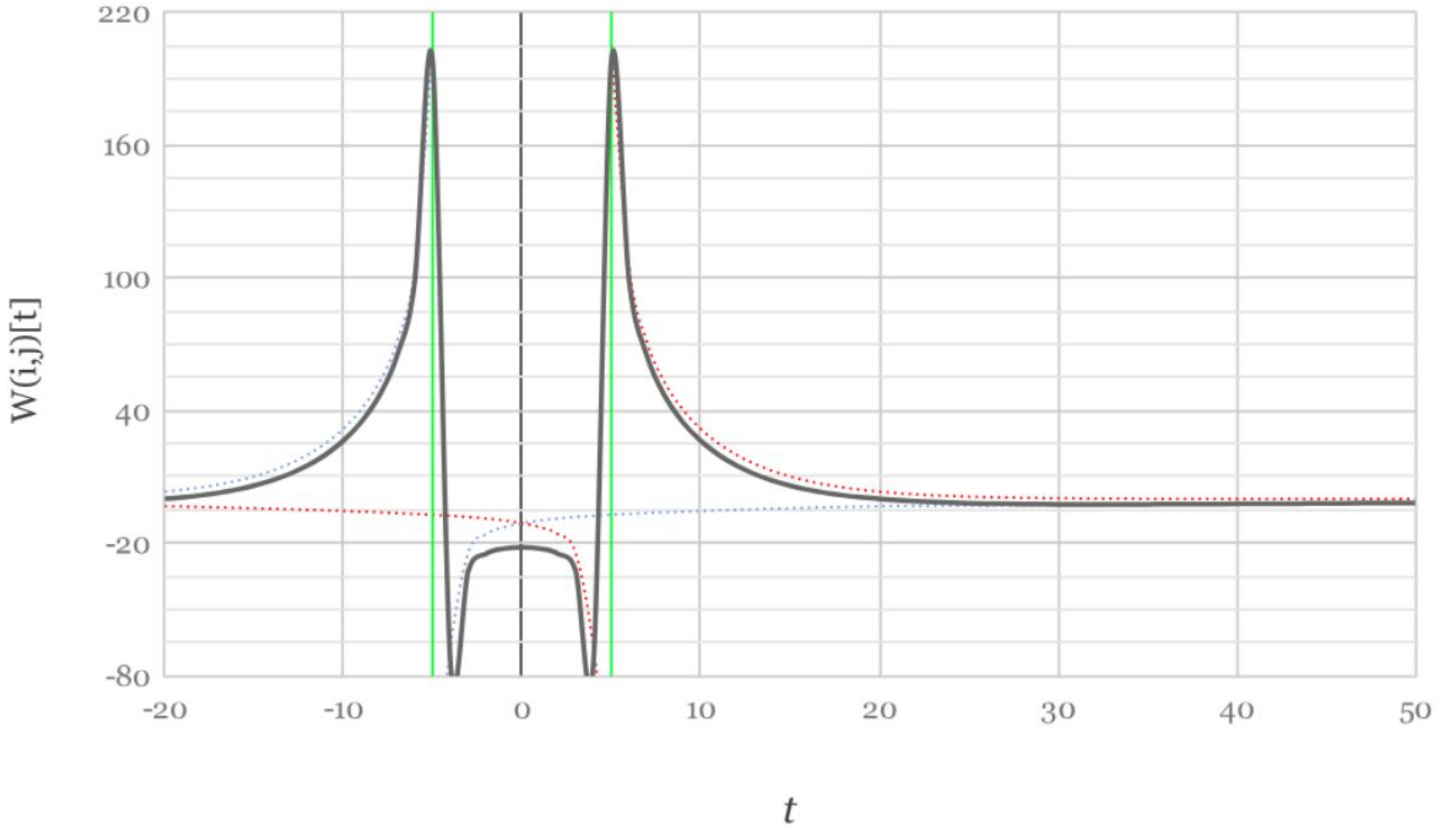
$$\hat{W}_{j,i}^{[-\delta]} = \begin{cases} 0 & \delta = 0 \\ \sum_{k \in \mathcal{K}} u_{j,i,k} \lambda_k^{-\delta - d_{i,j}} & \delta \leq -d_{i,j} \\ \sum_{l \in \mathcal{L}} -v_{j,i,k} \mu_l^{\delta} & otherwise \end{cases}$$

$$E_\theta(x_j^{[t]} | \mathbf{x}^{[:t-1]}) = -b_j x_j^{[t]} + \sum_{\delta=-\infty}^{t-1} (\mathbf{x}^{[\delta]})^T W_{:,j}^{[t-\delta]} x_j^{[t]}$$

$$(\mathbf{x}^{[\delta]})^T W_{:,j}^{[t-\delta]} = \sum_{i=1}^{N} x_i^{[\delta]} W_{i,j}^{[t-\delta]}$$

Allows for T $\rightarrow \infty$ with small no. of parameters

DyBM- Restrictions on Weights



DyBM - Energy Term

$$\alpha_{i,j,k}^{[t-1]} := \sum_{\delta=-\infty}^{-d_{i.j}} \lambda_k^{-\delta-d_{i,j}} x_i^{[t+\delta]}$$

$$\beta_{i,j,l}^{[t-1]} := \sum_{\delta=-d_{i.j}+1}^{-1} \mu_l^{\delta} x_i^{[t+\delta]}$$

$$\gamma_{i,l}^{[t-1]} := \sum_{\delta=-\infty}^{-1} \mu_l^{-\delta} x_i^{[t+\delta]}$$

$$\alpha_{i,j,k}^{[t]} \leftarrow \lambda_k(\alpha_{i,j,k}^{[t-1]} + x_j^{[t+d_{i,j}-1]})$$

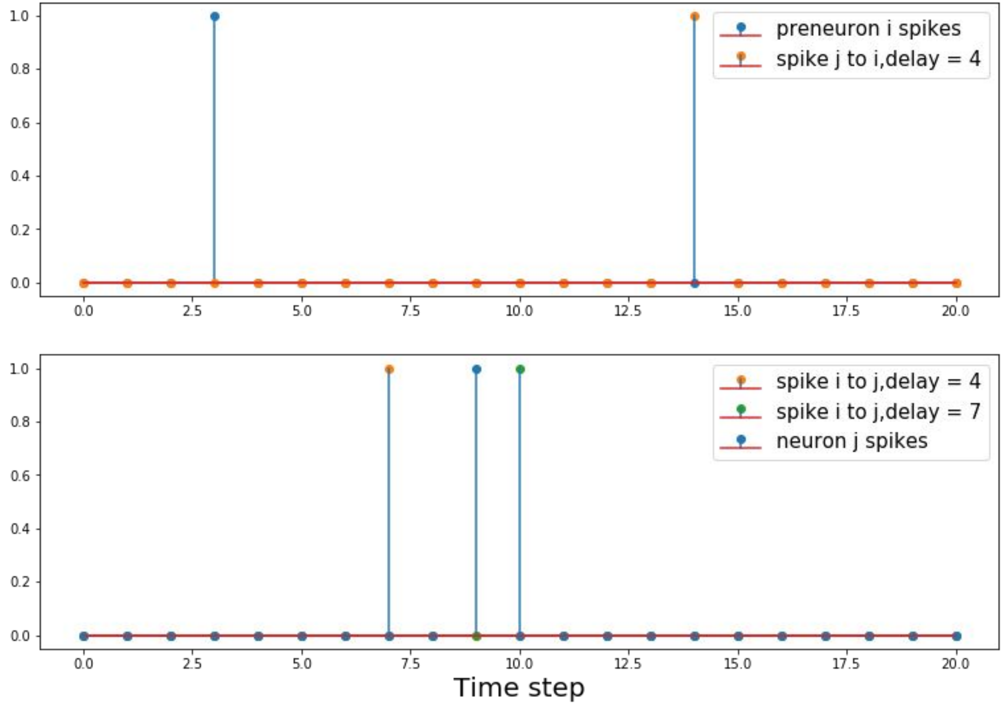
$$\gamma_{i,l}^{[t]} \leftarrow \mu_l(\gamma_{i,l}^{[t-1]} + x_i^{[t]})$$

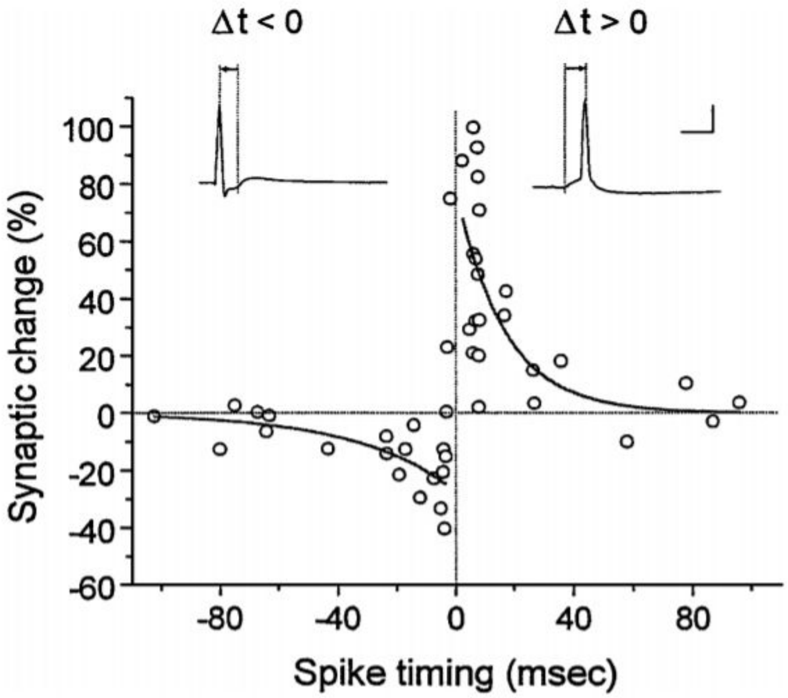
α- Synaptic Eligibility Trace

γ- Neural Eligibility Trace

$$E_\theta(\mathbf{x}^{[t]}|\mathbf{x}^{[:t-1]}) = \sum_{j=1}^{N}\Big\{-b_j - \sum_{i=1}^{N}\sum_{k\in\mathcal{K}} u_{i,j,k}\alpha_{i,j,k}^{[t-1]} + \sum_{i=1}^{N}\sum_{l\in\mathcal{L}} v_{j,i,k}\beta_{i,j,l}^{[t-1]} + \sum_{i=1}^{N}\sum_{l\in\mathcal{L}} v_{i,j,k}\gamma_{i,l}^{[t-1]}\Big\}x_j^{[t]}$$

Spike Time Dependent Plasticity





LTP and LTD in DyBM

$$E_{\theta,j}(x_j^{[t]}|\mathbf{x}^{[:t-1]}) = \Big\{ -b_j + \sum_{i=1}^{N}\sum_{k\in\mathcal{K}} u_{i,j,k}\alpha_{i,j,k}^{[t-1]} + \sum_{i=1}^{N}\sum_{l\in\mathcal{L}} v_{j,i,k}\beta_{i,j,l}^{[t-1]} + \sum_{i=1}^{N}\sum_{l\in\mathcal{L}} v_{i,j,k}\gamma_{i,l}^{[t-1]} \Big\} x_j^{[t]}$$

1st term- Higher the bias more the prob. of spiking
2nd term- LTP considering excitatory synapse(i,j)
3rd term- LTD considering excitatory synapse(i,j)
4th term- LTD considering excitatory synapse(j,i)

The reason for selecting that particular type of weight profile can be backtraced

DyBM- Weight Update Equation

$$b_j \leftarrow b_j + \eta\tau^{-1}(x_j^{[t]} - \langle X_j^{[t]}\rangle_\theta)$$

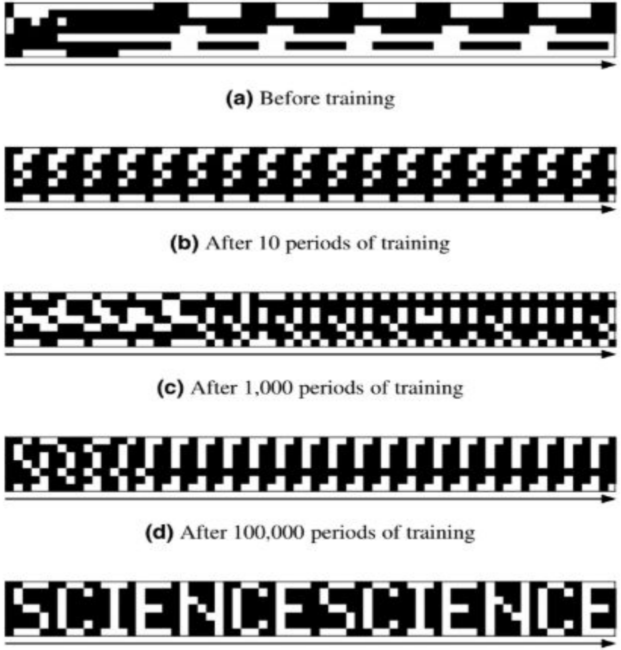
$$u_{i,j,k} \leftarrow u_{i,j,k} + \alpha_{i,j,k}^{[t-1]}\tau^{-1}(x_j^{[t]} - \langle X_j^{[t]}\rangle_\theta)$$

$$v_{i,j,l} \leftarrow v_{i,j,l} - \beta_{i,j,l}^{[t-1]}\tau^{-1}(x_j^{[t]} - \langle X_j^{[t]}\rangle_\theta) - \gamma_{j,l}^{[t-1]}\tau^{-1}(x_j^{[t]} - \langle X_j^{[t]}\rangle_\theta)$$

$$\begin{aligned}
\langle X_j^{[t]}\rangle_\theta &= \mathbb{E}_{\theta,j}(x_j^{[t]}|\mathbf{x}^{[:t-1]}) \\
&= 1 \times P_{\theta,j}(1|\mathbf{x}^{[:t-1]}) + 0 \times P_{\theta,j}(0|\mathbf{x}^{[:t-1]}) \\
&= \frac{e^{-\tau^{-1}E_{\theta,j}(1|\mathbf{x}^{[:t-1]})}}{1 + e^{-\tau^{-1}E_{\theta,j}(1|\mathbf{x}^{[:t-1]})}}
\end{aligned}$$

Weight Update is like STDP with Homeostatic Plasticity term which does not allow weights to blow up

DyBM- Memorising an Image



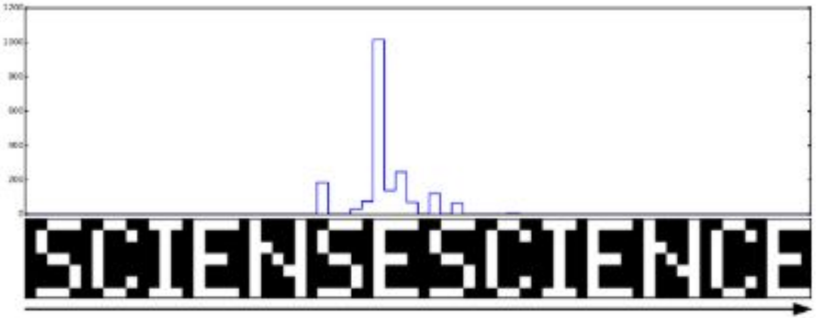
(a) Before training

(b) After 10 periods of training

(c) After 1,000 periods of training

(d) After 100,000 periods of training

(e) After 130,000 periods of training



Can memorise a sequence and detect an anomaly and its location if slightly perturbed sequence is presented

Gaussian DyBM

- Neurons in DyBM can only take binary values
- Extend Gaussian Boltzmann Machine to a Gaussian DyBM
- Sample for a Gaussian distribution

$$p(x_j^{[t]} | \mathbf{x}^{[t-T,t-1]}) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(x_j^{[t]} - \mu_j^{[t]})^2}{2\sigma_j^2}} \qquad \mu_j^{[t]} = b_j^{[t]} + \sum_{\delta=1}^{T} \sum_{i=1}^{N} w_{i,j}^{[\delta]} x_i^{[t-\delta]} \qquad w_{i,j}^{[\delta]} = \sum_{k=1}^{K} \lambda^{\delta - d_{i,j}} u_{i,j,k}$$

$$\mu_j^{[t]} = b_j^{[t]} + \sum_{\delta=1}^{\delta - d_{i,j}} \sum_{i=1}^{N} w_{i,j}^{[\delta]} x_i^{[t-\delta]} + \sum_{i=1}^{N} \sum_{k=1}^{K} \alpha_{i,j,k}^{[t-1]} u_{i,j,k} \qquad \alpha_{i,j,k}^{[t-1]} = \sum_{\delta=d_{i,j}}^{\infty} \lambda^{\delta - d_{i,j}} x_i^{[t-\delta]}$$

No restriction for $\boldsymbol{\delta} < d_{i,j}$ as in case of DyBM. Allows us to model G-DyBM as Extended VAR

RNN Gaussian DyBM

- Include a RNN layer with M dimensions for predicting bias (N dimensional) term
- Helpful when time series is highly non-linear
- Here RNN is similar to an echo state network

$$\mathbf{b}^{[t]} = \mathbf{b}^{[t-1]} + \mathbf{A}^\top \Psi^{[t]}$$

$$b_j \leftarrow b_j + \eta \frac{(x_j^{[t]} - \mu_j^{[t]})}{\sigma_j^2},$$

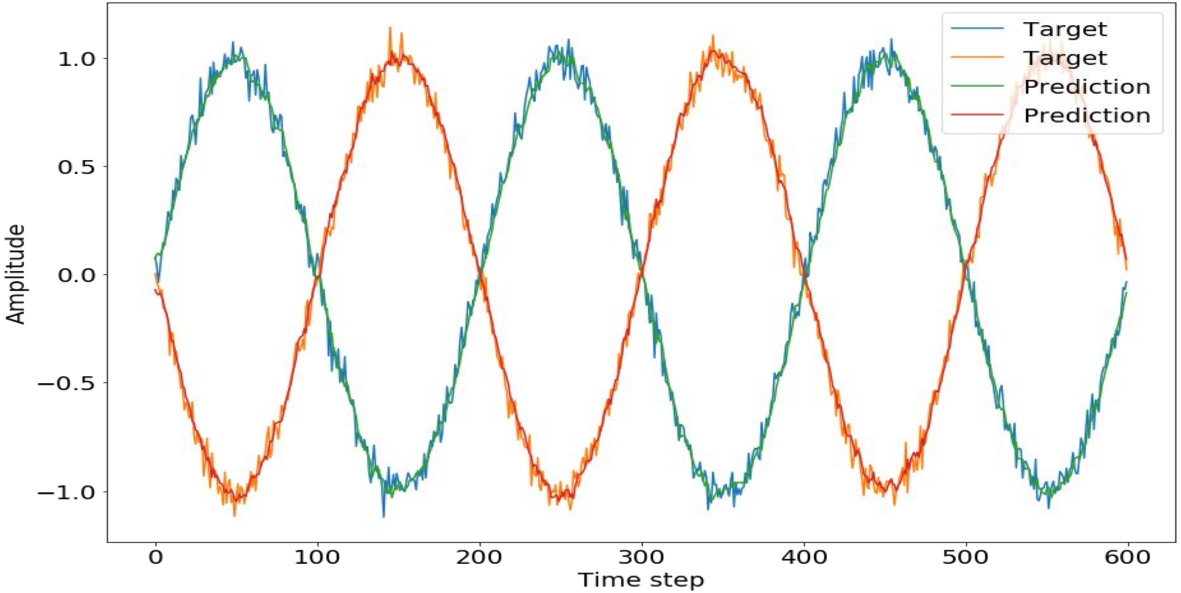
$$\sigma_j \leftarrow \sigma_j + \eta \left(\frac{(x_j^{[t]} - \mu_j^{[t]})^2}{\sigma_j^2} - 1 \right) \frac{1}{\sigma_j},$$

$$w_{i,j}^{[\delta]} \leftarrow w_{i,j}^{[\delta]} + \eta \frac{(x_j^{[t]} - \mu_j^{[t]})}{\sigma_j^2} x_i^{[t-\delta]},$$

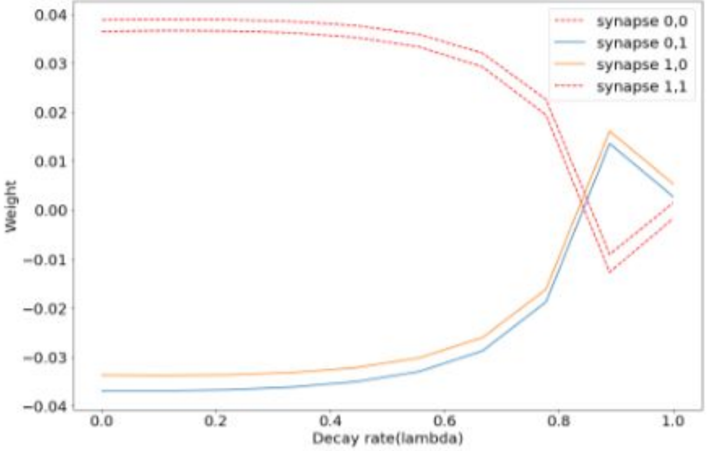
$$u_{i,j,k} \leftarrow u_{i,j,k} + \eta \frac{(x_j^{[t]} - \mu_j^{[t]})}{\sigma_j^2} \alpha_{i,j,k}^{[t-1]}$$

$$A_{l,j} \leftarrow A_{l,j} + \eta' \frac{(x_j^{[t]} - \mu_j^{[t]})}{\sigma_j^2} \psi_l^{[t]},$$

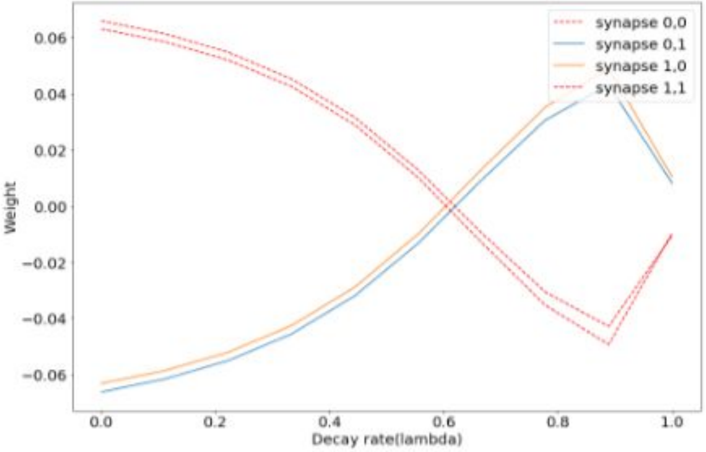
Multidimensional Sine Wave



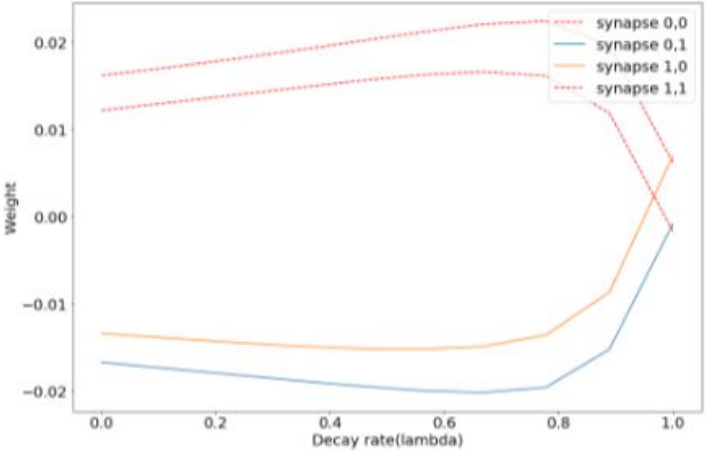
T=200
No. of traces = 10
FIFO length=7



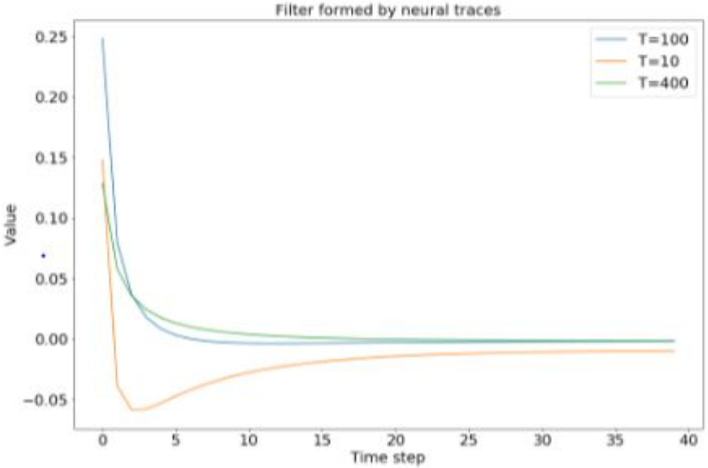
(a) Weights $u_{i,j,k}$ of Synaptic traces, T=100



(b) Weights $u_{i,j,k}$ of Synaptic traces, T=10

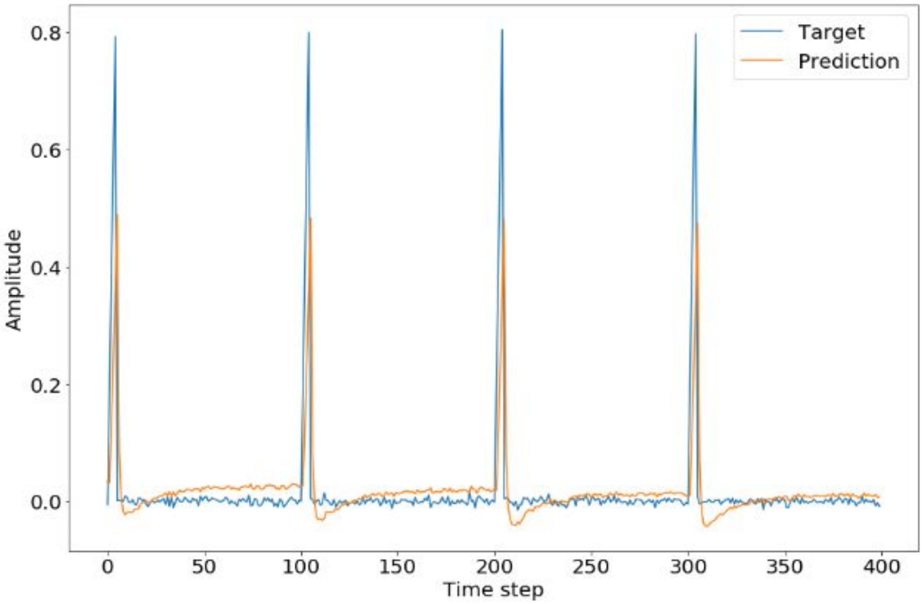


(c) Weights $u_{i,j,k}$ of Synaptic traces, T=400

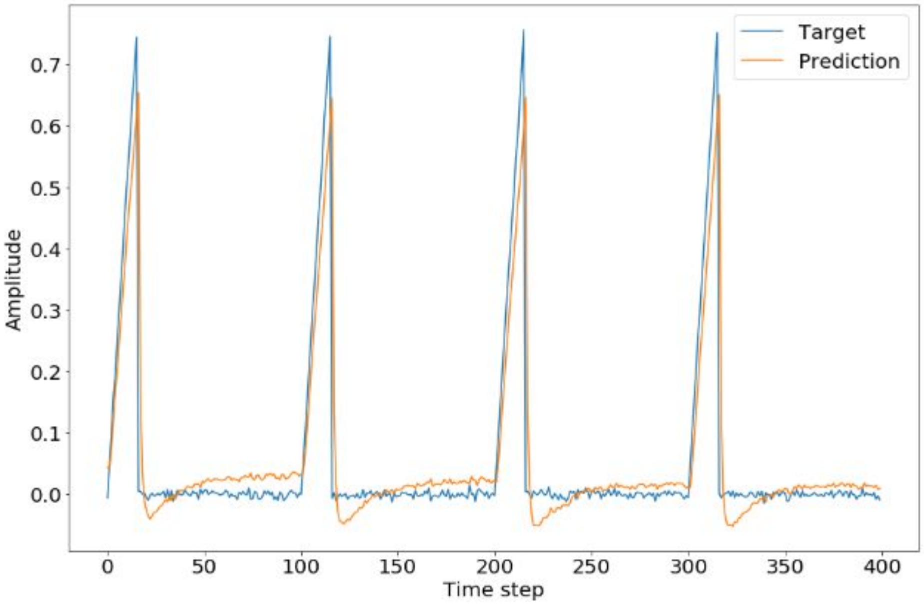


(d) Filter Response

Spike Train



(a) Spike Train Input, Spike Width=5



(b) Spike Train Input, Spike Width=15

Experiments

Dataset	Variable(s)	Link
Multidimensional Noisy Sine wave	5-dimensional sine wave with Gaussian noise	Synthetic
PACIFIC Exchange rate service (2007-2019)	Top 10 foreign currencies and INR w.r.t USD	http://fx.sauder.ubc.ca/data.html
Monthly sunspot number, Zurich (1749-1983)	Sun-spot number	http://www.sidc.be/silso/datafiles-old

DyBM network for experiments

Index
0
1
2

Number
58.0
62.6
70.0

CAD	EUR	JPY	GBP	CHF	AUD	HKD	NZD	KRW	MXN	INR
1.1713	0.75911	117.518	0.51075	1.22278	1.2719	7.8117	1.44	941.56	11.1872	44.084
1.1753	0.75865	116.841	0.51397	1.21742	1.2765	7.8149	1.4519	943.26	11.1827	44.118
1.1811	0.76372	116.022	0.51992	1.22191	1.2941	7.8152	1.4743	951.73	11.1741	44.436

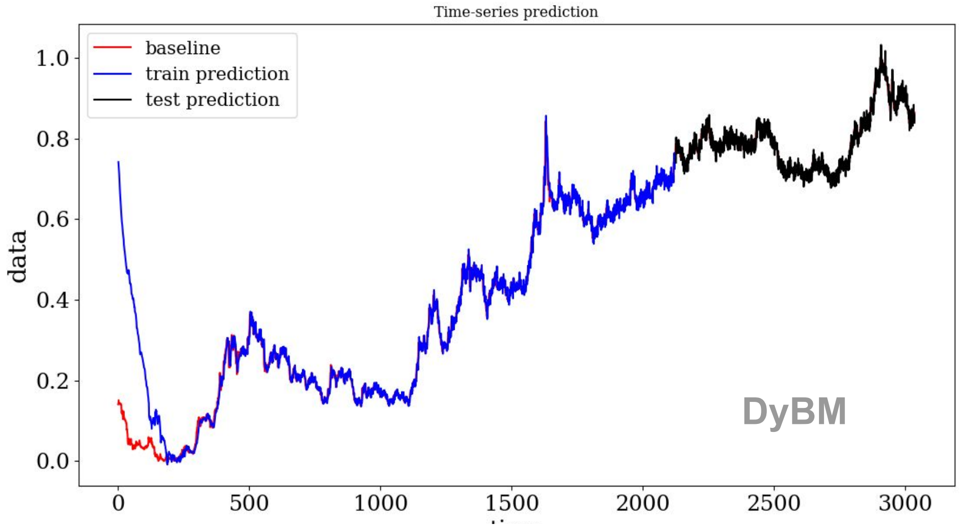
#Neurons = Dim(data)

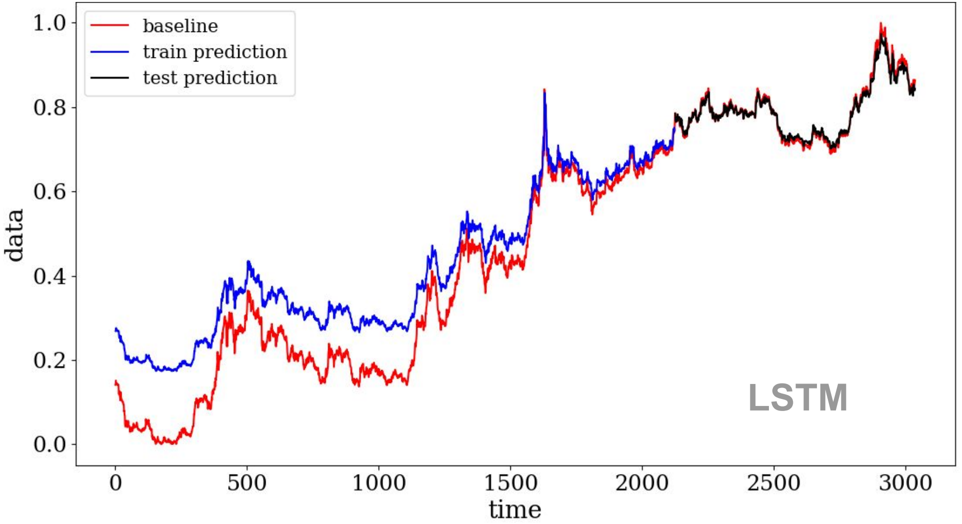
Decay = [0, 0.05, 0.16,..., 0.5]

Delay = 10-20

INR/USD - 2007-2019

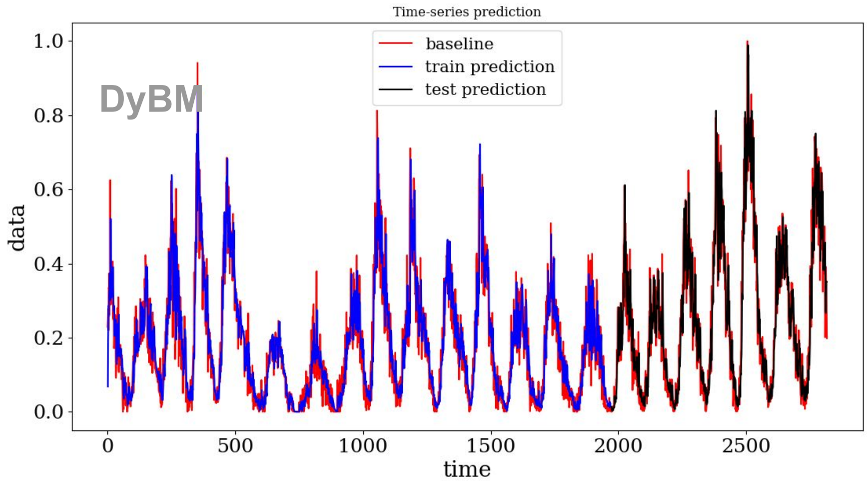
	DyBM	LSTM
Mean train error	0.08196	0.09791
Mean test error	0.01513	0.01082
Per epoch time to learn	0.80660	67.41648

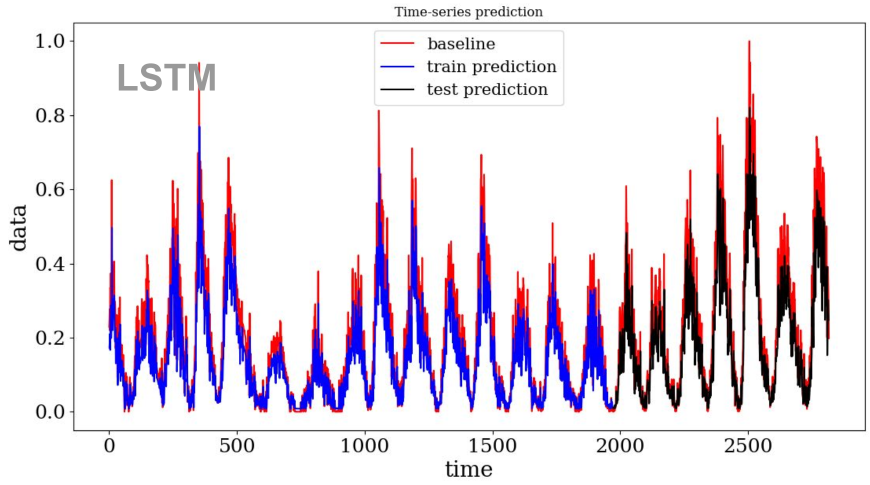




Sunspot number - Zurich

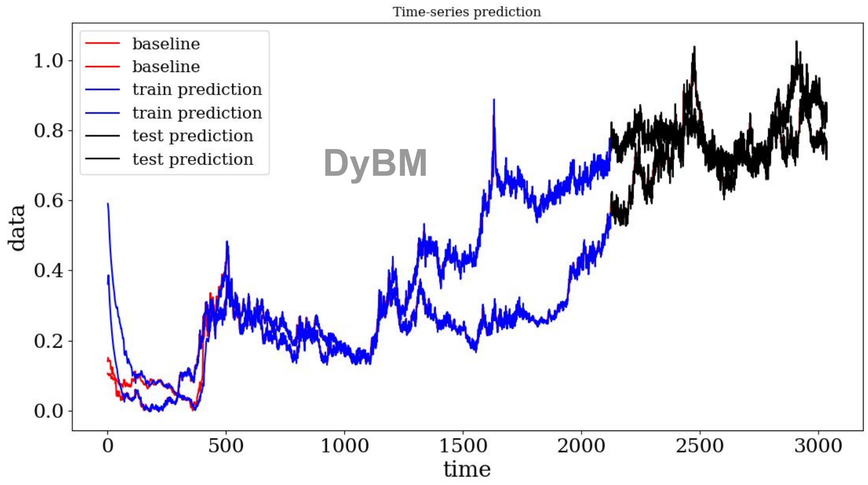
	DyBM	LSTM
Mean train error	0.07036	0.07810
Mean test error	0.08432	0.09655
Per epoch time to learn	0.80092	35.33157

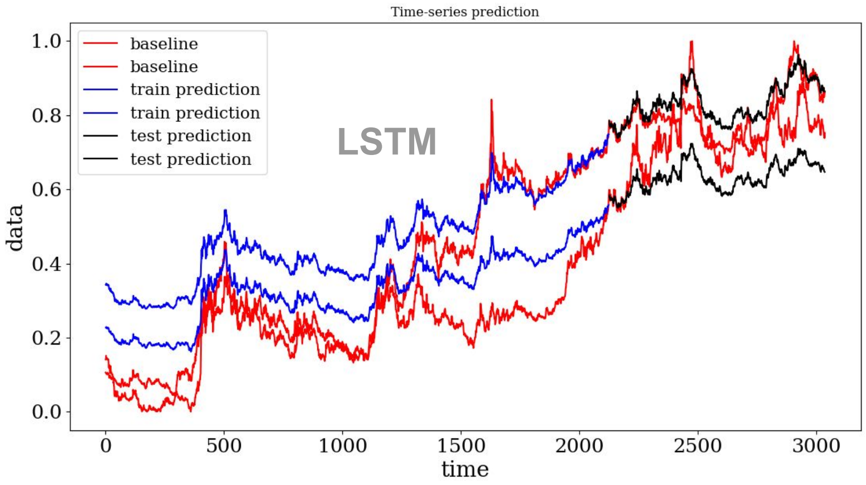




2-D time-series of exhange rates of MXN and INR

	DyBM	LSTM
Mean train error	0.05623	0.13958
Mean test error	0.03628	0.09103
Per epoch time to learn	0.98261	111.95985





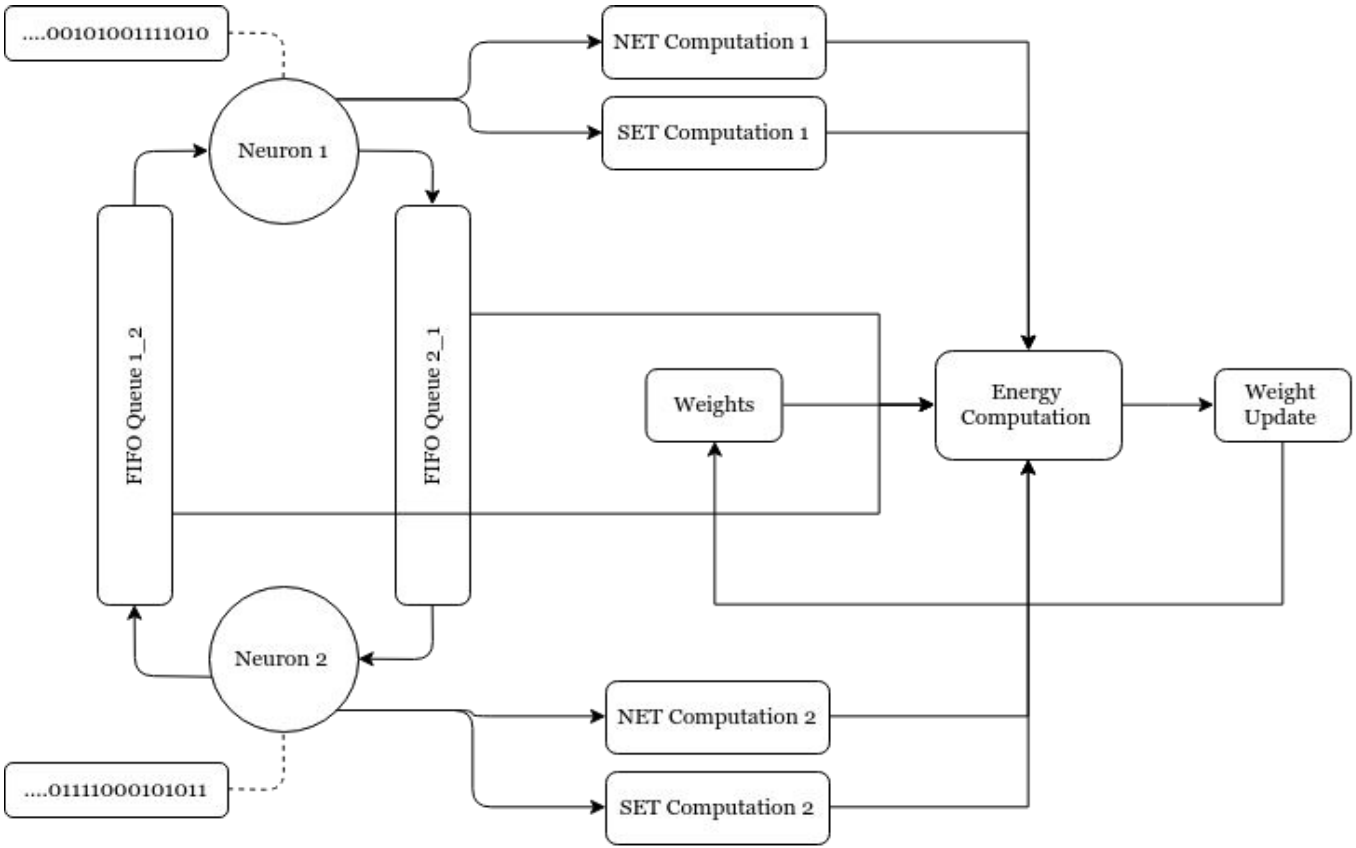
Conclusions

- DyBM learns a time-series much faster as compared to LSTM
- Thus the test error is consistently better in case of DyBM
- Time/epoch is always 15-25 times better in DyBM
- Very few resources in DyBM as compared to LSTM framework
- Customizable delay and decay in DyBM are ideal for very high-dimensional data
- Predicting uncertainty of the model in its prediction
- Highly suitable for online learning applications

Other Extensions of DyBM

- Delay Pruning- Same as dropout for Artificial Neural Network
- Functional DyBM- DyBM for infinite dimensonal data using gaussian process and Reproducing Hilbert kernel
- Time Discounted Convolution
- Second order Moments prediction
- Hidden layers

Mapping to Hardware



Mapping to Hardware

- Buesing mapped a network of spiking neurons to a Boltzmann Machine
- Neftci simulated Restricted Boltzmann Machine completely for hardware using LIF neurons and synapses
- Can something similar be done with DyBM??
- The weight update term already has LTP and LTD terms. How to include synaptic and neural traces?