CERIAS

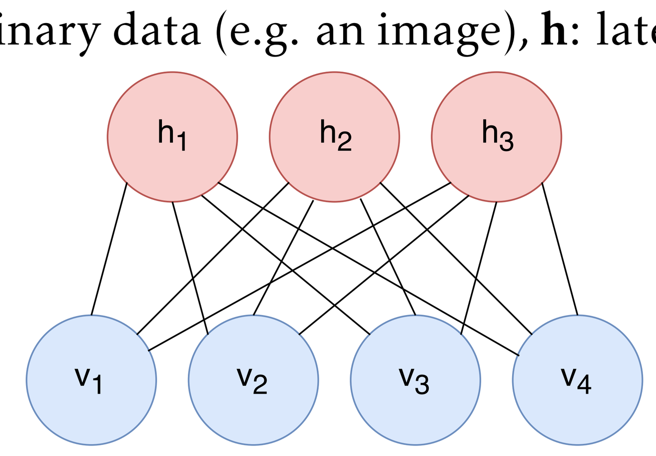
The Center for Education and Research in Information Assurance and Security

From Monte Carlo to Las Vegas: Understanding if Undirected Neural Networks can Really Generate Fake Images

Pedro Savarese[1], Mayank Kakodkar[2], Bruno Ribeiro[2]
[1] Toyota Technological Institute at Chicago [2] Purdue University

Restricted Boltzmann Machines

- Generative model, capable of learning latent representations
- **v**: observable binary data (e.g. an image), **h**: latent binary vectors



$$p(\mathbf{x} = (\mathbf{v}, \mathbf{h})) = \frac{1}{Z} e^{-E(\mathbf{x})}, \quad Z = \sum_{\mathbf{x}} e^{-E(\mathbf{x})}, \quad E(\mathbf{x}) = -\mathbf{v}^T \mathbf{W} \mathbf{h}$$

Maximum Likelihood Training

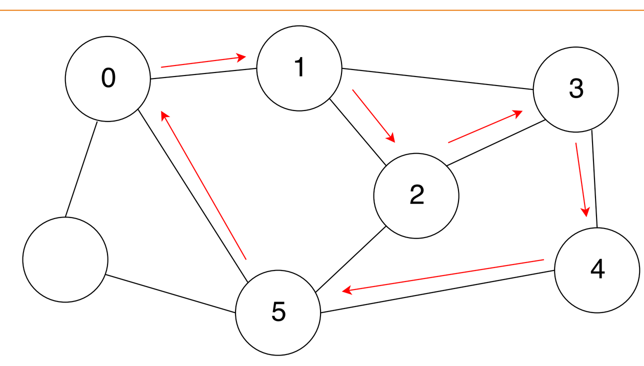
$$\nabla_{\mathbf{W}} \mathcal{L} = \left[\frac{1}{N} \sum_{n=1}^{N} \mathbf{v}_n \mathbb{E}[\mathbf{h}|\mathbf{v}_n]^T \right]_{\text{easy}} - \left[\mathbb{E}[\mathbf{v}\mathbf{h}^T] \right]_{\text{intractable}}$$

Markov-Chain Monte Carlo

- Sample $\mathbf{x}^{(i+1)} \sim p(\mathbf{x}|\mathbf{x}^{(i)})$, randomly initialized $\mathbf{x}^{(0)}$
- $\mathbb{E}[\mathbf{v}\mathbf{h}^T]$ estimated from $\mathbf{v}^{(K)} {\mathbf{h}^{(K)}}^T$, with fixed K
- Asymptotic unbiased-ness requires steady state, not guaranteed
- Contrastive Divergence: initialize $\mathbf{x}^{(0)}$ from Training data

Markov-Chain Las Vegas

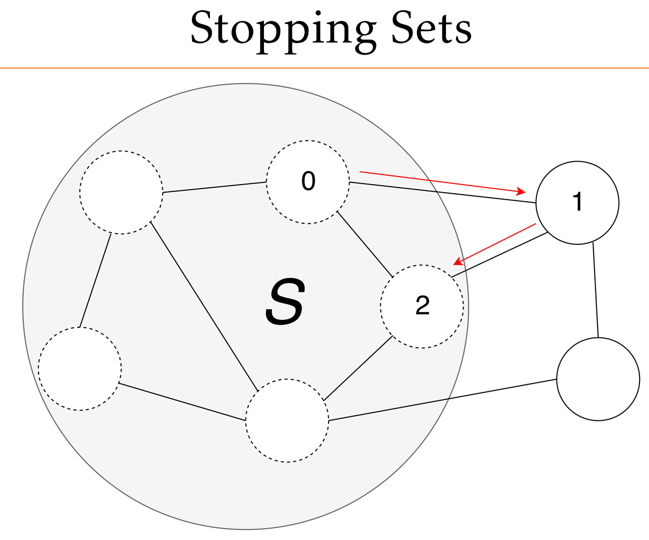
Strong Markov Property
Tours are Sample Paths from the Steady State Distribution



$$T := \mathbf{x}^{(0)} \to \mathbf{x}^{(1)} \to \mathbf{x}^{(2)} \to \mathbf{x}^{(3)} \to \mathbf{x}^{(4)} \to \mathbf{x}^{(5)} \to \mathbf{x}^{(6)} = \mathbf{x}^{(0)}$$

- $$\left(e^{-E(\mathbf{x}^{(0)})} \sum_{\mathbf{x} \equiv (\mathbf{v}, \mathbf{h}) \in T} \mathbf{v}\mathbf{h}^T \right)$$
  is an unbiased estimate of $Z \cdot \mathbb{E}[\mathbf{v}\mathbf{h}^T]$.
- Similarly, $e^{-E(\mathbf{x}^{(0)})}|T|$ is an estimate of the partition function Z
- Exactly Sampled from the steady state, but Random Running Time
- Uses all states in the chain to compute estimate

Stopping Sets



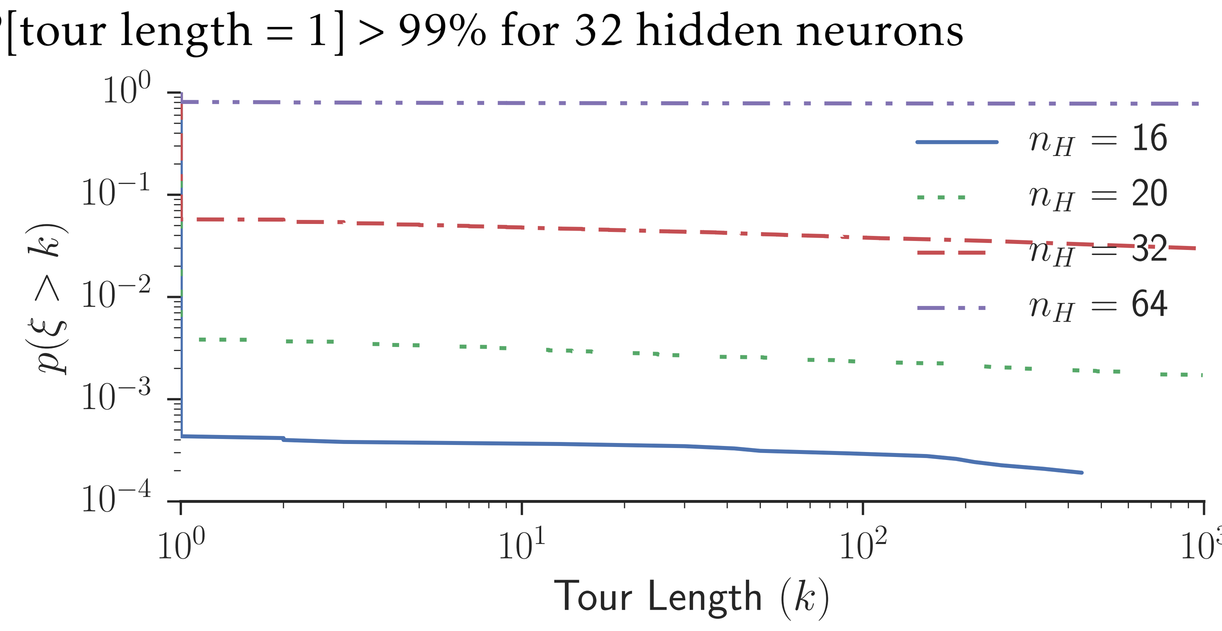
- Relax tour stopping condition: require $\mathbf{x}^{(i)} \in S$ instead of $\mathbf{x}^{(i)} = \mathbf{x}^{(0)}$

$$\uparrow \Pr(S) \implies \downarrow \text{running time}$$

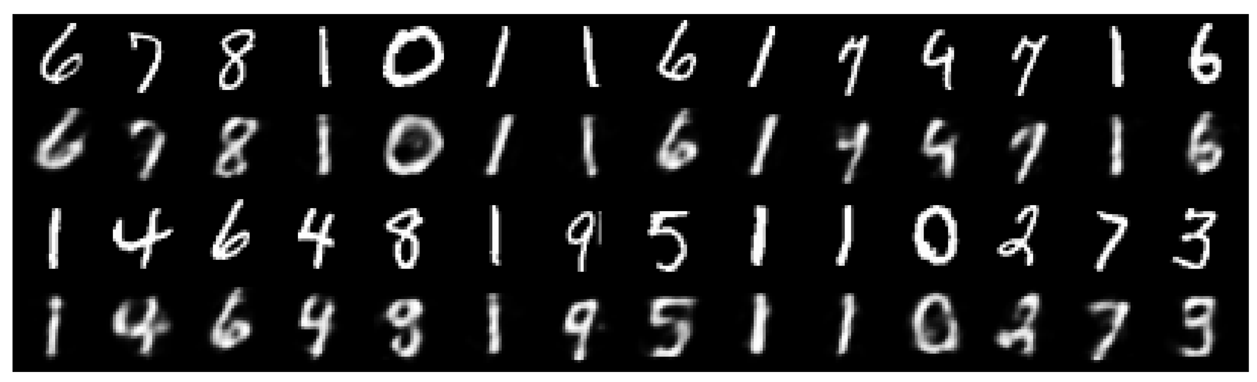
- Increases exact stopping set gradient computation time

Tour Behavior

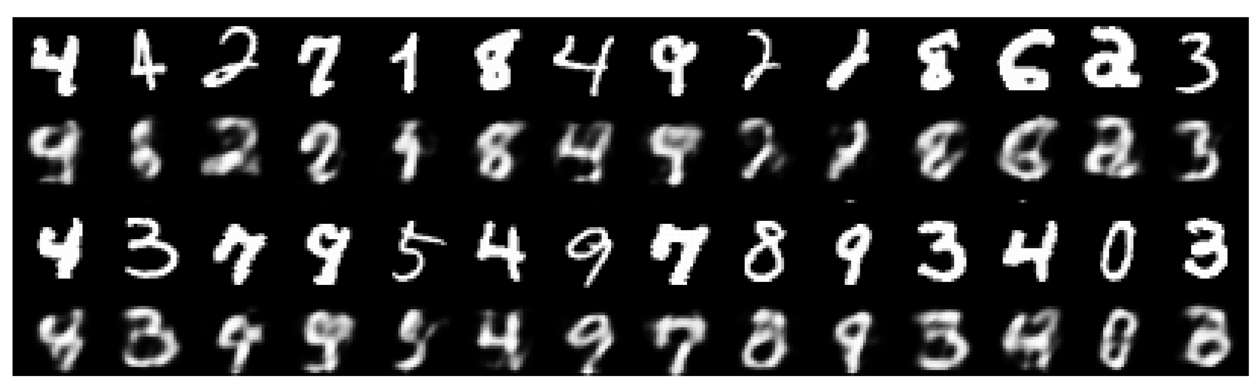
- Extremely Heavy Tail
- P[tour length = 1] > 99% for 32 hidden neurons



- Shorter tours $\implies$ Memorized data



- Longer tours $\implies$ Better generalization



- Ignoring long tours aids generalization

Gradient Estimates (Las Vegas Slope)

Stopping Set ≡ Training Data (high probability states)

- Use Tours which finish in $\leq K$ steps $\implies$ limit max. running time
- Heavy tail allows ignoring long tours
- Better training of RBMs on MNIST

	$n_H = 32$		$n_H = 25$	
Method	**Training**	**Testing**	**Training**	**Testing**
CD-1	-167.3 (2.7)	-166.6 (2.8)	-169.8 (2.6)	-169.0 (2.6)
PCD-1	-153.0 (4.9)	-152.1 (4.7)	-147.8 (0.5)	-147.0 (0.5)
LVS-1	**-134.0 (1.0)**	**-133.3 (1.0)**	**-138.3 (1.3)**	**-137.5 (1.4)**
CD-10	-154.3 (3.3)	-153.4 (3.3)	-156.4 (0.5)	-155.6 (0.5)
PCD-10	-139.3 (3.2)	-138.5 (3.3)	-147.4 (0.5)	-146.7 (0.5)
LVS-10	**-133.3 (1.0)**	**-132.6 (1.0)**	**-138.1 (1.1)**	**-137.4 (1.2)**

MNIST test log-likelihood (higher is better)

Method Comparison

	MLE	MCMC	MCLV	LVS
Running Time	Intractable	Fixed	Random	Fixed
Gradients	Unbiased	Biased	Unbiased	Controllable bias
States per estimate	All	1	All in tour	All in tour
Confidence intervals	Yes	No	Yes	Yes

Future Work

- Apply techniques to directed models (Variational models and GANs)

Code Repository

https://github.com/PurdueMINDS/MCLV-RBM