Inductive Principles for Restricted Boltzmann Machine Learning

Benjamin Marlin
Department of Computer Science
University of British Columbia

Joint work with Kevin Swersky, Bo Chen and Nando de Freitas

Introduction: The Big Picture

Some facts about maximum likelihood estimation:

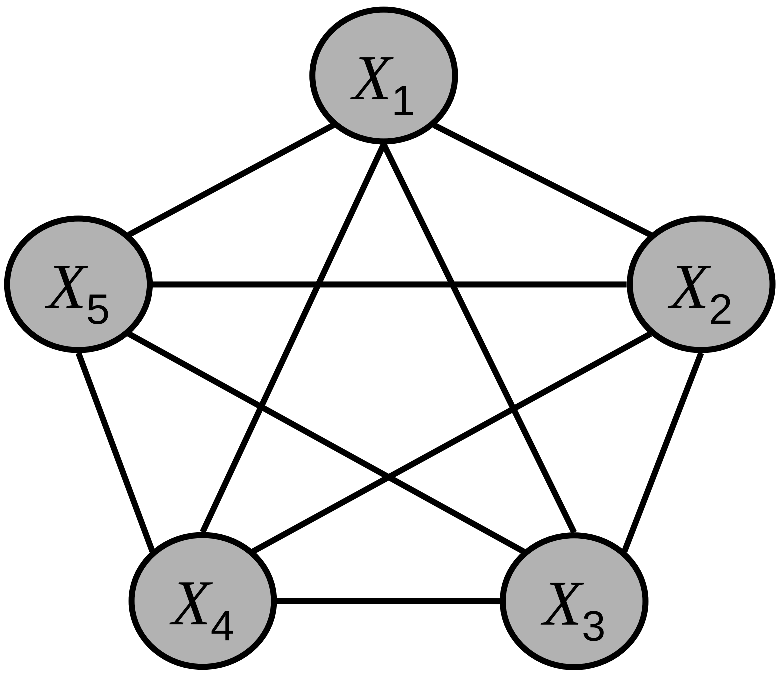
- ML is consistent (asymptotically unbiased)
- ML is statistically efficient (asymptotically lowest error)
- For certain model classes, computing the likelihood function can be **computationally intractable**.

This work studies alternative inductive principles for restricted Boltzmann machines that circumvent the computational intractability of the likelihood function at the expense of statistical consistency and/or efficiency.

Outline:

- Boltzmann Machines and RBMs
- Inductive Principles
    - Maximum Likelihood
    - Contrastive Divergence
    - Pseudo-Likelihood
    - Ratio Matching
    - Generalized Score Matching
- Experiments
- Demo

Introduction: Boltzmann Machines



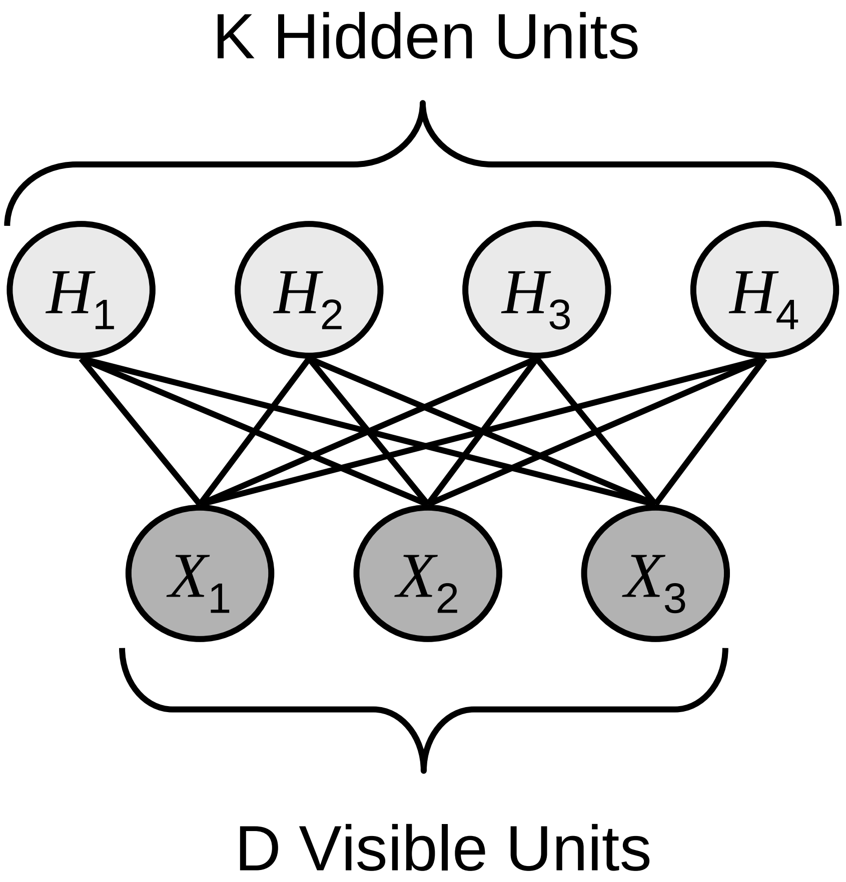
$$E_\theta(\boldsymbol{x}) = -(\boldsymbol{x}^T W \boldsymbol{x} + \boldsymbol{x}^T b)$$

$$P_\theta(\boldsymbol{x}) = \frac{1}{\mathcal{Z}} \exp\left(-E_\theta(\boldsymbol{x})\right)$$

$$\mathcal{Z} = \sum_{\boldsymbol{x}' \in \mathcal{X}} \exp\Big(-E_\theta(\boldsymbol{x}')\Big)$$

- A Boltzmann Machine is a Markov Random Field on D binary variables defined through a quadratic energy function.

Introduction: Restricted Boltzmann Machines



- A Restricted Boltzmann Machine (RBM) is a Boltzmann Machine with a bipartite graph structure.

- Typically one layer of nodes are fully observed variables (the visible layer), while the other consists of latent variables (the hidden layer).

Introduction: Restricted Boltzmann Machines

- The joint probability of the visible and hidden variables is defined through a bilinear energy function.

$$E_\theta(\boldsymbol{x}, \boldsymbol{h}) = -(\boldsymbol{x}^T W \boldsymbol{h} + \boldsymbol{x}^T b + \boldsymbol{h}^T c)$$

$$P_\theta(\boldsymbol{x}, \boldsymbol{h}) = \frac{1}{\mathcal{Z}} \exp\left(-E_\theta(\boldsymbol{x}, \boldsymbol{h})\right)$$

$$\mathcal{Z} = \sum_{\boldsymbol{x}' \in \mathcal{X}} \sum_{\boldsymbol{h}' \in \mathcal{H}} \exp\Big(-E_\theta(\boldsymbol{x}', \boldsymbol{h}')\Big)$$

Introduction: Restricted Boltzmann Machines

- The bipartite graph structure gives the RBM a special property: the visible variables are conditionally independent given the hidden variables and vice versa.

$$P_\theta(x_d = 1|\boldsymbol{h}) = \frac{1}{1 + \exp(-(\sum_{k=1}^{K} W_{dk}h_k + x_d b_d))}$$

$$P_\theta(h_k = 1|\boldsymbol{x}) = \frac{1}{1 + \exp(-(\sum_{d=1}^{D} W_{dk}x_d + h_k c_k))}$$

Introduction: Restricted Boltzmann Machines

- The marginal probability of the visible vector is obtained by summing out over all joint states of the hidden variables.

$$P_\theta(\boldsymbol{x}) = \frac{1}{\mathcal{Z}} \sum_{h \in \mathcal{H}} \exp\left(-E_\theta(\boldsymbol{x}, \boldsymbol{h})\right)$$

- This sum can be carried out analytically yielding an equivalent model defined in terms of a "free energy".

$$P_\theta(\boldsymbol{x}) = \frac{1}{\mathcal{Z}} \exp\left(-F_\theta(\boldsymbol{x})\right)$$

$$F_\theta(\boldsymbol{x}) = -\left(\boldsymbol{x}^T b + \sum_{k=1}^{K} \log\left(1 + \exp\left(\boldsymbol{x}^T W_k + c_k\right)\right)\right)$$

Introduction: Restricted Boltzmann Machines

• This construction eliminates the latent, hidden variables, leaving a distribution defined in terms of the visible variables.

• However, computing the normalizing constant (partition function) still has exponential complexity in D.

$$\mathcal{Z} = \sum_{\boldsymbol{x}' \in \mathcal{X}} \exp\left(-F_\theta(\boldsymbol{x}')\right)$$

• **This work is about inductive principles for RBM learning that circumvent the intractability of the partition function.**

Outline:

- Boltzmann Machines and RBMs
- Inductive Principles
    - Maximum Likelihood
    - Contrastive Divergence
    - Pseudo-Likelihood
    - Ratio Matching
    - Generalized Score Matching
- Experiments
- Demo

Stochastic Maximum Likelihood

- Exact maximum likelihood learning is intractable in an RBM. Stochastic ML estimation can instead be applied, usually using a simple block Gibbs sampler.

$$f^{ML}(\theta) = \sum_{\boldsymbol{x} \in \mathcal{X}} P_e(\boldsymbol{x}) \log P_\theta(\boldsymbol{x})$$

$$\nabla f^{ML} \approx -\left(\frac{1}{N} \sum_{n=1}^{N} \nabla F_\theta(\boldsymbol{x}_n) - \frac{1}{S} \sum_{s=1}^{S} \nabla F_\theta(\tilde{\boldsymbol{x}}_s) \right)$$

Contrastive Divergence

• The contrastive divergence principle results in a gradient that looks identical to stochastic maximum likelihood. The difference is that CD samples from the T-step Gibbs distribution.

$$f^{CD}(\theta) = \sum_{\boldsymbol{x} \in \mathcal{X}} P_e(\boldsymbol{x}) \log \left(\frac{P_e(\boldsymbol{x})}{P_\theta(\boldsymbol{x})} \right) - Q_\theta^t(\boldsymbol{x}) \log \left(\frac{Q_\theta^t(\boldsymbol{x})}{P_\theta(\boldsymbol{x})} \right)$$

$$\nabla f^{CD} \approx -\frac{1}{N} \left(\sum_{n=1}^{N} \nabla F_\theta(\boldsymbol{x}_n) - \nabla F_\theta(\tilde{\boldsymbol{x}}_n) \right)$$

Pseudo-Likelihood

- The principle of maximum pseudo-likelihood is based on optimizing a product of one-dimensional conditional densities under a log loss.

$$f^{PL}(\theta) = \sum_{\boldsymbol{x}\in\mathcal{X}} \sum_{d=1}^{D} P_e(\boldsymbol{x}) \log P_\theta(x_d|\boldsymbol{x}_{-d})$$

$$\nabla f^{PL} = \frac{-1}{N} \sum_{n,d} P_\theta(\overline{\boldsymbol{x}}_{dn}^{d}|\boldsymbol{x}_{-dn}) \left(\nabla F_\theta(\boldsymbol{x}_n) - \nabla F_\theta(\overline{\boldsymbol{x}}_n^d)\right)$$

Ratio Matching

- The ratio matching principle is very similar to pseudo-likelihood, but is based on minimizing a squared difference between one dimensional conditional distributions.

$$f^{RM}(\theta) = \sum_{\boldsymbol{x}\in\mathcal{X}} \sum_{d=1}^{D} \sum_{\xi\in\{0,1\}} P_e(\boldsymbol{x})\Big(P_\theta(X_d=\xi|\boldsymbol{x}_{-d}) - P_e(X_d=\xi|\boldsymbol{x}_{-d})\Big)^2$$

$$\nabla f^{RM} = \frac{2}{N}\sum_{n=1}^{N}\sum_{d=1}^{D} g(u_{dn})^3 u_{dn}\left(\nabla F_\theta(\boldsymbol{x}_n) - \nabla F_\theta(\bar{\boldsymbol{x}}_n^d)\right)$$

$$g(u) = \frac{1}{1+u}, \qquad u_{dn} = P_\theta(\boldsymbol{x}_n)/P_\theta(\bar{\boldsymbol{x}}_n^d)$$

Generalized Score Matching

- The generalized score matching principle is similar to ratio matching, except that the difference between inverse one dimensional conditional distributions is minimized.

$$f^{GSM}(\theta) = \sum_{\boldsymbol{x} \in \mathcal{X}} \sum_{d=1}^{D} P_e(\boldsymbol{x}) \left(\frac{1}{P_\theta(x_d|\boldsymbol{x}_{-d})} - \frac{1}{P_e(x_d|\boldsymbol{x}_{-d})} \right)^2$$

$$\nabla f^{GSM} = \frac{2}{N} \sum_{n=1}^{N} \sum_{d=1}^{D} (u_{dn}^{-2} - u_{dn}) \left(\nabla F_\theta(\boldsymbol{x}_n) - \nabla F_\theta(\bar{\boldsymbol{x}}_n^d) \right)$$

$$g(u) = u^{-2} - 2u, \quad u_{dn} = P_\theta(\boldsymbol{x}_n) / P_\theta(\bar{\boldsymbol{x}}_n^d)$$

Gradient Comparison

$$\nabla f^{ML} \approx -\left(\frac{1}{N}\sum_{n=1}^{N} \nabla F_\theta(\boldsymbol{x}_n) - \frac{1}{S}\sum_{s=1}^{S} \nabla F_\theta(\tilde{\boldsymbol{x}}_s)\right)$$

$$\nabla f^{CD} \approx -\frac{1}{N}\left(\sum_{n=1}^{N} \nabla F_\theta(\boldsymbol{x}_n) - \nabla F_\theta(\tilde{\boldsymbol{x}}_n)\right)$$

$$\nabla f^{PL} = \frac{-1}{N}\sum_{n,d} P_\theta(\bar{\boldsymbol{x}}_{dn}^{d}|\boldsymbol{x}_{-dn})\left(\nabla F_\theta(\boldsymbol{x}_n) - \nabla F_\theta(\bar{\boldsymbol{x}}_n^d)\right)$$

$$\nabla f^{RM} = \frac{2}{N}\sum_{n=1}^{N}\sum_{d=1}^{D} g(u_{dn})^3 u_{dn}\left(\nabla F_\theta(\boldsymbol{x}_n) - \nabla F_\theta(\bar{\boldsymbol{x}}_n^d)\right)$$

$$\nabla f^{GSM} = \frac{2}{N}\sum_{n=1}^{N}\sum_{d=1}^{D} (u_{dn}^{-2} - u_{dn})\left(\nabla F_\theta(\boldsymbol{x}_n) - \nabla F_\theta(\bar{\boldsymbol{x}}_n^d)\right)$$

Outline:

- Boltzmann Machines and RBMs
- Inductive Principles
  - Maximum Likelihood
  - Contrastive Divergence
  - Pseudo-Likelihood
  - Ratio Matching
  - Generalized Score Matching
- Experiments
- Demo

Experiments:

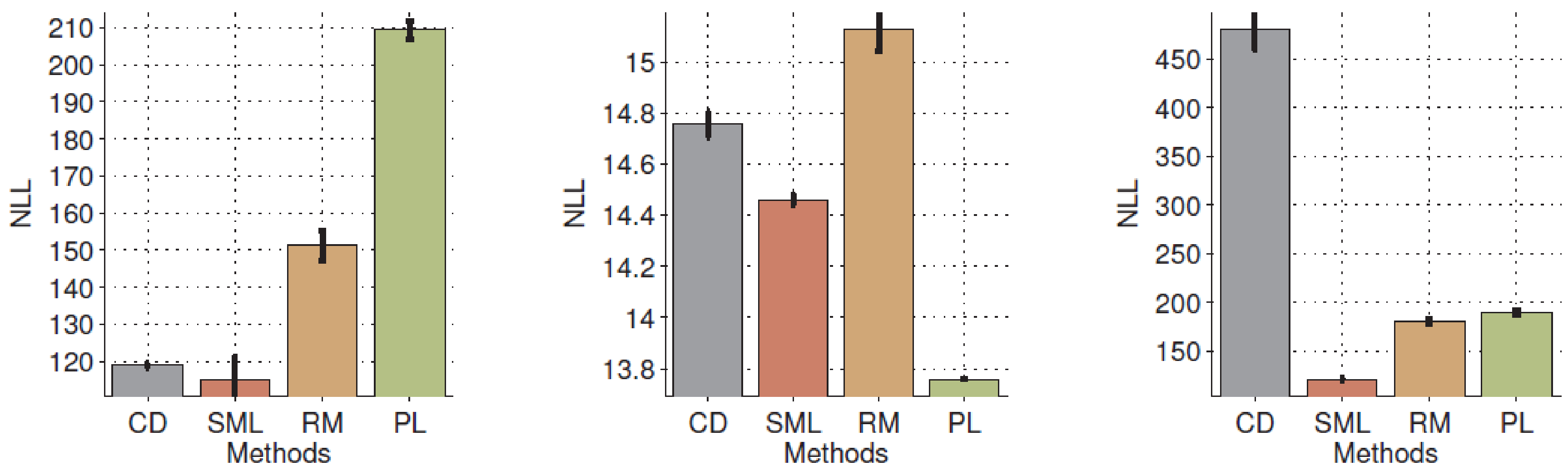
Data Sets:

- MNIST handwritten digits
- 20 News Groups
- CalTech 101 Silhouettes

Evaluation Criteria:

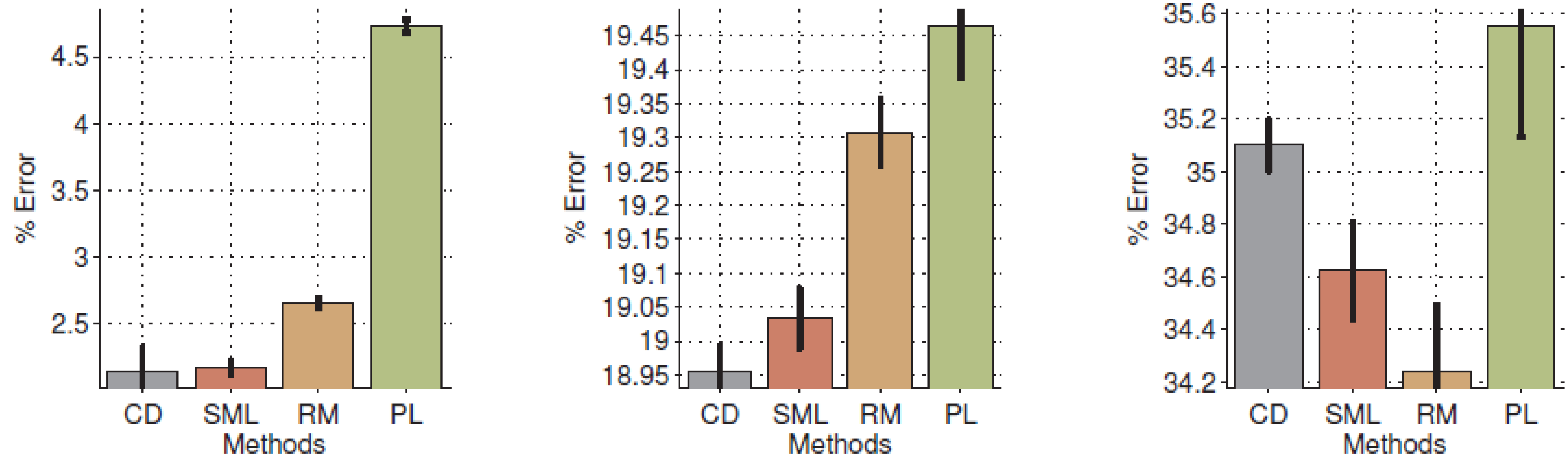
- Log likelihood (using AIS estimator)
- Classification error
- Reconstruction error
- De-noising
- Novelty detection

Experiments: Log Likelihood



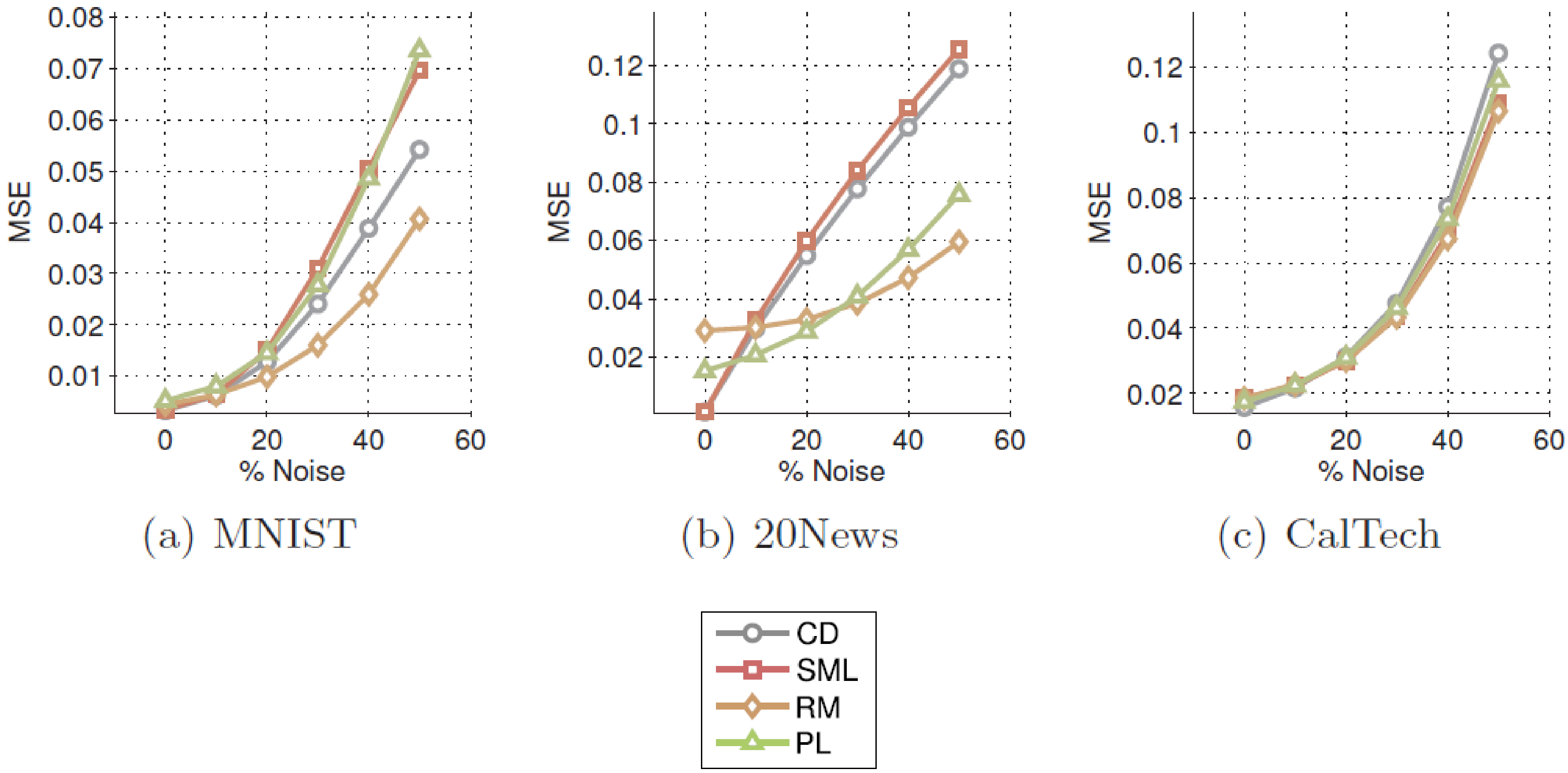
(a) MNIST (b) 20News (c) CalTech

Experiments: Classification Error



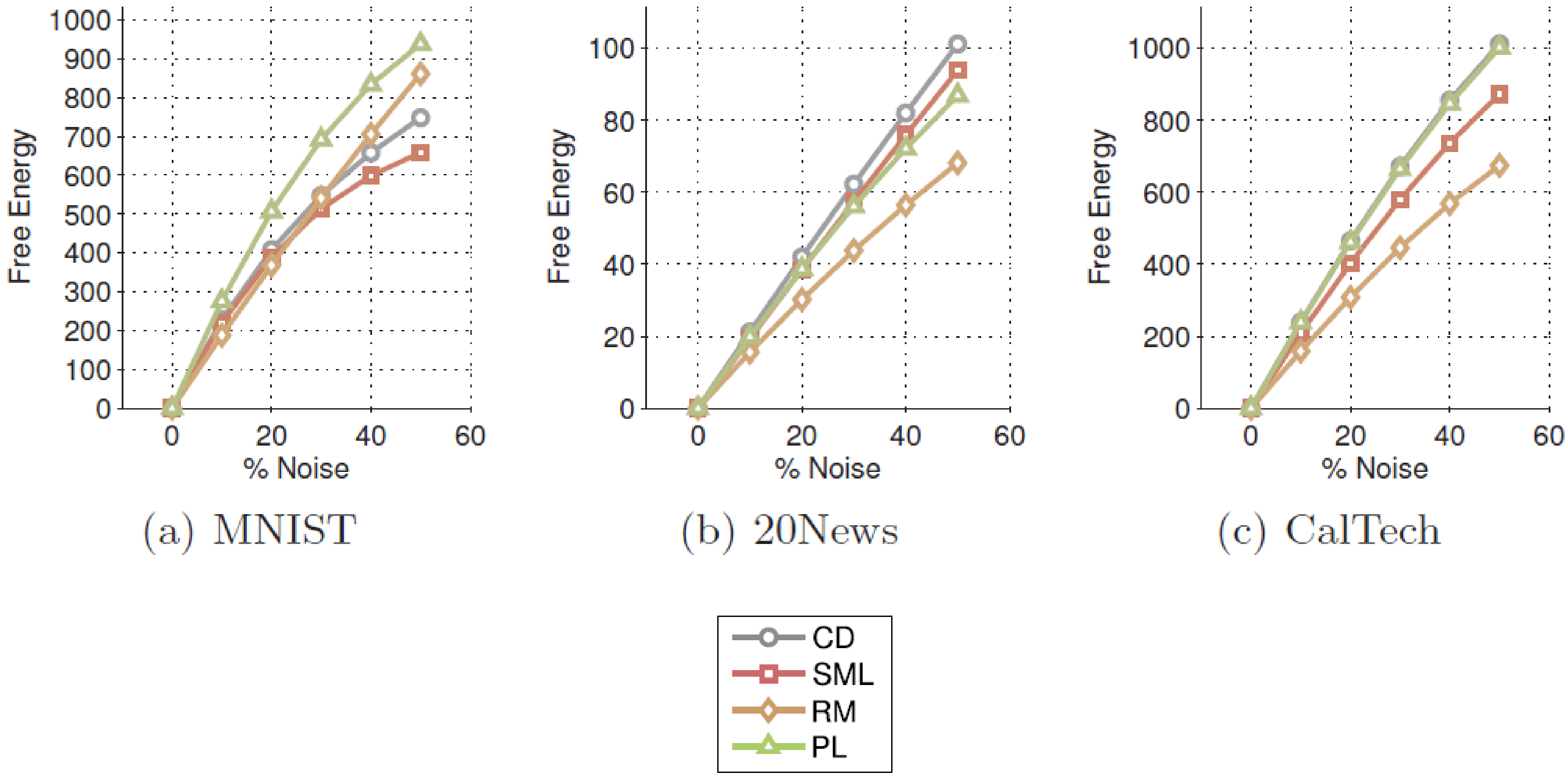
(a) MNIST (b) 20News (c) CalTech

Experiments: De-noising



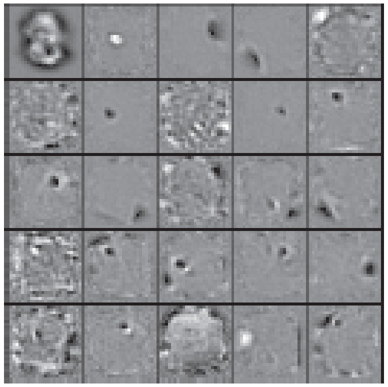
(a) MNIST (b) 20News (c) CalTech

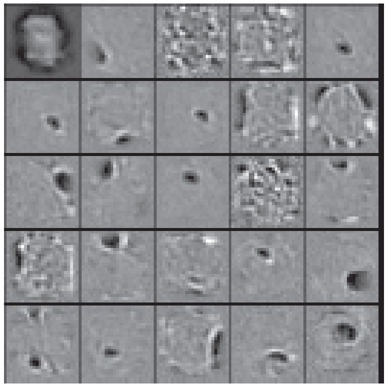
Experiments: Novelty Detection

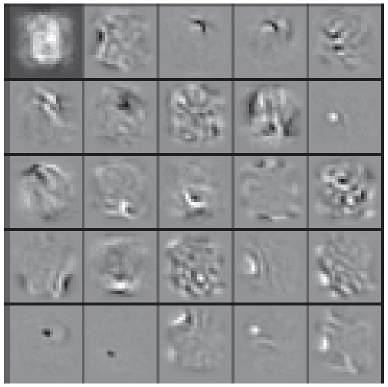


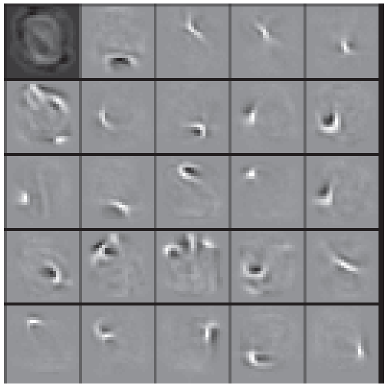
(a) MNIST (b) 20News (c) CalTech

Experiments: Learned Weights on MNIST


(a) CD


(b) SML


(c) PL


(d) RM

Discussion:

• As the underlying theory suggests, SML obtains the best test set log likelihoods.

• CD and SML perform very well on MNIST classification, which is consistent with past results.

• Ratio matching obtains the best de-noising results, which is consistent with the observation on MNIST that the filters have more local structure than those produced by SML, CD, and PL.

• PL and RM are actually **slower** than CD and SML due to the need to consider all one-neighbors for each data case.