Inductive Principles for Restricted Boltzmann Machine Learning

Benjamin Marlin

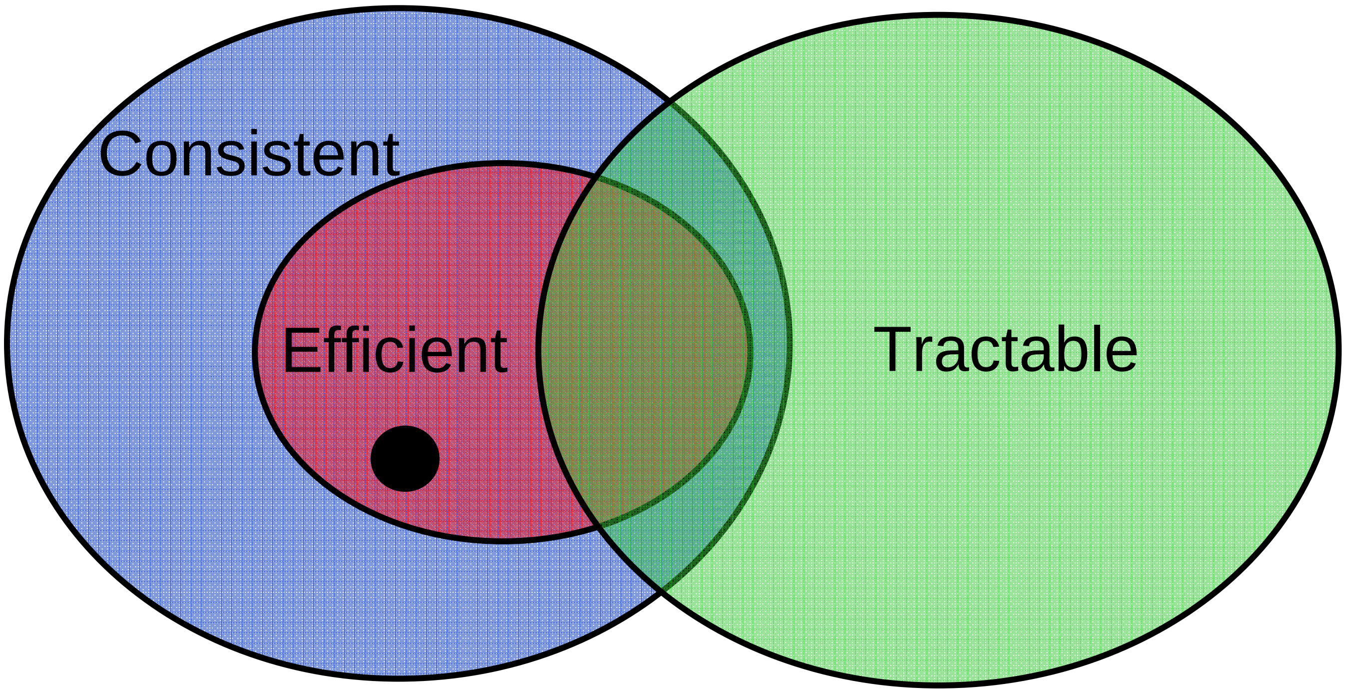
Department of Computer Science

University of British Columbia

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

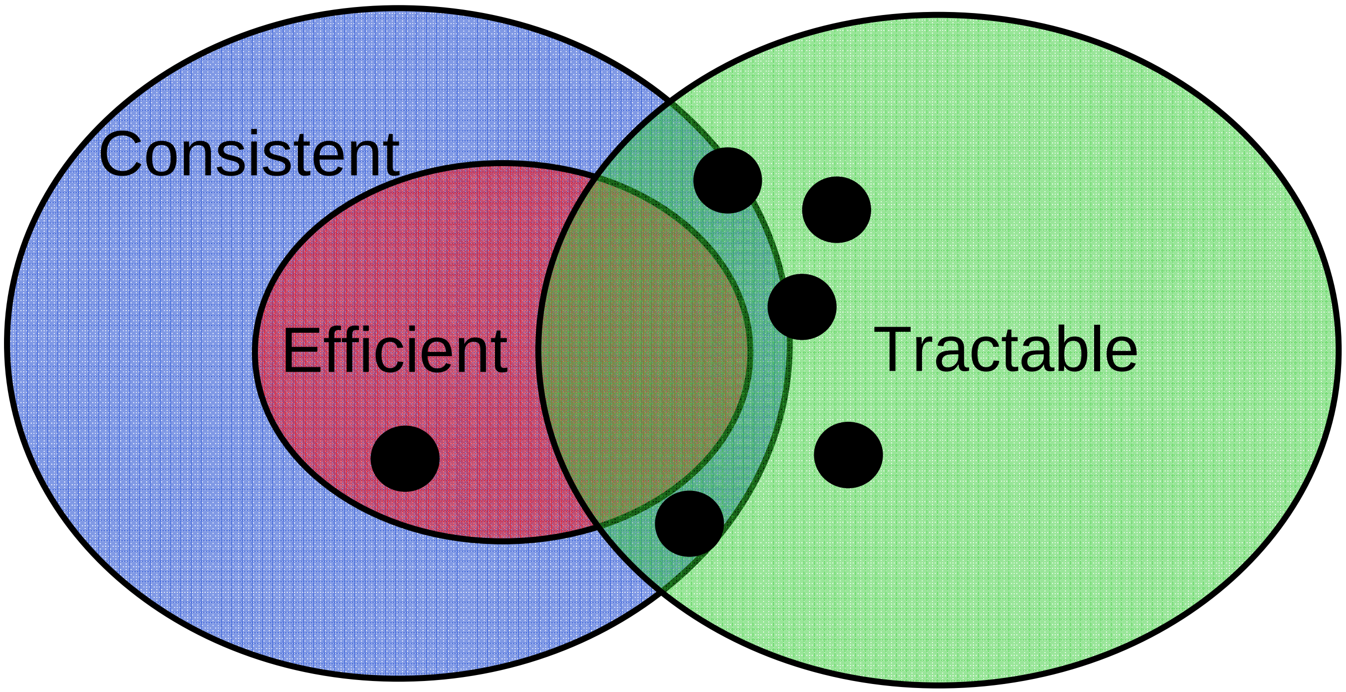
Introduction: Maximum Likelihood

- Maximum Likelihood Estimation is statistically consistent and efficient but is not computationally tractable for many models of interest like RBM's, MRF's, CRF's due to the partition function.



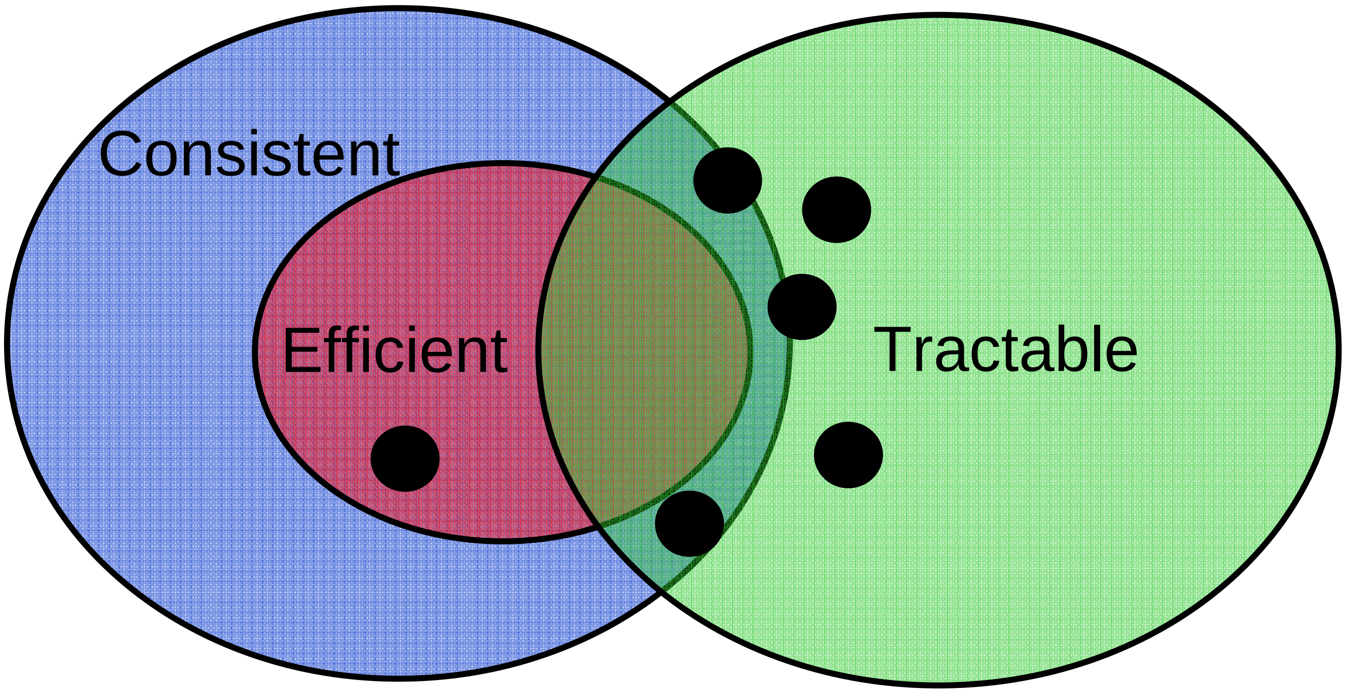
Introduction: Alternative Estimators

• Recent work has seen the proposal of many new estimators that trade consistency/efficiency for computational tractability including RM, SM, GSM, MPF, NCE, NLCE.



Introduction: Alternative Estimators

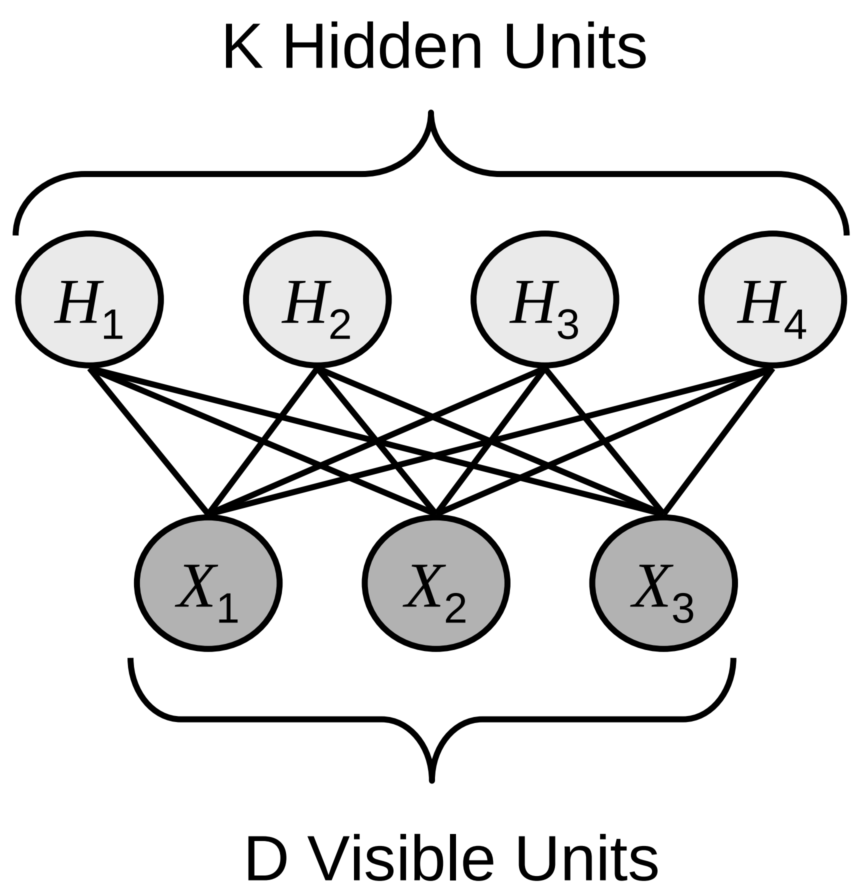
• Our main interest is uncovering the relationships between these estimators and studying their theoretical and empirical properties.



Outline:

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

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.

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

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

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

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.

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

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

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

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.

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

Outline:

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

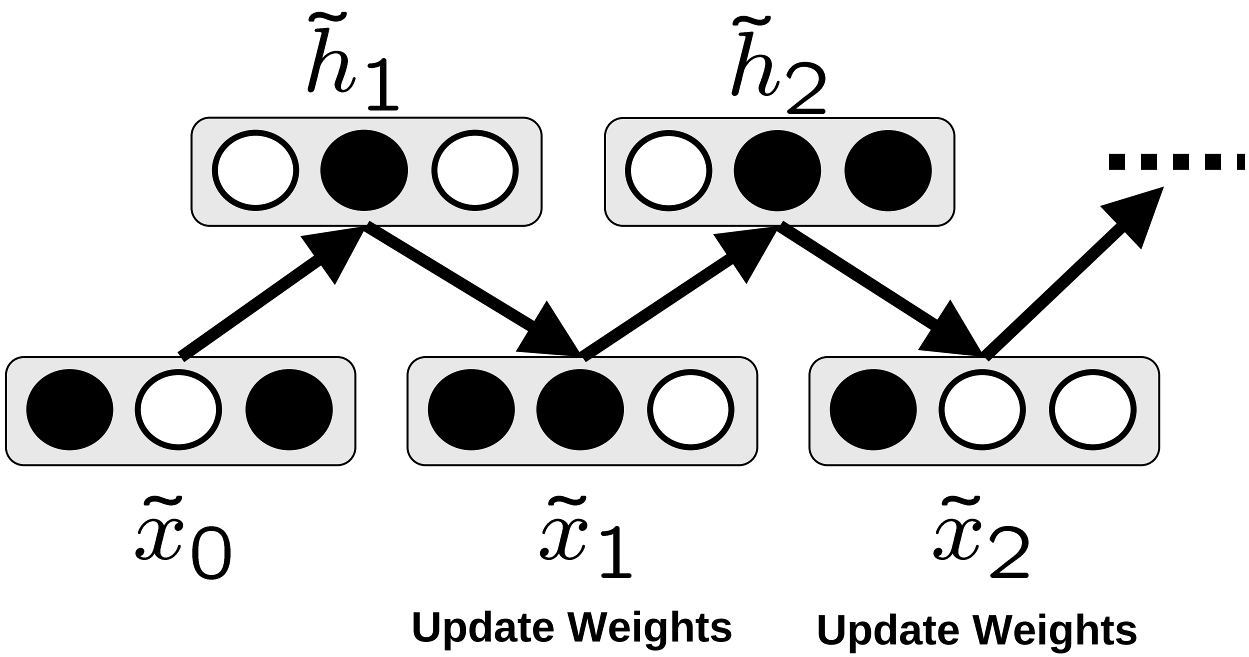
Stochastic Maximum Likelihood

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

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

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.



•L. Younes. Parametric inference for imperfectly observed Gibbsian fields. Prob. Th. and Related Fields, 82(4):625–645, 1989.
•T. Tieleman. Training restricted Boltzmann machines using approximations to the likelihood gradient. ICML 25, 2008.

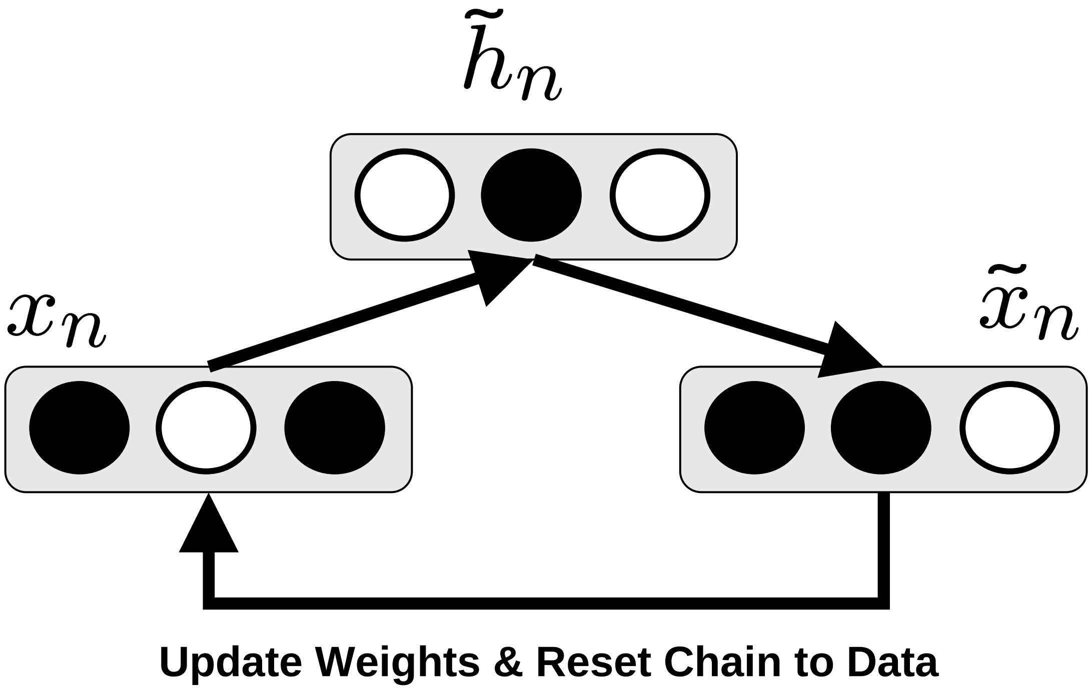
Contrastive Divergence

- 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.

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

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.



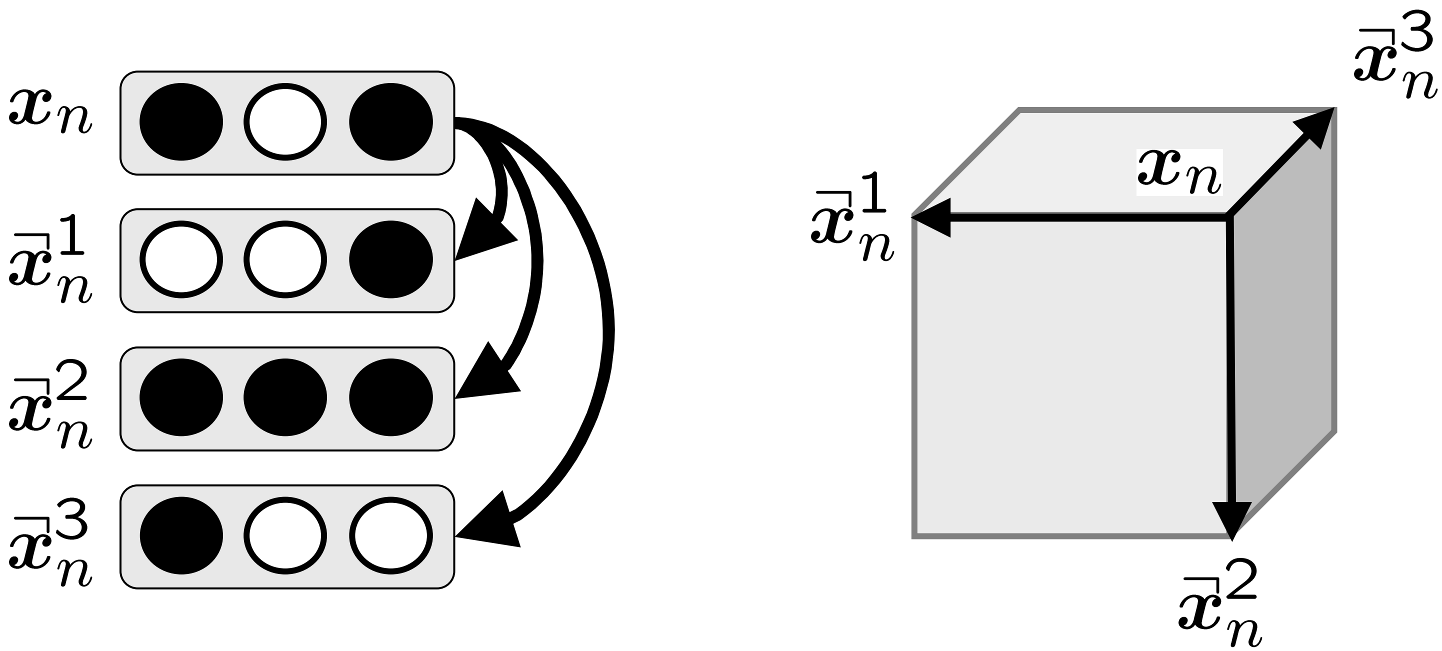
Pseudo-Likelihood

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

$$
\begin{aligned}
f^{PL}(\theta) &= \sum_{\boldsymbol{x}\in\mathcal{X}} \sum_{d=1}^{D} P_e(\boldsymbol{x}) \log P_\theta(x_d|\boldsymbol{x}_{-d}) \\
&= \frac{1}{N}\sum_{n,d} g_{PL}(r_{dn}) \\
g_{PL}(r) &= -\log(1+r^{-1}) \\
r_{dn} &= P_\theta(\boldsymbol{x}_n)/P_\theta(\bar{\boldsymbol{x}}_n^d)
\end{aligned}
$$

Pseudo-Likelihood

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



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.

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

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{d=1}^{D} g_{RM}(r_{dn})$$

$$g_{RM}(r) = (1+r)^{-2}$$

Aapo Hyvarinen. Some extensions of score matching. Computational Statistics & Data Analysis, 51(5):2499–2512, 2007.

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.

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

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{d=1}^{D} g_{GSM}(r_{dn})$$

$$g_{GSM}(r) = r^{-2} - 2r$$

Siwei Lyu. Interpretation and generalization of score matching. In Uncertainty in Artificial Intelligence 25, 2009.

Minimum Probability Flow

- Minimize the flow of probability from data states to non-data states (as we've just seen!).

$$f^{MPF}(\theta) = \sum_{\boldsymbol{x} \in \mathcal{X}} \sum_{d=1}^{D} P_e(\boldsymbol{x}) \log P_\theta^{(\epsilon)}(\boldsymbol{x})$$

$$\approx \frac{1}{N} \sum_{n=1}^{N} \sum_{d=1}^{D} I[P_e(\bar{\boldsymbol{x}}_n^d) = 0] g_{MPF}(r_{dn})$$

$$g_{MPF}(r) = r^{-1/2}$$

Jascha Sohl-Dickstein, Peter Battaglino, Michael R. DeWeese. Minimum Probability Flow Learning.

Comparison: Gradients

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

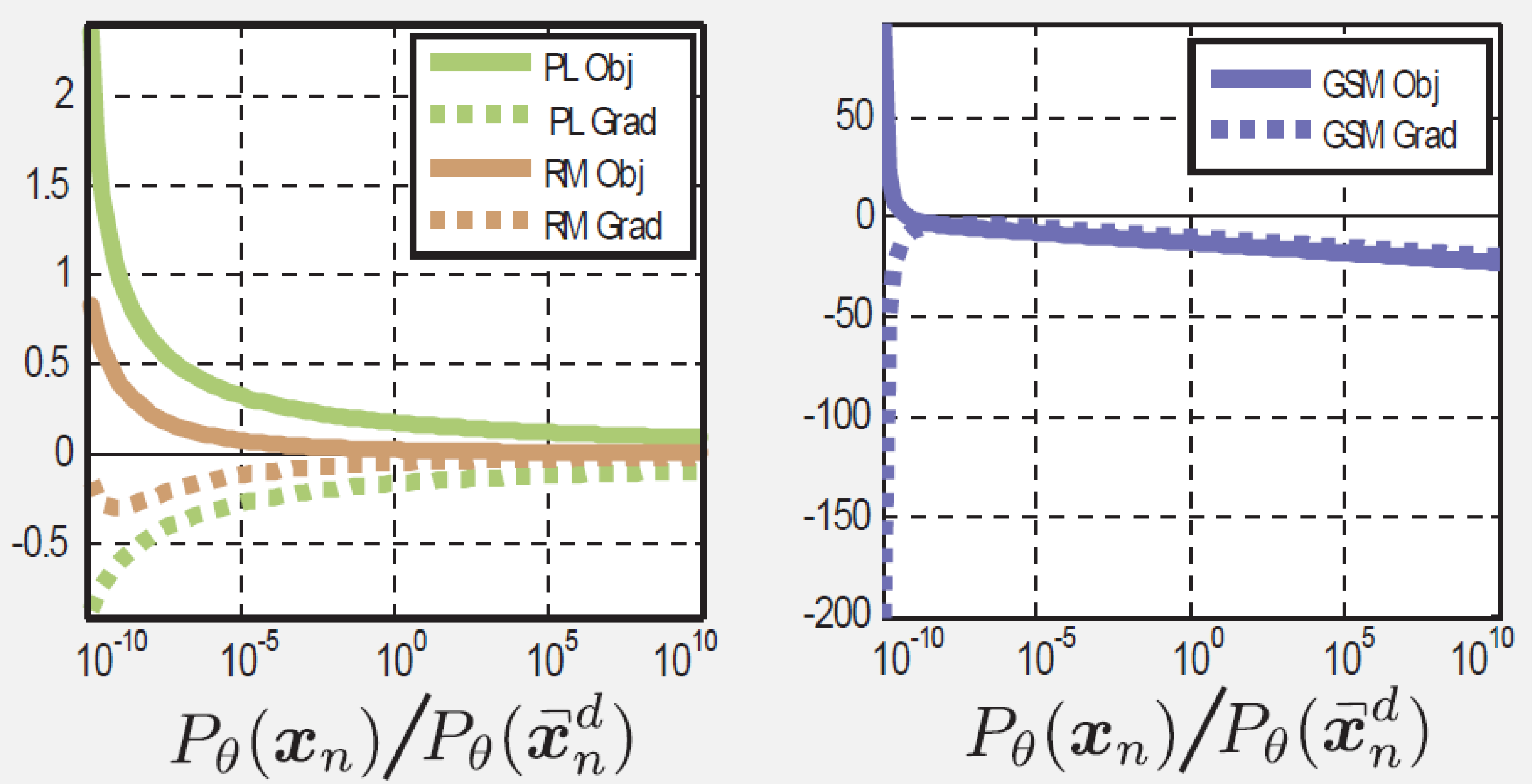
$$\nabla f^{CD} \approx -\frac{1}{N}\sum_{n=1}^{N}\left(\nabla F_\theta(\boldsymbol{x}_n) - \nabla F_\theta(\tilde{\boldsymbol{x}}_n)\right)$$

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

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

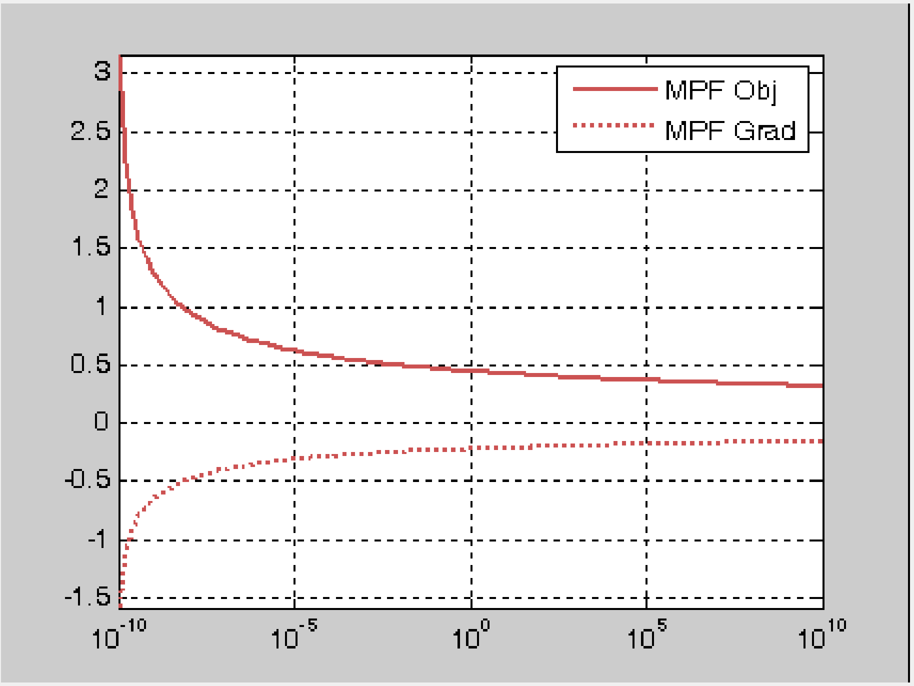
$$\nabla f^{GSM} = \frac{2}{N}\sum_{n=1}^{N}\sum_{d=1}^{D} g'_{GSM}(r_{dn}) r_{dn}\left(\nabla F_\theta(\boldsymbol{x}_n) - \nabla F_\theta(\bar{\boldsymbol{x}}_n^d)\right)$$

Comparison: Weighting Functions



Comparison: Weighting Functions

What about MPF?



$$P_\theta(\boldsymbol{x}_n) / P_\theta(\bar{\boldsymbol{x}}_n^d)$$

Comparison: A Manifold of Estimators?

Dimensions:

1. Neighborhood structure around data configurations.

2. Form of loss function on the probability ratio.
   - Smooth
   - Monotonically decreasing
   - Bounded below
   - Others?

Covers: PL, GLP, NLCE, RM, GSM, MPF

Limitations: No good for missing data/explicit latent variables.

Outline:

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

Experiments:

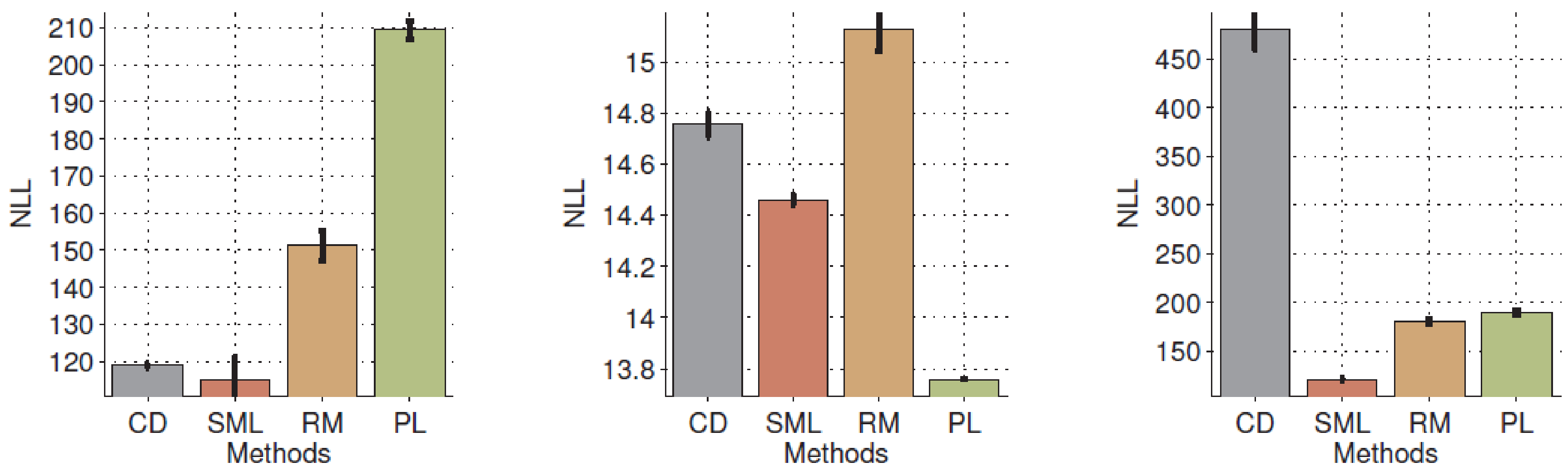
Data Sets:

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

Evaluation Criteria:

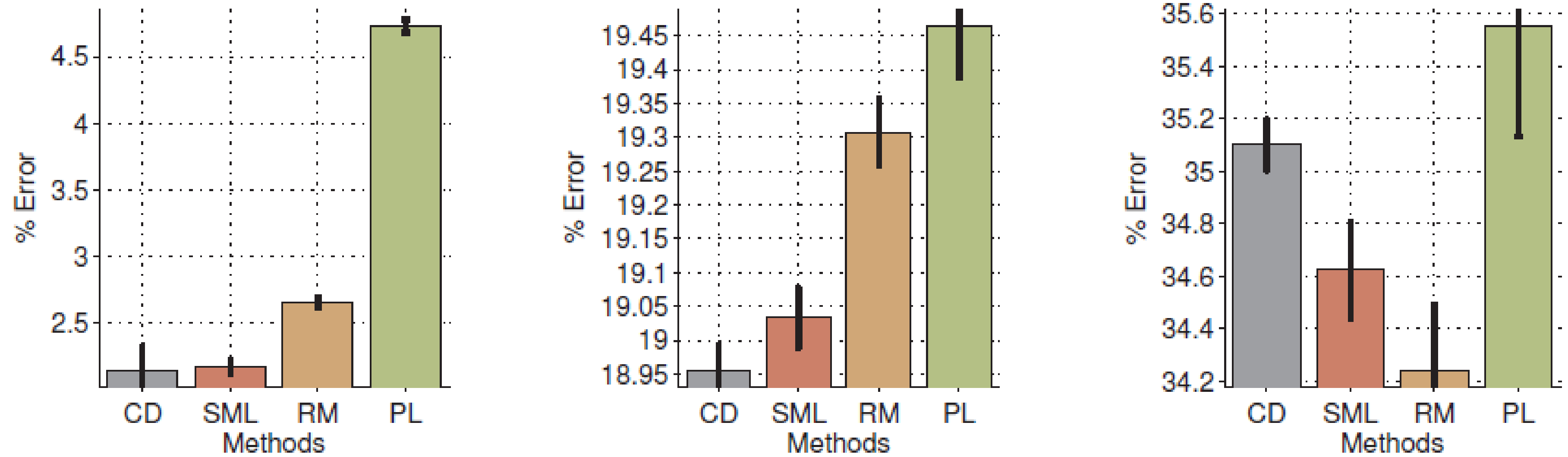
- Log likelihood (using AIS estimator)
- Classification error
- Reconstruction error
- De-noising
- Novelty detection

Experiments: Log Likelihood



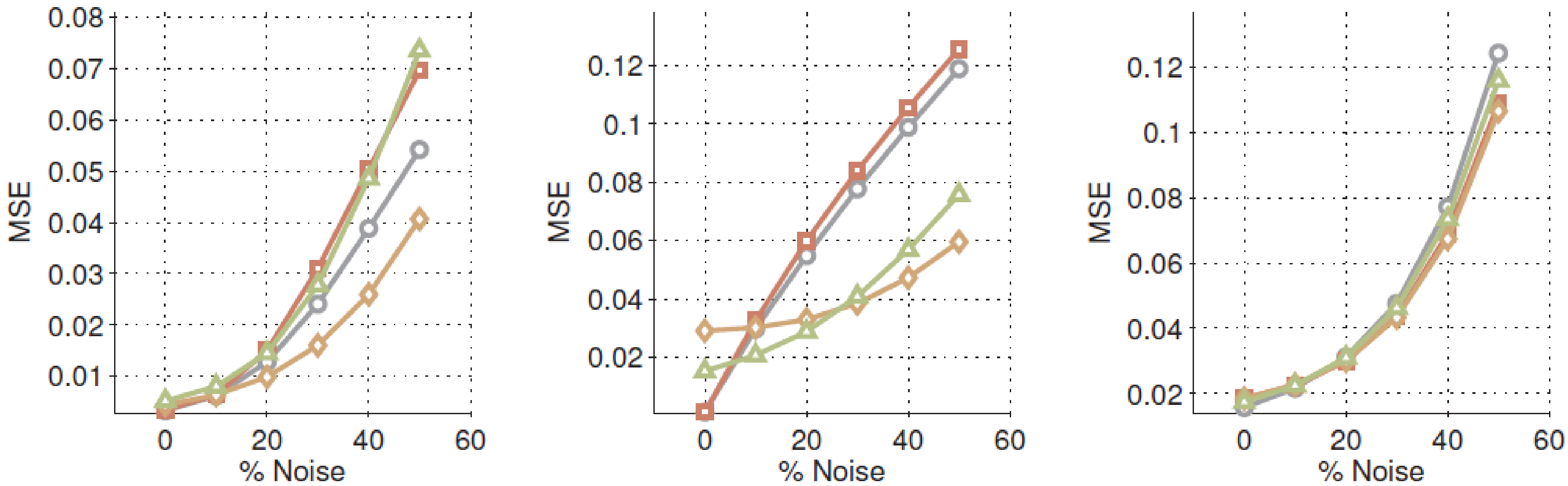
(a) MNIST (b) 20News (c) CalTech

Experiments: Classification Error

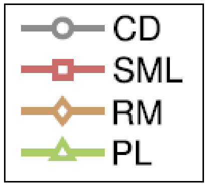


(a) MNIST (b) 20News (c) CalTech

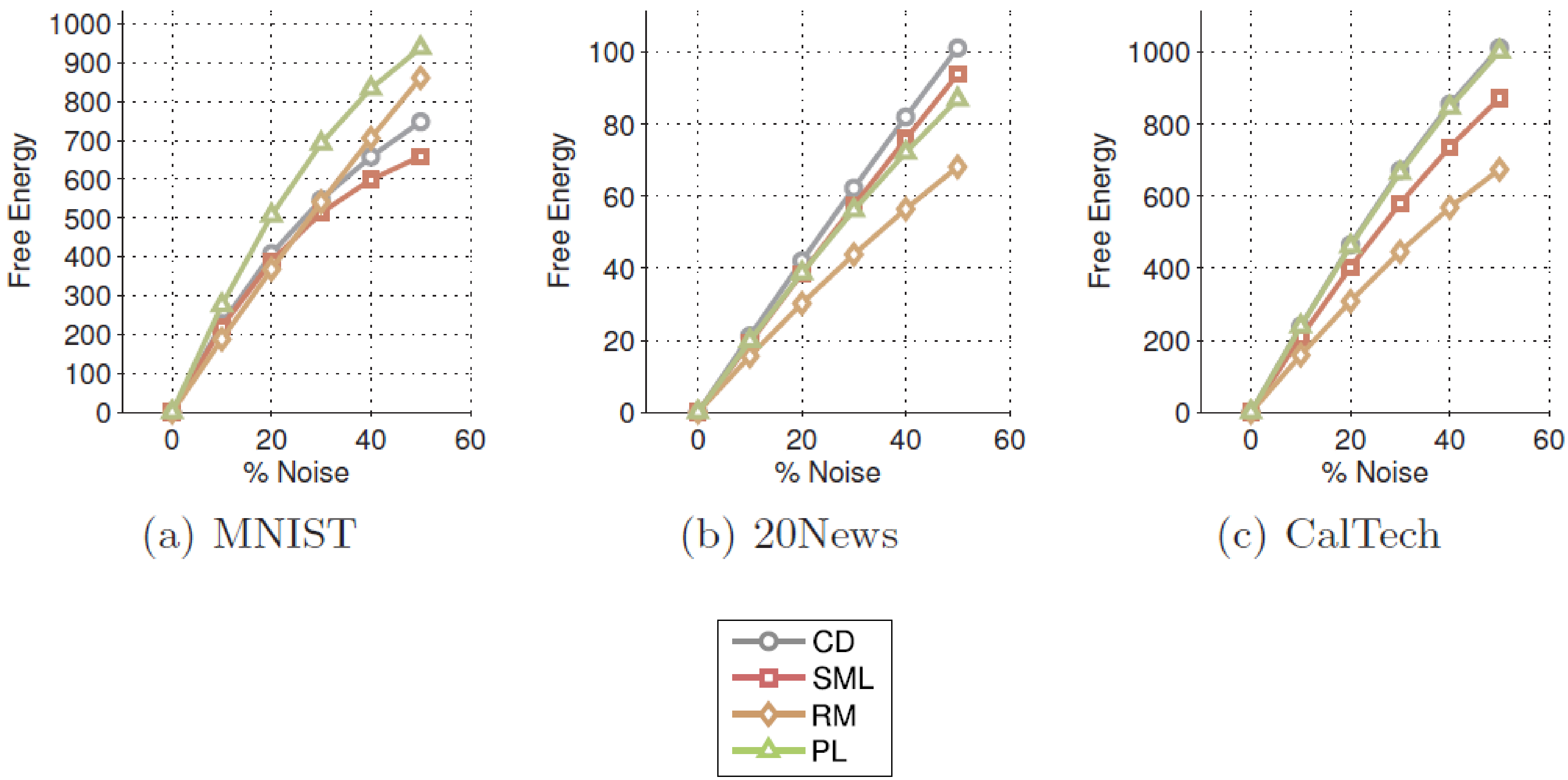
Experiments: De-noising



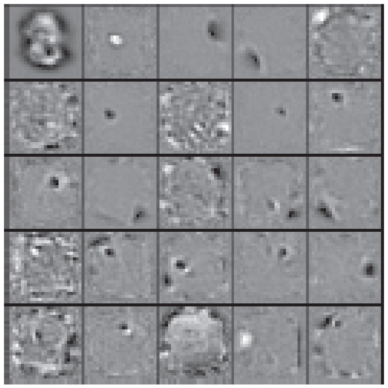
(a) MNIST (b) 20News (c) CalTech

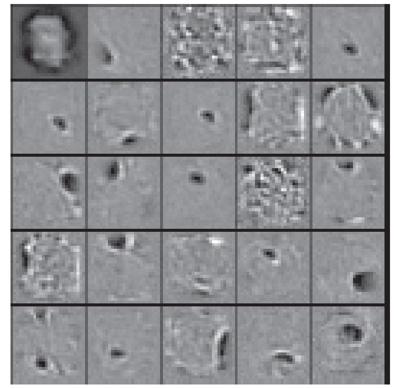


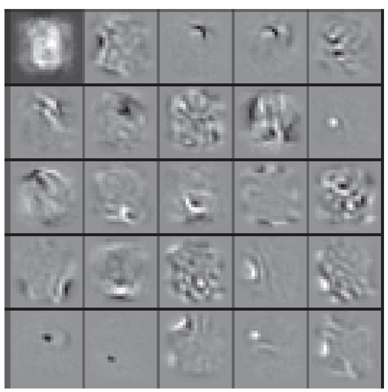
Experiments: Novelty Detection

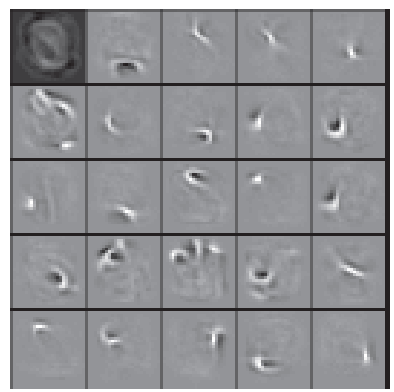


Experiments: Learned Weights on MNIST


(a) CD


(b) SML


(c) PL


(d) RM