Inductive Principles for Restricted Boltzmann Machine Learning

Benjamin M. Marlin, Kevin Swersky, Bo Chen and Nando de Freitas

Department of Computer Science, University of British Columbia, Vancouver, Canada

1.0 Introduction

- The prevalence of maximum likelihood as an inductive principle is based on two key properties: asymptotic consistency and asymptotic efficiency.

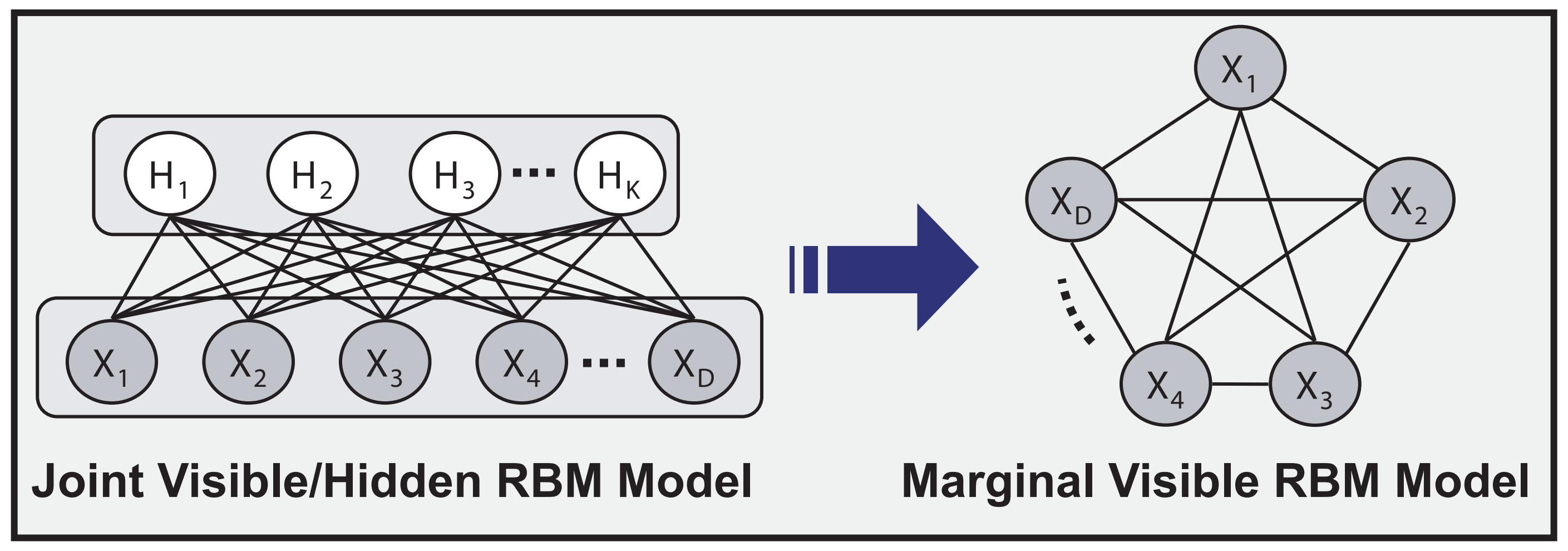
- For a large class of models including conditional random fields, Markov random fields, and restricted Boltzmann machines, simply computing exact likelihoods can be computationally intractable.

- There are two basic approaches to dealing with intractable likelihoods: (1) Approximately maximize the likelihood. (2) Select an alternative inductive principle.

- In this work, we present a study of several alternative inductive principles for learning restricted Boltzmann machines including contrastive divergence (CD), pseudo-likelihood (PL), ratio matching (RM), and generalized score matching (GSM). We compare to stochastic maximum likelihood estimation (SML).

2.0 RBM's

A restricted Boltzmann machine is a Markov random field with two layers of nodes called the "visible layer" and the "hidden layer". The visible layer consists of D binary data variables X, while the hidden layer consists of K binary latent variables H. The graph is bipartite with connections between the visible and hidden layers only. The binary hidden variables can be analytically integrated away yielding a marginal distribution on the visible variables only.



Joint Visible/Hidden RBM Model

Marginal Visible RBM Model

- Energy: $E_\theta(\boldsymbol{x},\boldsymbol{h}) = -(\boldsymbol{x}^T W\boldsymbol{h} + \boldsymbol{x}^T b + \boldsymbol{h}^T c)$
- Joint: $P_\theta(\boldsymbol{x},\boldsymbol{h}) = \frac{1}{\mathcal{Z}}\exp\left(-E_\theta(\boldsymbol{x},\boldsymbol{h})\right)$
- Conditional: $P_\theta(x_d = 1|\boldsymbol{h}) = \frac{1}{1+\exp(-(\sum_{k=1}^{K} W_{dk}h_k + b_d))}$

  $P_\theta(h_k = 1|\boldsymbol{x}) = \frac{1}{1+\exp(-(\sum_{d=1}^{D} W_{dk}x_d + c_k))}$
- Free Energy: $F_\theta(\boldsymbol{x}) = -\left(\boldsymbol{x}^T b + \sum_{k=1}^{K}\log\left(1+\exp\left(\boldsymbol{x}^T W_k + c_k\right)\right)\right)$
- Marginal: $P_\theta(\boldsymbol{x}) = \frac{1}{\mathcal{Z}}\exp\left(-F_\theta(\boldsymbol{x})\right) \qquad \mathcal{Z} = \sum_{\boldsymbol{x}'\in\mathcal{X}}\exp\left(-F_\theta(\boldsymbol{x}')\right)$

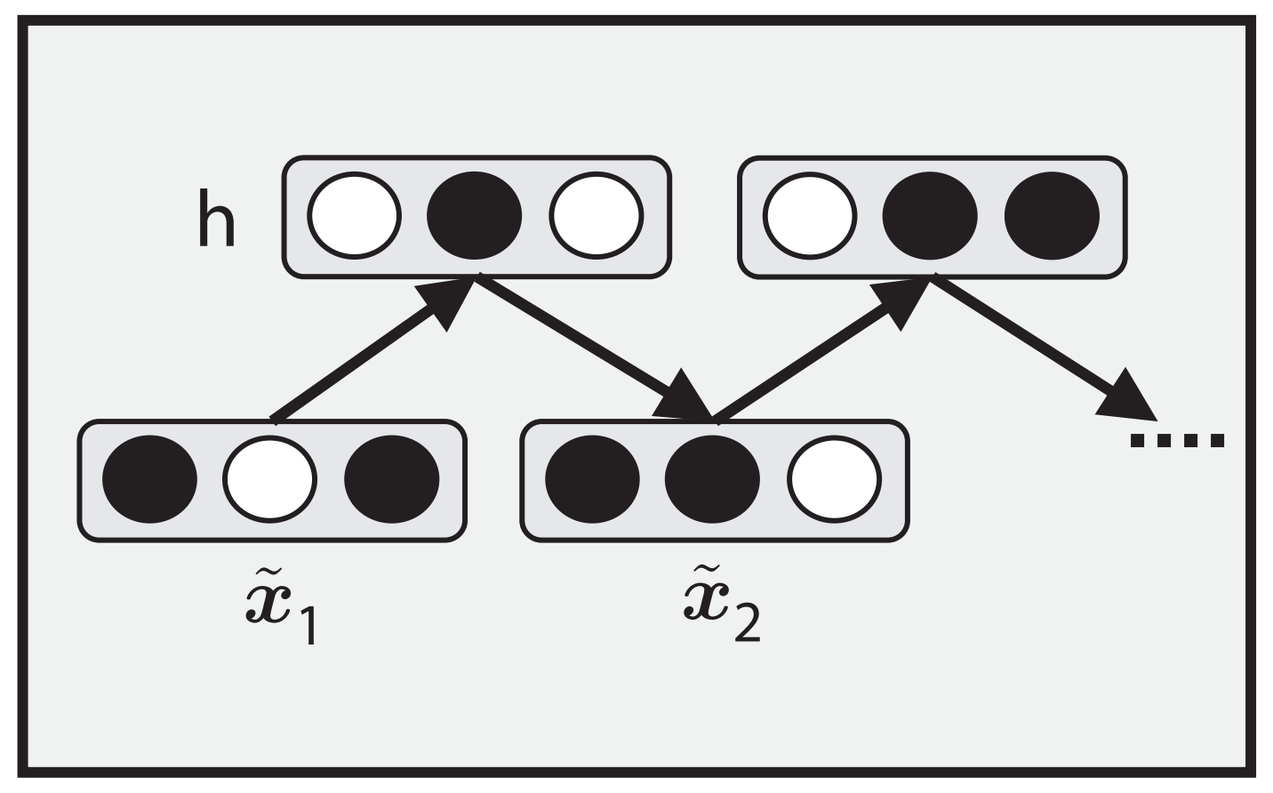
3.0 Inductive Principles

Maximum Likelihood: Maximize the likelihood of the data given the parameters. Equivalently, minimize the KL divergence from the data distribution to the model distribution. Exact ML is intractable so stochastic approximations are often used instead.

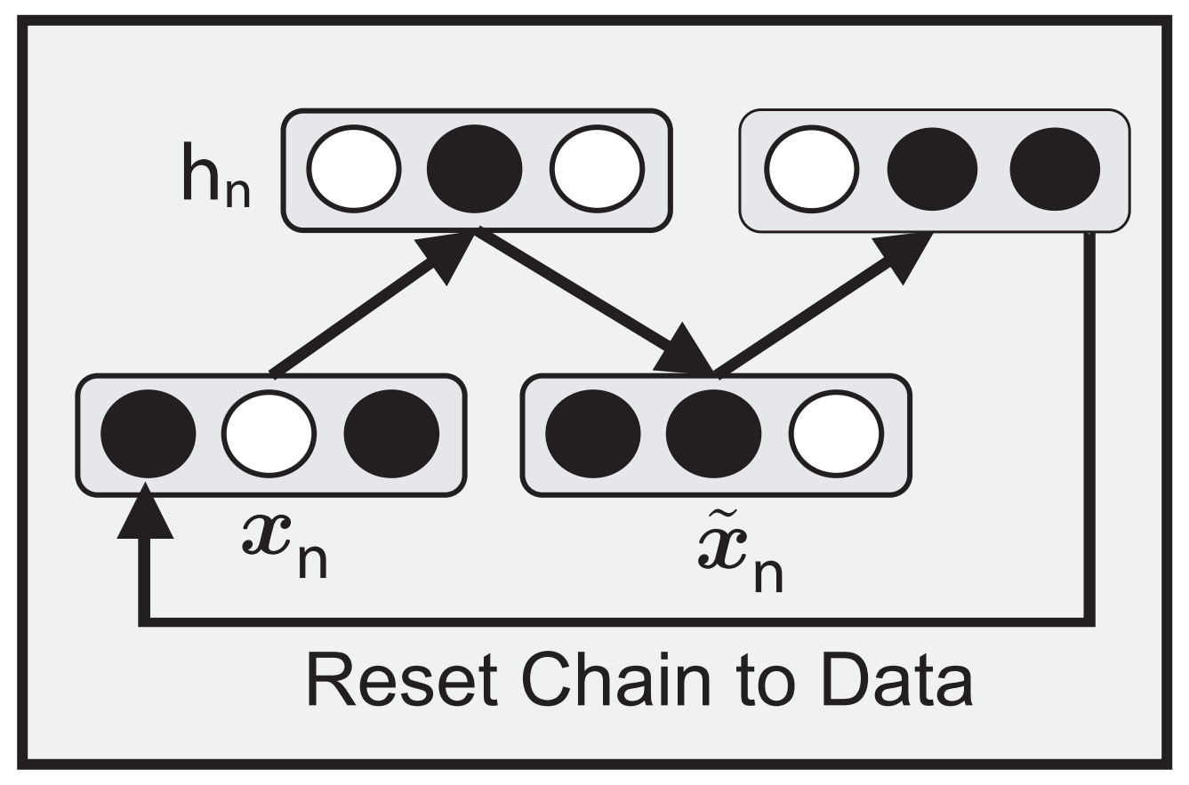
- Objective: $f^{ML}(\theta) = \sum_{\boldsymbol{x}\in\mathcal{X}} P_e(\boldsymbol{x})\log P_\theta(\boldsymbol{x})$
- Gradient: $\nabla f^{ML} \approx -\left(\frac{1}{N}\sum_{n=1}^{N}\nabla F_\theta(\boldsymbol{x}_n) - \frac{1}{S}\sum_{s=1}^{S}\nabla F_\theta(\tilde{\boldsymbol{x}}_s)\right)$

Contrastive Divergence: Minimize the difference in KL divergence from the data to the model distribution and the T-step Gibbs distribution to the model distribution.

- Objective: $f^{CD}(\theta) = \sum_{\boldsymbol{x}\in\mathcal{X}} P_e(\boldsymbol{x})\log\left(\frac{P_e(\boldsymbol{x})}{P_\theta(\boldsymbol{x})}\right) - Q_\theta^t(\boldsymbol{x})\log\left(\frac{Q_\theta^t(\boldsymbol{x})}{P_\theta(\boldsymbol{x})}\right)$
- Gradient: $\nabla f^{CD} \approx -\frac{1}{N}\left(\sum_{n=1}^{N}\nabla F_\theta(\boldsymbol{x}_n) - \nabla F_\theta(\tilde{\boldsymbol{x}}_n)\right)$



Stochastic ML Sampling



CD Sampling

Pseudo-Likelihood, Ratio Matching, & Generalized Score Matching: All three inductive principles can be cast in terms of one-dimensional conditional probabilities or ratios of probabilities of one-neighbours. The gradients differ only by a weighting function.

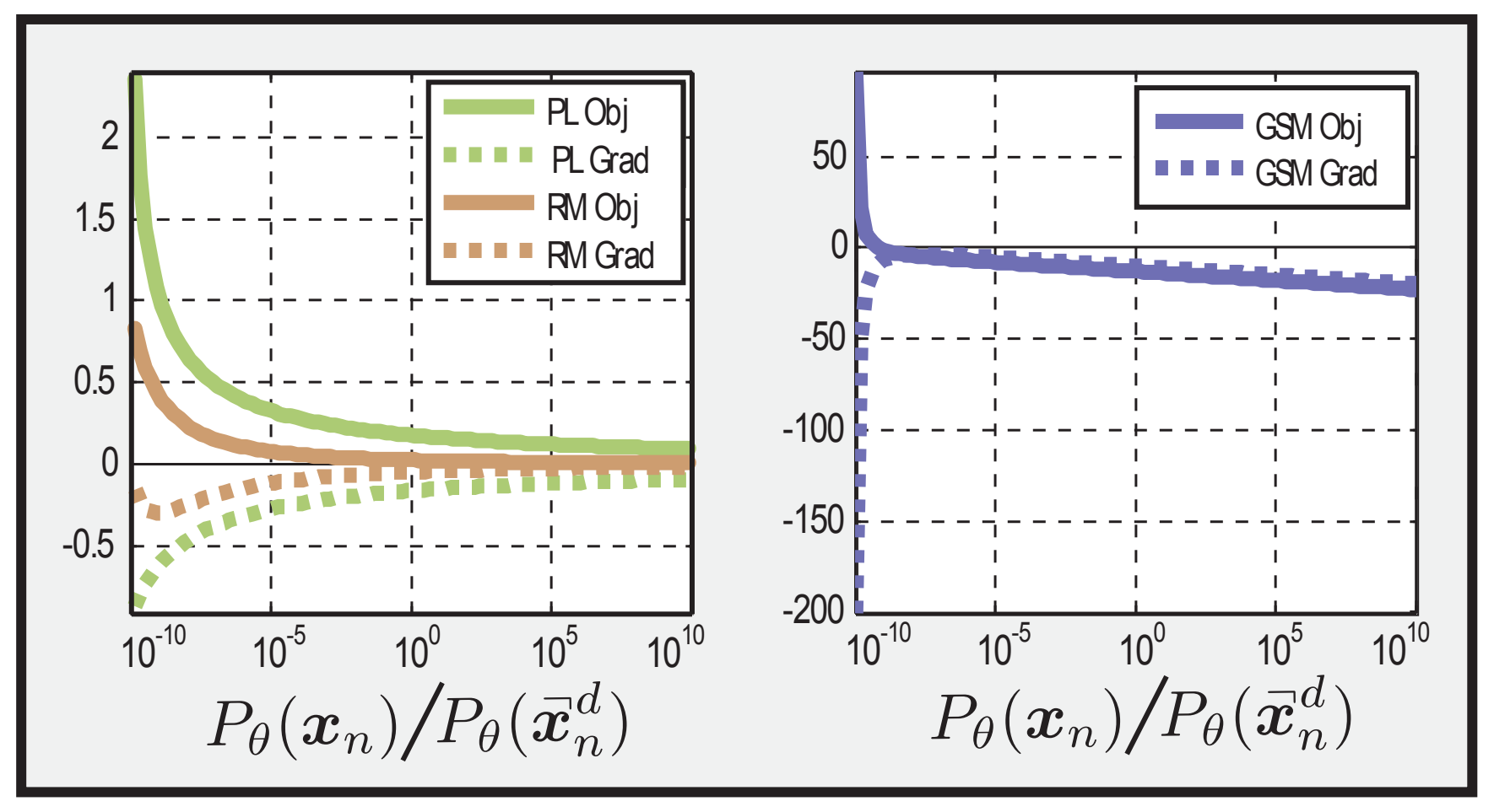
- Objective: $f^{PL}(\theta) = \sum_{\boldsymbol{x}\in\mathcal{X}}\sum_{d=1}^{D} P_e(\boldsymbol{x})\log P_\theta(x_d|\boldsymbol{x}_{-d})$

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

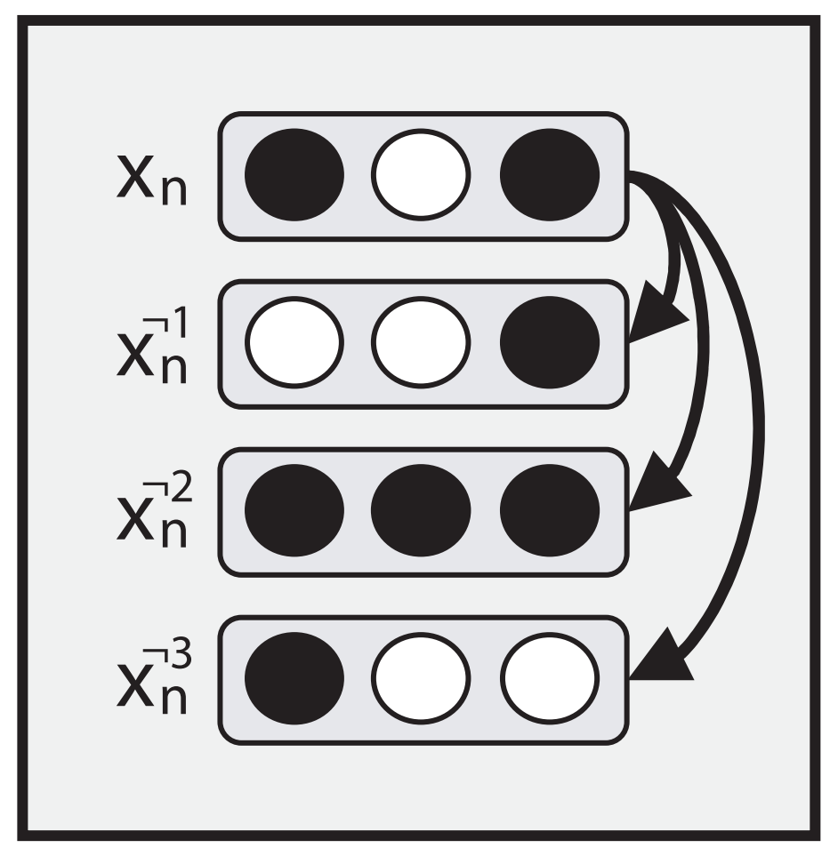
$$f^{GSM}(\theta) = \sum_{\boldsymbol{x}\in\mathcal{X}}\sum_{d=1}^{D} P_e(\boldsymbol{x})\Big(\frac{1}{P_\theta(x_d|\boldsymbol{x}_{-d})} - \frac{1}{P_e(x_d|\boldsymbol{x}_{-d})}\Big)^2$$

- Gradient: $\nabla f^* = \frac{1}{N}\sum_{n\;1}^{N}\sum_{d\;1}^{D} g^*\Big(\frac{P_\theta(\boldsymbol{x}_n)}{P_\theta(\bar{\boldsymbol{x}}^d)}\Big)(\nabla F_\theta(\boldsymbol{x}_n) - \nabla F_\theta(\bar{\boldsymbol{x}}_n^d))$

$$\nabla F_\theta(\boldsymbol{x}) = \{-\boldsymbol{x}E[\boldsymbol{h}|\boldsymbol{x}]^T,\ -\boldsymbol{x},\ -E[\boldsymbol{h}|\boldsymbol{x}]\}$$

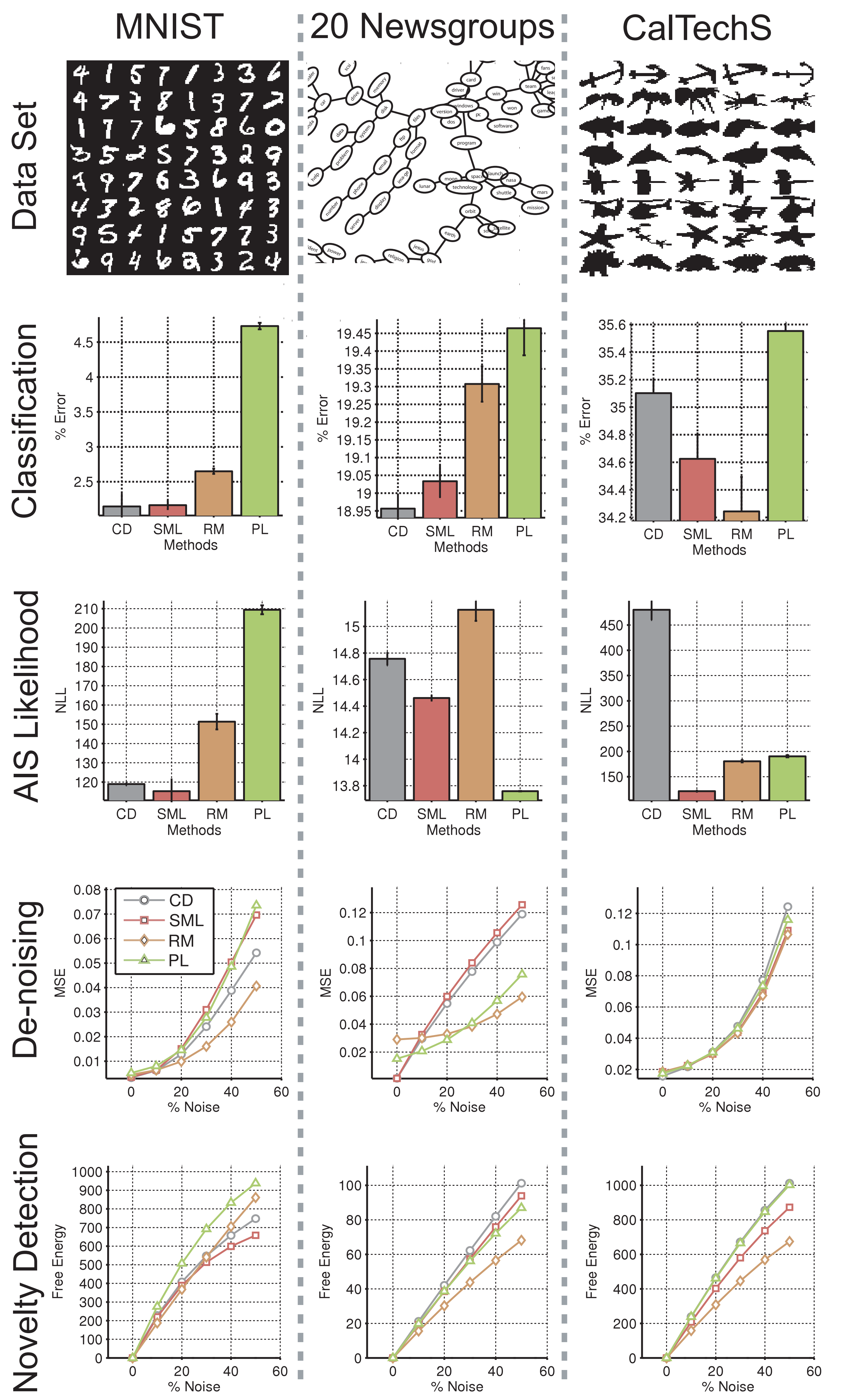


Weighting Functions



One-Neighbours

4.0 Experiments & Results



5.0 Conclusions

- The gradients of all the methods we consider differ only in the distribution of "fantasy" data and the choice of weighting function.

- These differences are meaningful as the methods exhibit very different theoretical and empirical characteristics.

- Careful implementations of ratio matching and pseudo likelihood are still an order of magnitude slower than SML and CD due to considering all possible one-neighbours for each training case.

- Taking computation time into account, SML is the most attractive method.