Unbiased Contrastive Divergence Algorithm for Training Energy-Based Latent Variable Models

Yixuan Qiu

Department of Statistics and Data Science, Carnegie Mellon University

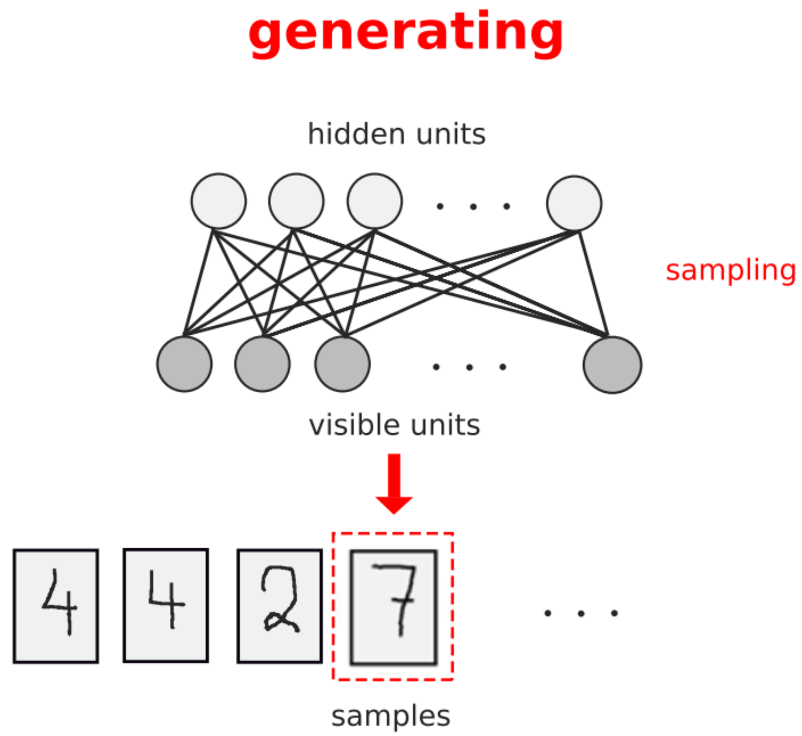
ICLR

Joint work with Prof. Lingsong Zhang & Prof. Xiao Wang

Department of Statistics, Purdue University

Motivation

1. Energy-based models (EBM) are widely used in deep generative models, e.g. restricted Boltzmann machines (RBM)
2. They are typically trained using the contrastive divergence (CD, Hinton, 2002) algorithm
3. But many papers have given examples that CD may diverge
4. The cause is the use of a biased estimator of the gradient
5. We fix it by using an unbiased MCMC estimator



RBM: Fischer and Igel (2004)

Diagnosis of CD

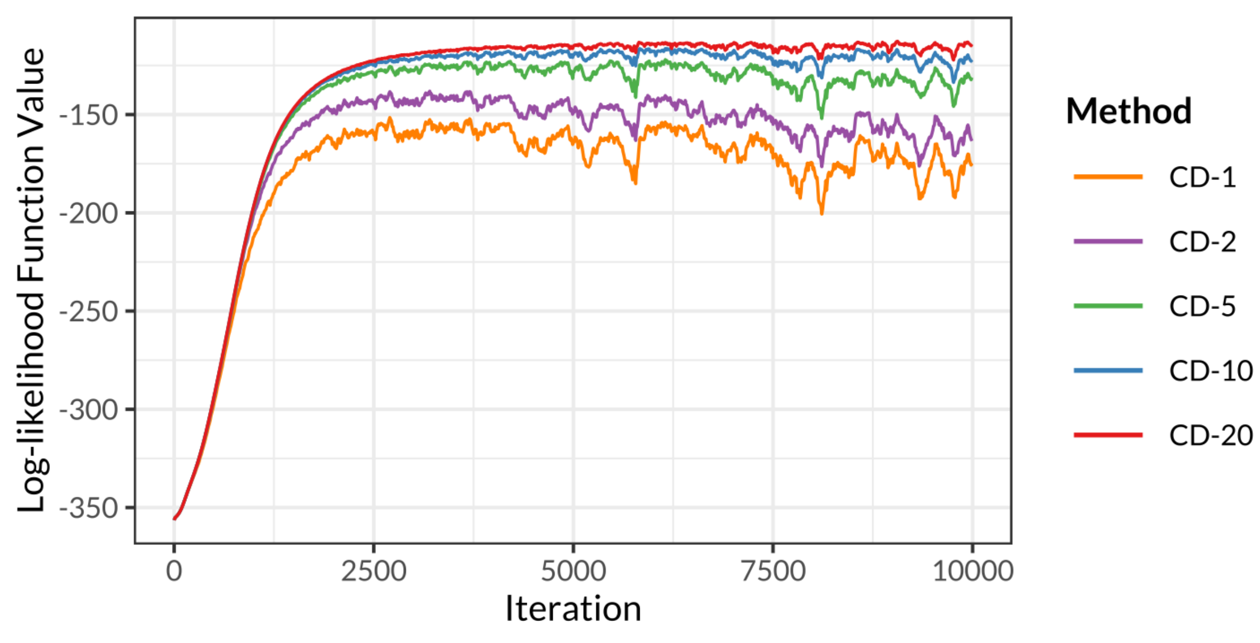
- For RBM, the gradient of log-likelihood function has a nice form

$$\frac{\partial \ell(\theta)}{\partial w_{ij}} = \underbrace{\mathbf{E}_{\text{data}}(v_i h_j)}_{\text{Simple}} - \underbrace{\mathbf{E}_{\text{model}}(v_i h_j)}_{\text{CD approximates it by running a } k\text{-step MCMC}}$$

Simple

CD approximates it by running a k-step MCMC

- CD gradient is a biased estimator for the true one
- Errors will accumulate during the training process
- This is a consequence of using a finite-step MCMC to approximate the limit

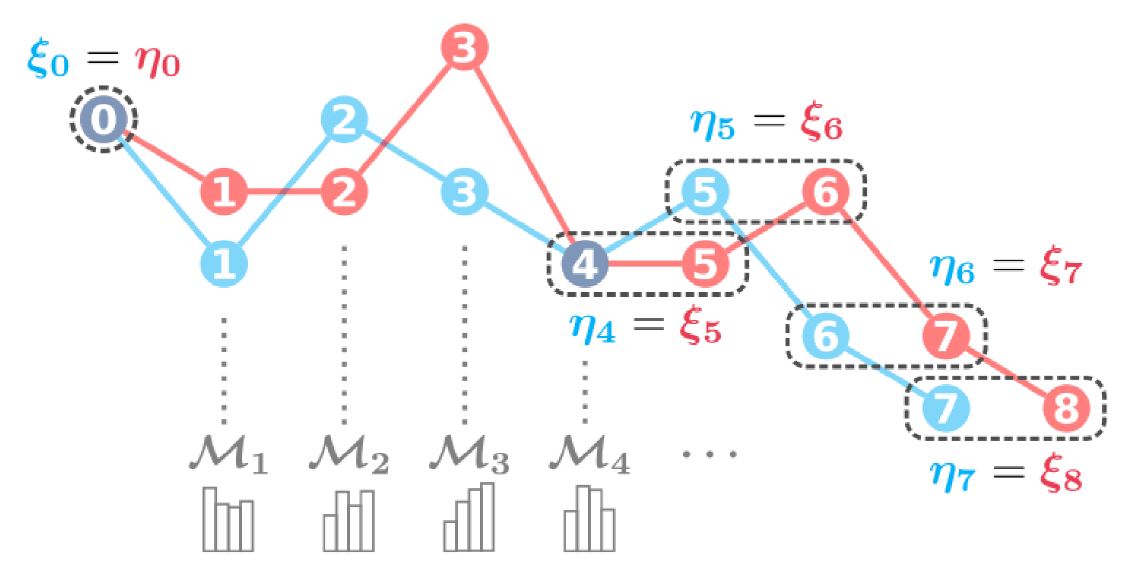


The Unbiased MCMC Estimator

- A relatively new topic in statistics and machine learning community
- Two seminal works: Glynn & Rhee (2014) and Jacob et al. (2017)
- Want to estimate $\mu = \lim_{k \to \infty} \mathbf{E}(X_k)$, but every X_k is biased
- Write μ as a telescoping sum $\mu = \mathbf{E}(X_k) + \sum_{t=k+1}^{\infty} \{\mathbf{E}(X_t) - \mathbf{E}(X_{t-1})\}$
- If we have another sequence Y_k such that:
    1. X_k and Y_k have the same marginal distribution
    2. $X_t = Y_{t-1}$ for all $t \geq \tau$, where τ is some random time
- Then $\mu = \mathbf{E}\{X_k + \sum_{t=k+1}^{\tau-1} (X_t - Y_{t-1})\}$
- Key idea: use two coupled chains to cancel the tail series

How to Find Y_k?

- Obviously, we cannot take $Y_k = X_k$ or let them be independent
- X_k and Y_k need to be correlated in such a way that:
  1. $\mathbf{P}(X_k = Y_{k-1}) > 0$
  2. They have identical marginal distributions
- Such a technique is called coupling
- We have developed:
  - Specialized algorithm for RBM
  - General method for other EBMs



Unbiased Contrastive Divergence (UCD)

- Replace the second term by the unbiased MCMC estimator

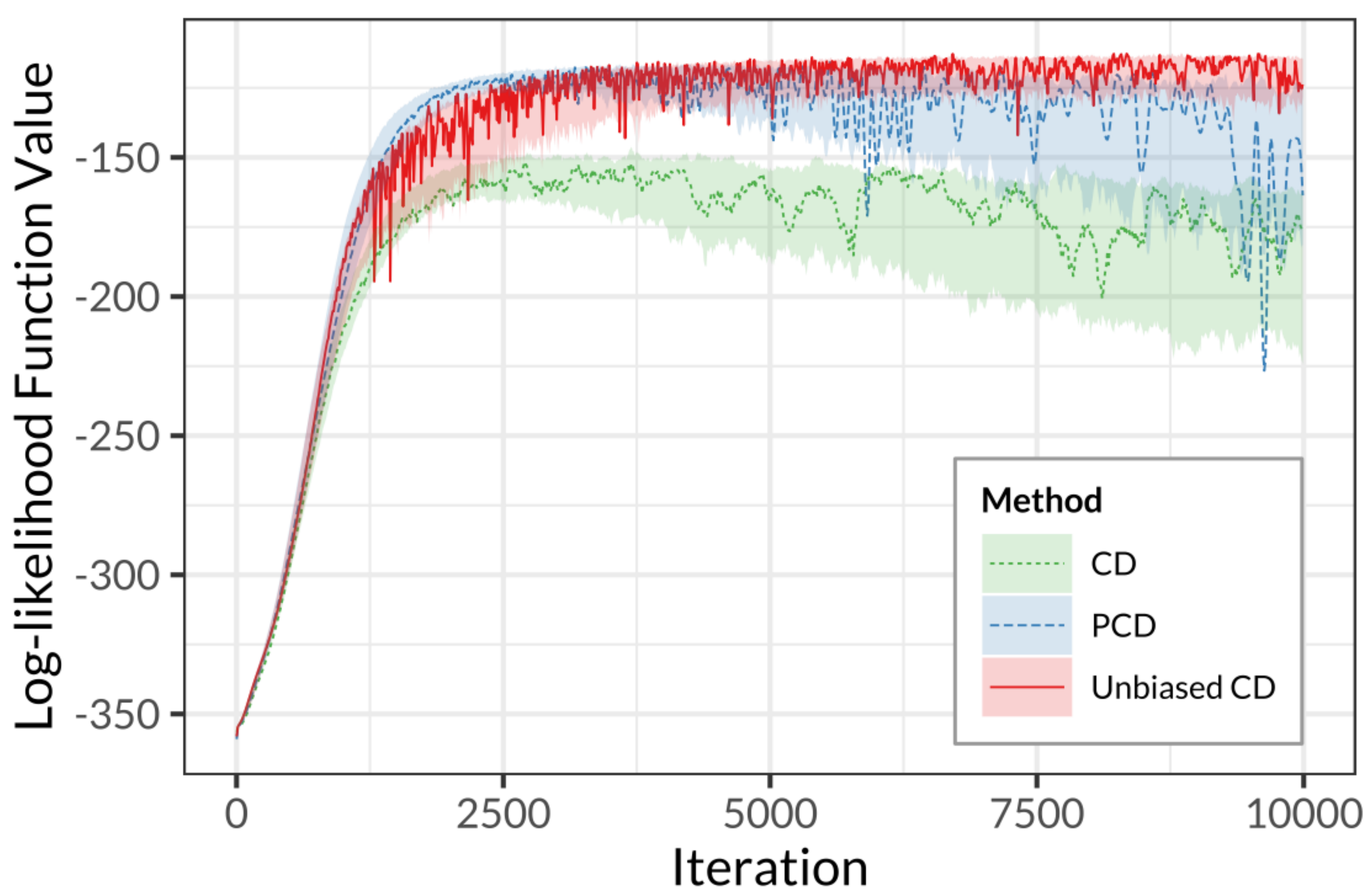
$$\frac{\partial \ell(\theta)}{\partial w_{ij}} = \mathbf{E}_{\text{data}}(v_i h_j) - \mathbf{E}_{\text{model}}(v_i h_j)$$

UCD constructs an unbiased estimator for the second term

- We develop theorems to show:
  1. The estimator has a finite variance → So we can apply SGD
  2. τ has a finite expectation → So we can compute in finite time

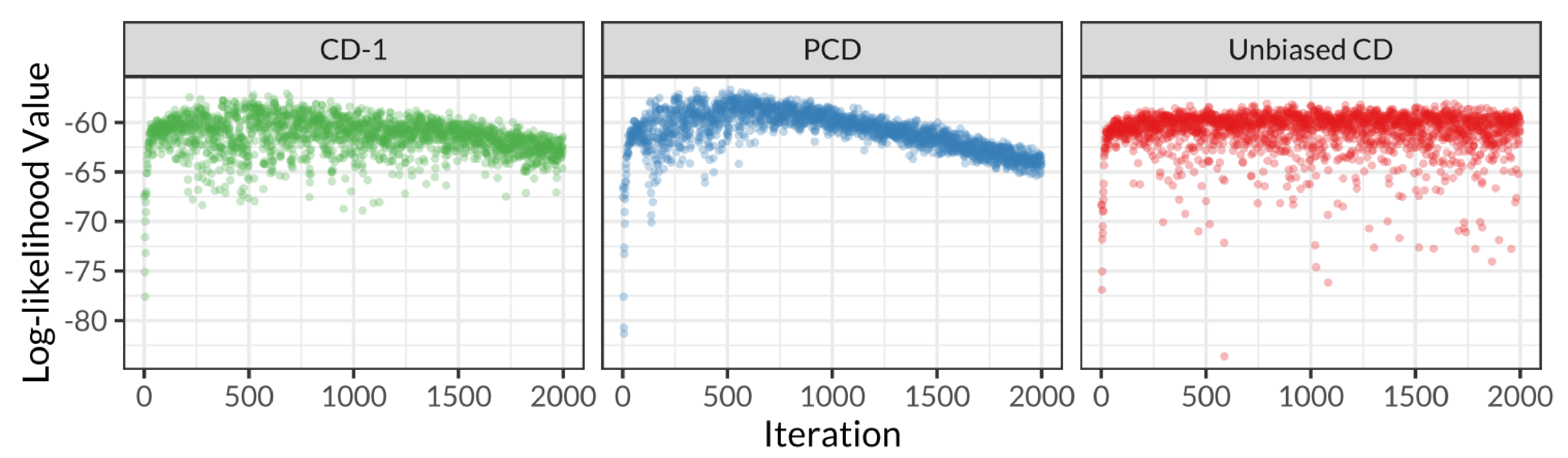
Experiments

- Bars-and-stripes data



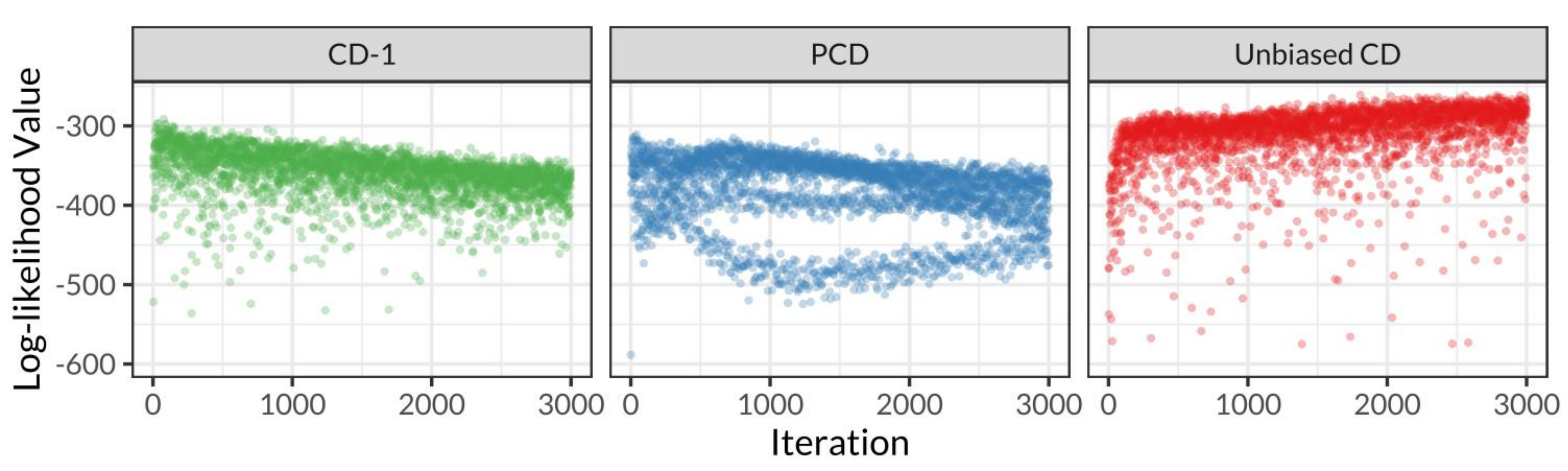
Experiments

- Synthetic RBM model data



Experiments

- Fashion-MNIST



Thanks for Listening

- The algorithm has been implemented in the cdtau package
- Written in efficient C++, with R interface
- Python interface under development

https://github.com/yixuan/cdtau