Understanding Minimum Probability Flow Learning for RBMs under Various Kinds of Dynamics

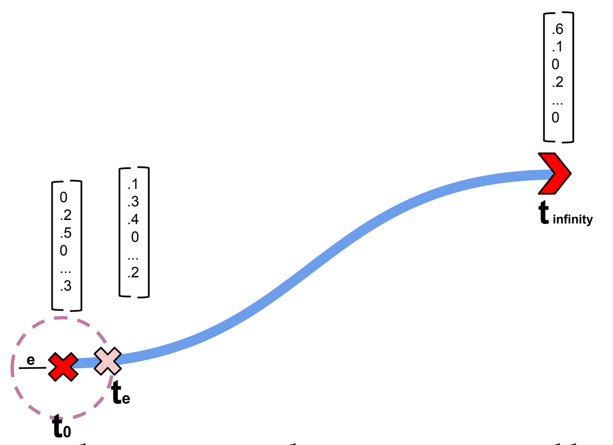
{ Daniel Jiwoong Im, Ethan Buchman, and Graham W. Taylor } University of Guelph, School of Engineering

Introduction

- We investigate the performance of Minimum Probability Flow (MPF) learning for training RBMs.
- Unlike CD, with its focus on approximating an intractable partition function via Gibbs sampling, MPF proposes a tractable, consistent, objective function defined in terms of a Taylor expansion of the KL divergence with respect to sampling dynamics.
- We propose a more general form for the sampling dynamics in MPF, and explore the consequences of different choices for these dynamics for training RBMs.

Dynamics of the model

The key intuition behind MPF is that we introduce explicit dynamics over the model, yielding an equilibrium distribution as a function of dynamics.



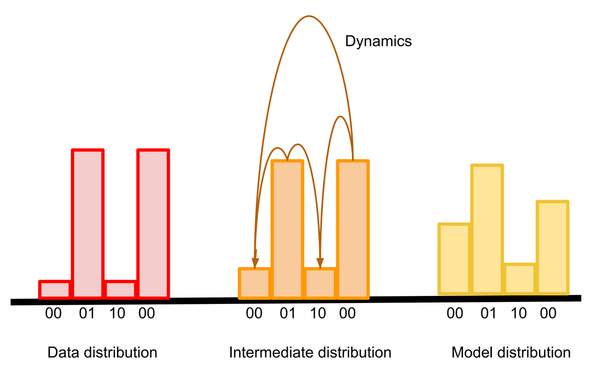
For example, the initial states will evolve over time towards the states under some dynamics.

Using *the Master equation*,

$$\dot{p}_i = \sum_{j \neq i} [\Gamma_{ij} p_j^{(t)} - \Gamma_{ji} p_i^{(t)}] \tag{5}$$

where Γ_{ij} is the probability flow rate from state j to state i, to describe above phenomenon with the assumptions of :

- Continuous time Markov chain,
- Converges to equilibrium state.



Minimum Probability Flow

The objective of MPF is to minimize the KL divergence between the data distribution and the distribution after evolving an infinitesimal amount of time ϵ under the dynamics.

$$\theta_{\text{MPF}} = \text{argmin}_\theta J(\theta), \;\; J(\theta) = D_{KL}(p^{(0)} || p^{(\epsilon)}(\theta))$$

Approximating $J(\theta)$ up to a first order Taylor expansion with respect to time t,

$$J(\theta) = \frac{\epsilon}{|\mathcal{D}|} \sum_{j \in \mathcal{D}} \sum_{i \notin \mathcal{D}} \Gamma_{ij} \tag{1}$$

where we want Γ to satisfy detailed balance,

$$\Gamma_{ji} p_i^{(\infty)}(\theta) = \Gamma_{ij} p_j^{(\infty)}(\theta). \tag{2}$$

One way is to choose Γ to be

$$\Gamma_{ij} = g_{ij} \exp\left(\frac{1}{2}(F_j(\theta) - F_i(\theta)\right). \tag{3}$$

where g_{ij} is the connectivity between state j and i. Choosing sparse g_{ij} $\forall i, j$ allows faster computation!

Probability Flow Rates

Here are different probability flow matrix dynamics, we explore:

1. **One-bit flip** - data states are connected to all other states 1-bit flip away:

$$g_{ij} = \begin{cases} 1, & \text{if state } i, j \text{ differs by single bit flip} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

2. **Factored MPF** - use a probability distribution, such that g_{ij} is the probability that state j is connected to state i. $J(\theta) = J_{\mathcal{D}}(\theta) J_{\mathcal{S}}(\theta)$

$$J_{\mathcal{D}}(\theta) = \left(\frac{1}{|\mathcal{D}|} \sum_{\mathbf{x} \in \mathcal{D}} \exp\left[\frac{1}{2}\left(F(\mathbf{x};\theta) - F(\mathbf{x};\theta^{n-1})\right)\right]\right);$$

$$J_{\mathcal{S}}(\theta) = \left(\frac{1}{|\mathcal{S}|} \sum_{\mathbf{x}' \in \mathcal{S}} \exp\left[\frac{1}{2}\left(-F(\mathbf{x}';\theta) + F(\mathbf{x}';\theta^{n-1})\right)\right]\right)$$

3. **Persistent MPF** - Applied persistent sampling in MPF connectivity functions.

Experiments

MNIST - small models, exact log likelihood computation

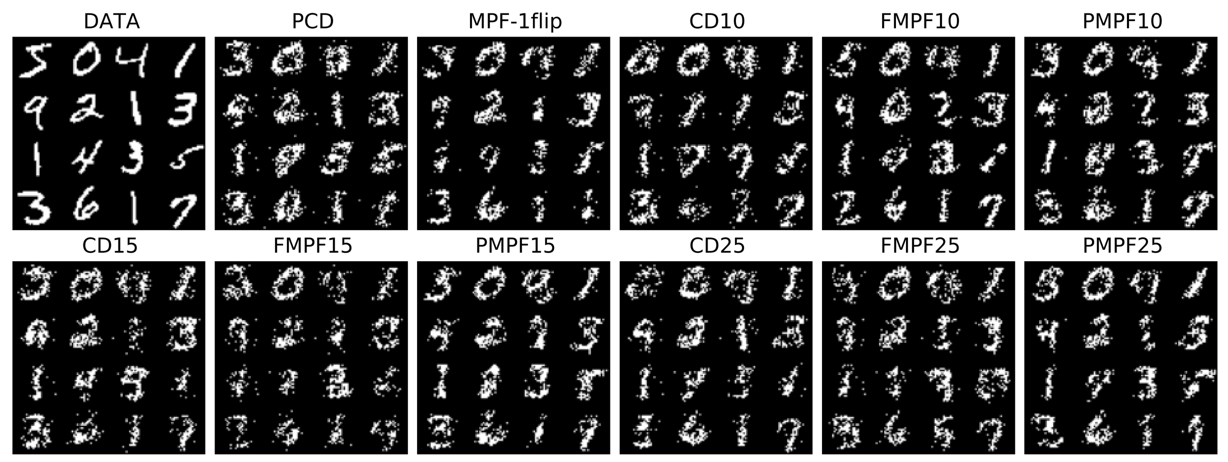


Table 1: Experimental results on MNIST using 11 RBMs with 20 hidden units each. The average training and test log-probabilities over 10 repeated runs with random parameter initializations are reported.

Method	Average log Test	Average log Train	Time (sec)
CD1	-145.63 ± 1.30	-146.62 ± 1.72	831
PCD	**-136.10 ± 1.21**	**-137.13 ± 1.21**	2620
MPF-1flip	-141.13 ± 2.01	-143.02 ± 3.96	2931
CD10	-135.40 ± 1.21	-136.46 ± 1.18	17329
FMPF10	-136.37 ± 0.17	-137.35 ± 0.19	12533
PMPF10	-141.36 ± 0.35	-142.73 ± 0.35	11445
FPMPF10	**-134.04 ± 0.12**	**-135.25 ± 0.11**	22201
CD15	-134.13 ± 0.82	-135.20 ± 0.84	26723
FMPF15	-135.89 ± 0.19	-136.93 ± 0.18	18951
PMPF15	-138.53 ± 0.23	-139.71 ± 0.23	13441
FPMPF15	**-133.90 ± 0.14**	**-135.13 ± 0.14**	27302
CD25	-133.02 ± 0.08	-134.15 ± 0.08	46711
FMPF25	-134.50 ± 0.08	-135.63 ± 0.07	25588
PMPF25	-135.95 ± 0.13	-137.29 ± 0.13	23115
FPMPF25	**-132.74 ± 0.13**	**-133.50 ± 0.11**	50117

MNIST - larger models, log likelihood estimates

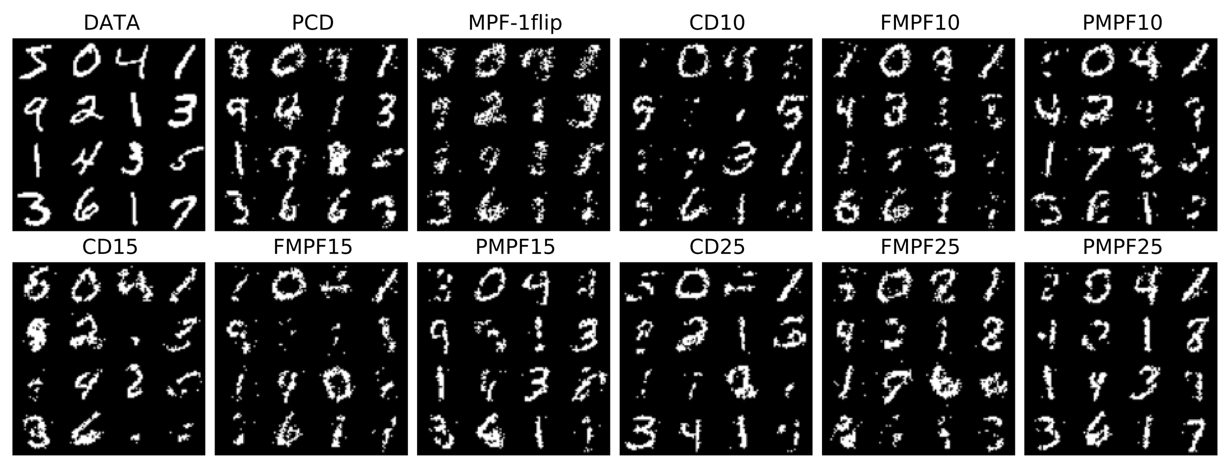


Table 2: Experimental results on MNIST using 11 RBMs with 200 hidden units each. The average estimated training and test log-probabilities over 10 repeated runs with random parameter initializations are reported. Likelihood estimates are made with CSL and AIS.

	CSL	AIS	
Method	Avg. log Test	Avg. log Test	Time (sec)
CD1	-138.63 ± 0.48	-98.75 ± 0.66	1258
PCD1	**-114.14 ± 0.26**	**-88.82 ± 0.53**	2614
MPF-1flip	-179.73 ± 0.085	-141.95 ± 0.23	4575
CD10	-117.74 ± 0.14	-91.94 ± 0.42	24948
FMPF10	-115.11 ± 0.09	-91.21 ± 0.17	24849
PMPF10	-114.00 ± 0.08	-89.26 ± 0.13	24179
FPMPF10	**-112.45 ± 0.03**	**-83.83 ± 0.23**	24354
CD15	-115.96 ± 0.12	-91.32 ± 0.24	39003
FMPF15	-114.05 ± 0.05	-90.72 ± 0.18	26059
PMPF15	-114.02 ± 0.11	-89.25 ± 0.17	26272
FPMPF15	**-112.58 ± 0.03**	**-83.27 ± 0.15**	26900
CD25	-114.50 ± 0.10	-91.36 ± 0.26	55688
FMPF25	-113.07 ± 0.06	-90.43 ± 0.28	40047
PMPF25	-113.70 ± 0.04	-89.21 ± 0.14	52638
FPMPF25	**-112.38 ± 0.02**	**-83.25 ± 0.27**	53379

Caltech-101 - largest models, log likelihood estimates

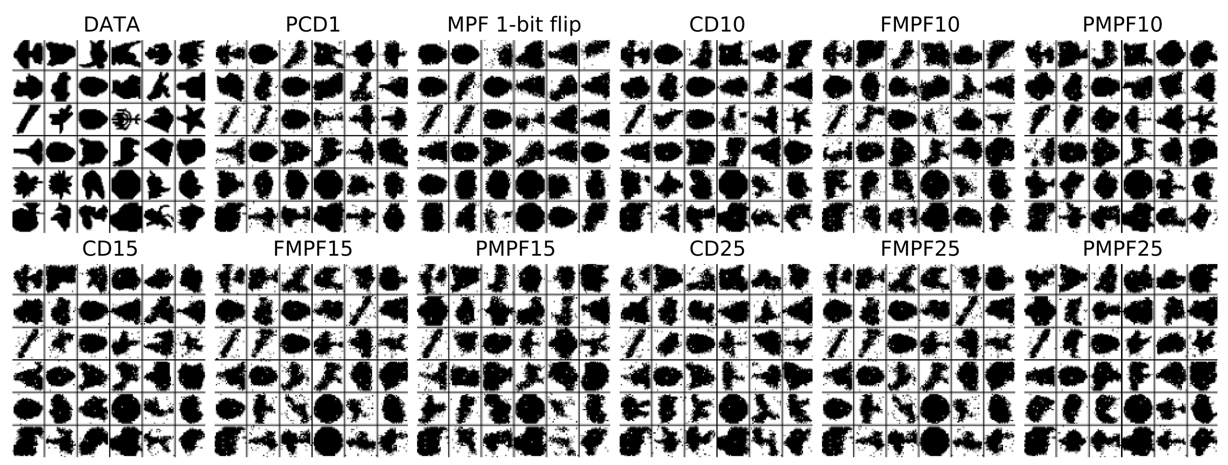


Table 3: Experimental results on Caltech-101 Silhouettes using 11 RBMs with 500 hidden units each. The average estimated training and test log-probabilities over 10 repeated runs with random parameter initializations are reported.

	CSL	AIS	
Method	Avg. log Test	Avg. log Test	Time (sec)
CD1	-251.30 ± 1.80	-141.87 ± 8.80	300
PCD1	**-199.89 ± 1.53**	**-124.56 ± 0.24**	784
MPF-1flip	-281.55 ± 1.68	-164.96 ± 0.23	505
CD10	-207.77 ± 0.92	-128.17 ± 0.20	4223
FMPF10	-211.30 ± 0.84	-135.59 ± 0.16	2698
PMPF10	-203.13 ± 0.12	-128.85 ± 0.15	7610
FPMPF10	**-200.36 ± 0.16**	**-123.35 ± 0.16**	11973
CD15	-205.12 ± 0.87	-125.08 ± 0.24	6611
FMPF15	-210.66 ± 0.24	-130.28 ± 0.14	3297
PMPF15	-201.47 ± 0.13	-127.09 ± 0.10	9603
20			
FPMPF15	**-198.59 ± 0.17**	**-122.33 ± 0.13**	18170
CD25	-201.56 ± 0.11	-124.80 ± 0.20	13745
FMPF25	-206.93 ± 0.13	-129.96 ± 0.07	10542
PMPF25	-199.53 ± 0.11	-127.81 ± 020	18550
FPMPF25	**-198.39 ± 0.0.16**	**-122.75 ± 0.13**	23998

Reference

Sohl-Dickstein, Jascha, Battaglino, Peter, and DeWeese, Michael R. Minimum probability flow learning. In Proceedings of the International Conference of Machine Learning (ICML), 2011

Code: https://github.com/jiwoongim/minimum_probability_flow_learning