Comparing Binary Hamiltonian Monte Carlo and Gibbs Sampling for Training Discrete MRFs with Stochastic Approximation

Hanchen Xiong, Sandor Szedmak, Justus Piater
UNIVERSITY OF INNSBRUCK, AUSTRIA

INTRODUCTION

The past decade witnessed a revival of learning MRFs with sampling-based approximations. In particular, persistent contrastive divergence (PCD) yields remarkable successes in many applications. PCD is a special case of Robbins-Monro's stochastic approximation procedure (SAP) with Gibbs sampling as transitions [1].

It was also recently pointed out in [2] that every discrete MRFs can be generally converted to a Boltzmann machine, which defines a binary distribution over the presence of discrete states of all variables. Therefore, learning discrete MRFs can be converted to learning Boltzmann machines.

This work empirically compared two versions of SAP on learning Boltzmann machines: one with Gibbs sampling as transitions, the other one with binary Hamiltonian Monte Carlo (BHMC), which is a recent extension of HMC [3].

HAMILTONIAN MONTE CARLO

Hamiltonian Monte Carlo (HMC)

1. A Metropolis algorithm with a proposal distribution analogous to Hamiltonian dynamics;
2. Compared to random walk in the standard Metropolis algorithm, HMC can propose a distant jump while still preserving a high acceptance rate.
3. HMC can yield more effective sampling by making use of gradient information of target distribution's density function.

Suppose that we are interested in sampling from $p(\mathbf{x})$ (where $\mathbf{x} \in \mathbb{R}^D$). An auxiliary variable $\mathbf{q} \in \mathbb{R}^D$ with $\mathbf{q} \sim \mathcal{N}(\mathbf{q}; \mathbf{0}, \mathbf{M})$ is introduced (usually $\mathbf{M} = c \cdot \mathbf{I}_D$). A Hamiltonian function can be constructed as:

$$\mathcal{H}(\mathbf{x}, \mathbf{q}) = U(\mathbf{x}) + K(\mathbf{q}) \tag{1}$$

where $U(\mathbf{x})$, $K(\mathbf{q})$ are negative logarithms of $p(\mathbf{x})$ and $p(\mathbf{q})$. The changes of $\mathbf{x}$ and $\mathbf{q}$ over time ν are:

$$\dot{\mathbf{x}}(\nu) = \frac{\partial \mathcal{H}}{\partial \mathbf{q}(\nu)} = \mathbf{M}^{-1}\mathbf{q}(\nu) \quad \dot{\mathbf{q}}(\nu) = -\frac{\partial \mathcal{H}}{\partial \mathbf{x}(\nu)} = -\frac{dU(\mathbf{x})}{d\mathbf{x}(\nu)} \tag{2}$$

Limitation: HMC can only be applied on continuous distributions of which the partial derivatives of the log density function can be computed. Therefore, applying HMC to sample from binary distributions (*e.g.* Boltzmann machines) is not straightforward.

ROBBINS-MONRO'S SAP

A MRF based on exponential family:

$$p(\mathbf{x}; \boldsymbol{\theta}) = \frac{\exp\{-E(\mathbf{x}; \boldsymbol{\theta})\}}{\mathbf{Z}} = \frac{\exp\{\boldsymbol{\theta}^\top \phi(\mathbf{x})\}}{\mathbf{Z}} \tag{7}$$

SAP Algorithm:

1. Training data set $\mathcal{D} = \{\mathbf{x}^1, \cdots, \mathbf{x}^M\}$. Randomly initialize model parameters $\boldsymbol{\theta}^0$ and N particles $\{\mathbf{s}^{0,1}, \cdots, \mathbf{s}^{0,N}\}$.
2. **for** $t = 0 : T$ **do**
3. **for** $n = 1 : N$ **do**
4. Sample $\mathbf{s}^{t+1,n}$ from $\mathbf{s}^{t,n}$ using transition operator $T_{\boldsymbol{\theta}}$.
5. **end for**
6. Update: $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \eta\Big[\frac{1}{M}\sum_{m=1}^{M} \phi(\mathbf{x}^m) - \frac{1}{N}\sum_{n=1}^{N} \phi(\mathbf{s}^{t+1,n})\Big]$
7. Decrease η.

BINARY HAMILTONIAN MONTE CARLO

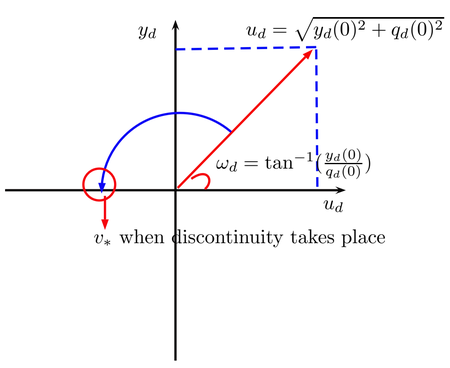
For a Boltzmann machine $p(\mathbf{x} \in \{-1, +1\}^D)$, an auxiliary, continuous variable $\mathbf{y} \in \mathbb{R}^D$ can be added with its conditional probability on $\mathbf{x}$ as a truncated Gaussian:

$$p(\mathbf{y}|\mathbf{x}) = \begin{cases} c \cdot \exp(-\frac{\mathbf{y}^\top \mathbf{y}}{2}) & \forall d \in [1, D], \text{sign}(y_d) = x_d \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$p(\mathbf{y}) = \sum_{\mathbf{x}} p(\mathbf{y}|\mathbf{x})p(\mathbf{x}; \boldsymbol{\theta}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x}; \boldsymbol{\theta}) \tag{4}$$

Since $\mathbf{y}$ is continuous, we can employ HMC to sample from $p(\mathbf{y})$ and $\mathbf{x} = \text{sign}(\mathbf{y})$. By substituting (4) into (1) and (2), we can have:

$$y_d(\nu) = u_d \sin(\omega_d + \nu) \quad q_d(\nu) = u_d \cos(\omega_d + \nu) \tag{5}$$



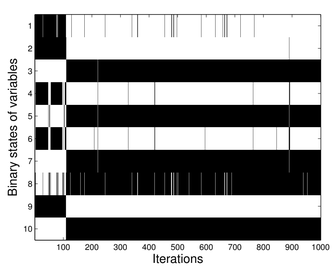
When y_d hits 0 at time ν_*, whether it will be reflected from the $y_d = 0$ or cross it depends on the sign of:

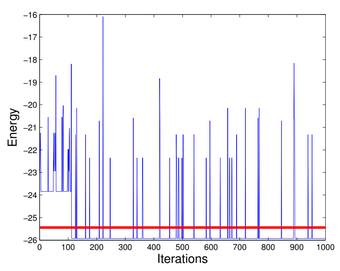
$$\frac{q_d^2(\nu_*^-)}{2} - (E(-x_d, \mathbf{x}_{\neg d}; \boldsymbol{\theta}) - E(x_d, \mathbf{x}_{\neg d}; \boldsymbol{\theta})) \tag{6}$$

where $q_d(\nu_*^-) = u_d$. (6) can be considered as **a pseudo Gibbs sampling**. When $E(-x_d, \mathbf{x}_{\neg d}; \boldsymbol{\theta}) - E(x_d, \mathbf{x}_{\neg d}; \boldsymbol{\theta}) > 0$, the probability of switching sign of x_d is lower than not. According to (6), as long as the energy raise is smaller than a threshold $u_d^2/2$, the switch still can take place.

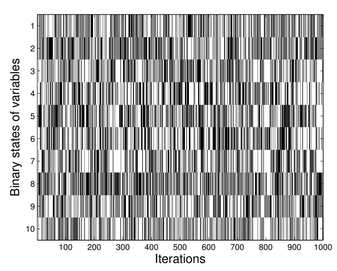
SAMPLING FROM BOLTZMANN MACHINES

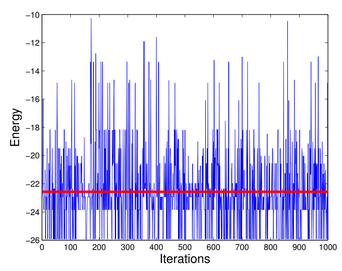
Gibbs Sampling





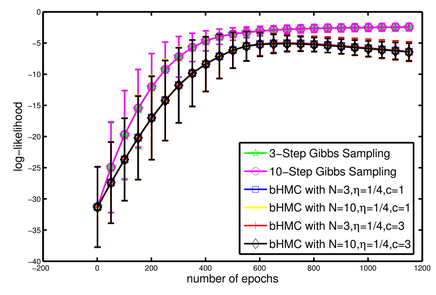
BHMC Sampling





Conclusion: BHMC samples more *"boldly"*, but less *"faithfully"* than Gibbs.

LEARNING MRFS



Conclusion: SAP with Gibbs sampling (*i.e.* PCD) is preferred.

REFERENCES

(1) Ruslan Salakhutdinov. Learning in markov random fields using tempered transitions. In NIPS, 2010.

(2) Yichuan Zhang, Charles Sutton, Amos Storkey, and Zoubin Ghahramani. Continu- ous relaxations for discrete hamiltonian monte carlo. In NIPS. 2012.

(3) Ari Pakman and Liam Paninski. Auxiliary-variable exact hamiltonian monte carlo sampler for binary distributions. In NIPS, 2013.

ACKNOWLEDGEMENT

This research has received funding from the EU FP7 Programme under grant agreement no. 270273, Xperience.