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

Hanchen Xiong Sandor Szedmak Justus Piater
Institute of Computer Science, University of Innsbruck, Technikerstr.21a A-6020, Innsbruck, Austria

Learning discrete Markov random fields (MRFs) has been an important yet challenging machine learning task. In general, learning Markov random fields (MRF) is intractable due to the presence of partition function. Persistent contrastive divergence (PCD), known as a state-of-the-art learning algorithm for Markov random fields (MRFs), is a Robbins-Monro's stochastic approximation procedure (SAP) with Gibbs sampling as transitions (Salakhutdinov, 2010). We conduct an empirical study on a SAP with an alternative transition: binary Hamiltonian Monte Carlo. Ususally sampling-based method (*e.g.* Markov chain Monte Carlo) is employed for approximation. However, when the disibution exhibts multiple modes, the standard Metropplois algorithm will lead to low mxing rate of Markov chains. Hamiltonian Monte Carlo (HMC) is a Metropolis algorithm with a proposal distribution analogous to Hamiltonian dynamics. Compared to random walk in the standard Metropolis algorithm, HMC can propose a distant jump while still preserving a high acceptance rate. 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:

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

HMC can yield more effective sampling by making use of gradient information of target distribution's density function. We can also see, from (2), that 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 discrete MRFs is not straightforward. However, the random variables are discrete in many applications (e.g. computer vision, natural language processing), Luckily, Zhang et al. (2012) pointed out that all discrete MRFs can be generally converted to Boltzmann machines (BMs):

$$p(\mathbf{x}; \boldsymbol{\theta}) = \frac{\exp\left(-E(\mathbf{x}; \boldsymbol{\theta})\right)}{\mathbf{Z}(\boldsymbol{\theta})}, \quad E(\mathbf{x}; \boldsymbol{\theta}) = -\sum_{i<j} x_i x_j W_{ij} \tag{3}$$

where $\mathbf{x} \in \{-1, +1\}^D$, $\boldsymbol{\theta} = \{W_{ij}\}$. In addition, Pakman and Paninski (2013) developed an extension of HMC for binary distribution to learn Bayesian regression with spike-and-slab prior. Here we attempt to apply this extended HMC, which we refer to as binary HMC (bHMC), for our purpose of learning Boltzmann machines (so also discrete MRFs). Assume that we are interested in sampling from 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:

$$p(\mathbf{y}|\mathbf{x}) = \begin{cases} c \cdot \exp(-\frac{\mathbf{y}^\top \mathbf{y}}{2}) & \forall d \in [1, D], \text{sign}(y_d) = x_d \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Hence, instead of sampling $\mathbf{x}$ directly, we can sample $\mathbf{y}$ first and take their signs as our desired samples. By making use of orthant consistency constraint in (4), we can have:

$$p(\mathbf{y}) = \sum_{\mathbf{x}} p(\mathbf{y}|\mathbf{x}) p(\mathbf{x}; \boldsymbol{\theta}) = p(\mathbf{y}|\mathbf{x}) p(\mathbf{x}; \boldsymbol{\theta}) \tag{5}$$

Since $\mathbf{y}$ is continuous, we can employ HMC to sample them from $p(\mathbf{y})$. By substituting (5) into (1) and (2), we can have:

$$y_d(\nu) = u_d \sin(\omega_d + \nu) \quad q_d(\nu) = u_d \cos(\omega_d + \nu) \tag{6}$$

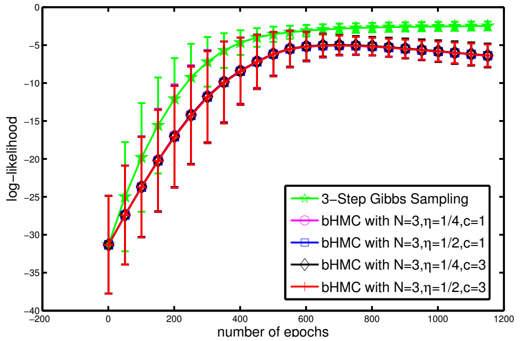
where $u_d = \sqrt{y_d(0)^2 + q_d(0)^2}$, and $\omega_d = \tan^{-1}\left(\frac{y_d(0)}{q_d(0)}\right)$. In addition, since (6) keeps Hamiltonian function (1) invariant, change of $\mathbf{y}, \mathbf{q}$ according to (6) are always accepted. It can be seen in (6) that (q_d, y_d) actually moves counterclockwisely along a circle with radius u_d. However, one issue arising from discontinuity of $p(\mathbf{y}|\mathbf{x})$ is that 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:

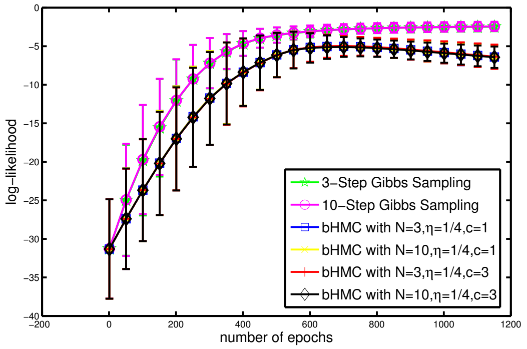
$$\frac{q_d^2(\nu_*^-)}{2} - \left(E(-x_d, \mathbf{x}_{\neg d}; \boldsymbol{\theta}) - E(x_d, \mathbf{x}_{\neg d}; \boldsymbol{\theta})\right) \tag{7}$$

where $q_d(\nu_*^-)$ is the q_d immediately before time ν_*, so it equlals to u_d. (7) 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 (7), as long as the energy raise is smaller than a threshold $u_d^2/2$, the switch still can take place. In addition, with different initializations of $(y_d(0), q_d(0))$, different y_d will hit 0 at different time ν_*^d, so bHMC is a randomly scheduled sampling. Finally, since $y_d(N\pi) = y(0)$ if y_d always gets reflected, traveling time is recommended as $T = (N + \eta)\pi$ so as to avoid degeneracy of samples ($\eta \in (0, 1)$). In conclusion, bHMC somehow resembles Gibbs sampling but with a different acceptance criterion. To verify its practical applicabilities, we compared it against Gibbs sampling in SAP (with different c, N, η) on training a toy Boltmann machine ($D = 10$), the results are presented as follows. Our empirical results suggest that the SAP with bHMC is inferior to the one with Gibbs sampling for learning discrete MRFs.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme





FP7/2007-2013 (Specific Programme Cooperation, Theme 3, Information and Communication Technologies) under grant agreement no. 270273, Xperience.

References

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

Ruslan Salakhutdinov. Learning in markov random fields using tempered transitions. In *NIPS*, 2010.

Yichuan Zhang, Charles Sutton, Amos Storkey, and Zoubin Ghahramani. Continuous relaxations for discrete hamiltonian monte carlo. In *NIPS*. 2012.