Tsinghua University

Joint Stochastic Approximation Learning of Helmholtz Machines

Haotian Xu, Zhijian Ou

Department of Electronic Engineering, Tsinghua University

Abstract

- **Helmholtz machines** : models with a pair of generative and inference models $p_\theta(x,h)$ and $q_\varphi(h|x)$.
- **JSA** : directly optimize **marginal log-likelihood** & simultaneously optimize **inclusive KL divergence**, in the framework of stochastic approximation (SA).
- To sample true posterior, treat inference network as proposal and construct two types of MCMC operators – **MIS** and **MTMIS.**
- JSA outperforms RWS with better log-likelihoods on MNIST.
- MTMIS enables larger move and improves mixing**.**

Related Work

Algorithm		$p_\theta(\boldsymbol{x},\boldsymbol{h})$			$q_\phi(\boldsymbol{h}\vert\boldsymbol{x})$		RV type	
		ML	V-LB	IS-LB	$KL(q\vert\vert p)$	$KL(p\vert\vert q)$	C	D
1	VAE [1]		√		√		√	
	NVIL [2]		√		√		√	√
	MuProp [3]		√		√		√	√
2	WS [4]		√			√	√	√
	RWS [5]			√		√	√	√
3	IWAE [6]			√	√		√	
	JSA	√				√	√	√

References

[1] Kingma, Diederik P and Welling, Max, Auto-Encoding Variational Bayes, ICLR, 2014.
[2] A. Mnih and K. Gregor, Neural variational inference and learning in belief networks, ICML, 2014.
[3] S. Gu, S. Levine, I. Sutskever, and A. Mnih, Muprop : unbiased back-propagation for stochastic neural networks, ICLR, 2016.
[4] G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal, The wake-sleep algorithm for unsupervised neural networks, Science, 1995.
[5] J. Bornschein and Y. Bengio, Reweighted wake-sleep, ICLR, 2015.
[6] Y. Burda, R. Grosse, and R. Salakhutdinov, Importance weighted auto-encoders, ICLR, 2016.

JSA Learning of Helmholtz Machines

Simultaneous equations to be solve by SA

$$\frac{\partial}{\partial\theta} logp_\theta(x) = E_{p_\theta(x|h)}\left[\frac{\partial}{\partial\theta} logp_\theta(x,h)\right] = 0$$

$$\frac{\partial}{\partial\varphi} KL\left(p_\theta(h|x)||q_\varphi(h|x)\right) = -E_{p_\theta(x|h)}\left[\frac{\partial}{\partial\varphi} logq_\varphi(h|x)\right] = 0$$

SA recursion for updating parameters

$$\begin{cases} \theta^{(t)} = \theta^{(t-1)} + \alpha_t \dfrac{\partial}{\partial\theta} logp_\theta(x,h^{(t)}) \\ \varphi^{(t)} = \varphi^{(t-1)} + \beta_t \dfrac{\partial}{\partial\varphi} logq_\varphi(h^{(t)}|x) \end{cases}$$

MCMC operators used in JSA

Given the current state $x^{(t)}$, target distribution $\pi(x)$;

Importance sampling weight $\omega(x) = \frac{\pi(x)}{g(x)}$

Metropolis Independence Sampler	Multiple-trial Metropolis Independence Sampler
• Draw $y \sim g(y)$	• Generate K I.I.D samples $y_j \sim g(y)$,W $= \sum_{j=1}^{K}\omega(y_j)$
1 vs K samples	• Draw y from $\{y_1,\ldots,y_K\}$ with the probability proportional to $\omega(y_j)$
• Set $x^{(t+1)} = y$ with probability $min\left\{1, \frac{\omega(y)}{\omega(x^{(t)})}\right\}$,	• Set $x^{(t+1)} = y$ with probability $min\left\{1, \frac{W}{W-\omega(y)+\omega(x^{(t)})}\right\}$

Results for SBN and categorical SBN(C) on MNIST dataset

Model	200	200-200	200-200-200	200-10(C)	200-200-10(C)	200-200-200-10(C)
	100,000 samples Negative log-likelihood estimated by importance sampling. (Negative log-likelihood variational bound)					
WS	116.3[5] (120.7[5])	106.9[5] (109.4[5])	101.3[5] (104.4[5])			
RWS	103.1[5] —	93.4[5] —	90.1[5] —	97.65 (109.41)	90.35 (99.71)	88.43 (96.09)
JSA-MIS	103.5 (112.7)	93.33 (101.00)	89.85 (97.04)	97.8 (106.83)	91.60 (98.04)	88.43 (96.09)
JSA-MTMIS	102.3 (116.37)	92.11 (101.88)	88.92 (98.20)	97.05 (110.39)	89.84 (98.93)	87.82 (96.58)
NVIL	— (113.1[2])	— (99.8[2])	— (96.7[2])			
MuProp	— (113.1[3])	— (100.4[3])	— (98.6[3])	— (107.8[3])		

Convergence curves of JSA-MIS, JSA-MTMIS and RWS for SBN 200-200