ORIGINAL RESEARCH

An effect of initial distribution covariance for annealing Gaussian restricted Boltzmann machines

Taichi Kiwaki *[1], Kazuyuki Aihara[2]

[1] *Graduate School of Engineering, The University of Tokyo, Tokyo, Japan*
[2] *Institute of Industrial Science, The University of Tokyo, Tokyo, Japan*

Received: October 26, 2014 **Accepted:** December 2, 2014 **Online Published:** January 20, 2015
DOI: 10.5430/air.v4n1p53 **URL:** http://dx.doi.org/10.5430/air.v4n1p53

Abstract

In this paper, we investigate an effect that the covariance of an initial distribution for annealed importance sampling (AIS) exerts on the estimation accuracy for the partition functions of Gaussian restricted Boltzmann machines (RBMs). A common choice for an AIS initial distribution is a Gaussian RBM (GRBM) with zero weight connections. Such an initial distribution does not show any covariance between variables. However, target distributions generally allow a finite covariance between variables. We propose a method to design the covariance matrix of an initial distribution for GRBMs. We empirically analyze the effect of the initial distribution covariance on the estimation accuracy of AIS. The proposed method for designing initial distributions outperforms conventional methods under various conditions.

Key Words: Annealed importance sampling, Initial distribution covariance, Gaussian restricted Boltzmann machines

1 Introduction

Many stochastic latent feature models are defined by unnormalized probability or density function, and the exact computation of the normalizing constant, or partition function, is usually intractable. This causes a problem when we compare different models or monitor training of models by checking the probabilities that models assign to validation data. Therefore, approximate inference for partition functions has attracted substantial research interest.[1–3] Annealed importance sampling (AIS) is commonly applied to model validation because unbiased estimates can be obtained with adequate computational resources.[2,4,5] If we do not choose the annealing parameters carefully, however, AIS can give inaccurate estimates.

AIS uses a tractable initial distribution to estimate the statistics of the intractable target distribution. For restricted Boltzmann machines (RBMs), a common choice for the initial distribution is an RBM with zero weight connections.[4,5] Particularly for Gaussian RBMs (GRBMs), the initial distribution is a Gaussian distribution that does not model any covariance between variables. However, target distributions generally can model a non-zero covariance between variables. Although the mathematical framework for running AIS with RBMs leaves enough flexibility for choosing a tractable RBM with non-zero weight connections, there is no established method for designing proper weights for an initial distribution.

In this paper, we propose an AIS algorithm for GRBMs in which an initial distribution is a multivariate normal distribution with a nondiagonal covariance matrix. We experimentally compare the proposed method with standard methods for AIS estimation. The proposed method for designing

*Correspondence: Taichi Kiwaki; Email: kiwaki@sat.t.u-tokyo.ac.jp; Address: Graduate School of Engineering, The University of Tokyo, Tokyo, Japan

the covariance matrix of an initial distribution outperforms the standard methods in almost all conditions we examined.

2 Gaussian restricted Boltzmann machines

RBMs are Markov random fields of a bipartite graph that consists of two layers of variables: a visible layer representing data, and a hidden layer representing latent variables.[6] GRBMs are one of the variants of RBMs with real-valued visible variables $\mathbf{v} \in \{0,1\}^M$. The energy of the state $\{\mathbf{h}, \mathbf{v}\}$ is

$$E(\mathbf{h}, \mathbf{v}; \theta) = -\sum_{i=1}^{M}\sum_{j=1}^{D}\frac{v_j}{\sigma_j}w_{ij}h_i - \sum_{i=1}^{M}a_i h_i + \sum_{j=1}^{D}\frac{(v_j - b_j)^2}{2\sigma_j^2} \tag{1}$$

where $\theta = \{W, \mathbf{a}, \mathbf{b}, \sigma\}$ are the model parameters: $W = (w_{ij})$ is the connection matrix between hidden units and visible units; a_i and b_j are hidden and visible biases and σ_j^2 are variances of visible variables within a single mode.

The probability density function of a GRBM over $\mathbf{v}$ is

$$p(\mathbf{v}; \theta) = \frac{p*(\mathbf{v}; \theta)}{Z(\theta)} = \frac{1}{Z(\theta)}\sum_{\mathbf{h}}\exp(-E(\mathbf{h}, \mathbf{v}; \theta)) \tag{2}$$

$$Z(\theta) = \sum_{\mathbf{h}}\int_{\mathbf{v}}\exp(-E(\mathbf{h}, \mathbf{V}; \theta))\prod_{j}^{D}\mathrm{d}v_j c \tag{3}$$

where p^* represents the unnormalized probability density, and $Z(\theta)$ is the partition function.

3 Annealed importance sampling

Suppose that we have a distribution defined on some space ν with a probability density function $p_B(\mathbf{v}) = p_B^*(\mathbf{v})/Z(\theta^B)$, where we can efficiently evaluate $p_B^*(\mathbf{v})$, for $\mathbf{v} \in \nu$, and we compute the partition function $Z(\theta^B)$. One method to estimate the partition function is importance sampling (IS). Assume that we have a tractable distribution defined by p_A s.t. $p_A(\mathbf{v}) \neq 0 \Leftarrow p_B(\mathbf{v}) \neq 0$, then we have the following Monte Carlo approximation: $Z(\theta^B) = \int \frac{P_B^*(\mathbf{v})}{P_A(\mathbf{v})}P_A(\mathbf{v})\mathrm{d}\mathbf{v} \approx \frac{1}{N}\sum_i^N \frac{P_B^*(\mathbf{v}_i)}{P_A(\mathbf{v}_i)}$, where $\mathbf{v}_i \sim p_A$. If $\mathbf{v}_i$ are i.i.d., this Monte Carlo approximation gives us an unbiased estimate for the partition function as $N \to \infty$. However, unless p_A and p_B are sufficiently close, as is not often the case, the estimate by IS can have a large variance and cannot be reliable.

The AIS algorithm alleviates this problem by considering a sequence of *annealed* intermediate distributions that bridges the gap between p_A and p_B.[7] When using AIS, we need to define this sequence, which we call a *path*, $\{p_k(\mathbf{v})\}$ for $k \in \{0, \ldots, K\}$, where the starting point of the path $p_0(\mathbf{v}) = p_A(\mathbf{v})$ is the tractable initial distribution, and the end point $p_K(\mathbf{v}) = p_B(\mathbf{v})$ is the intractable target distribution. For each $p_k(\mathbf{v})$, we also need to define a Markov Chain Monte Carlo (MCMC) transition operator T_k that renders p_k invariant. Algorithm 1 summarizes the procedure of AIS in which MCMC transitions and importance weight updates are alternatively performed.

Algorithm 1 The AIS algorithm

for $i = 1$ *to* N **do**
  $\mathbf{v}_0 \leftarrow$ sample from $p_0(\mathbf{v})$
  $w^{(i)} \leftarrow 1$
  for $k = 1$ *to* K **do**
    $w^{(i)} \leftarrow w^{(i)}\frac{p_k^*(\mathbf{v}_{k-1})}{p_{k-1}^*(\mathbf{v}_{k-1})}$
    $\mathbf{v}_k \leftarrow$ sample from $T_k(\mathbf{v}_k, \mathbf{v}_{k-1})$
  end
end
return $\hat{Z}(\theta^B) = Z(\theta^A)\sum_i^N w^{(i)}/N$

AIS actually belongs to a family of algorithms for partition function estimation based on the following equality:[1,8,9]

$$\log Z(\theta^B) - \log Z(\theta^A) = \int_0^1 E_\beta[\frac{\partial}{\partial\beta}\log p_\beta^*(\mathbf{v})]\mathrm{d}\beta \tag{4}$$

where $p_\beta(\cdot)$ is a continuously parameterized probability mass or density function with $\beta \in [0,1]$ s.t. $p_{\beta=0}(\cdot) = p_A(\cdot)$ and $p_{\beta=1}(\cdot) = p_B(\cdot)$. AIS is a finite difference approximation of the integral on the r.h.s. of Eq.**??** where the interval $[0,1]$ is partitioned by a monotonic sequence $\{\beta_k\}$ $(k = 0, \ldots, K)$ and $p_{\beta_k}(\cdot)$ is substituted by $p_k(\cdot)$. Although the equality of Eq.**??** originates from statistical physics, one should note that β can be any parameterization and is not necessarily the inverse temperature.

As with IS, AIS also produces an unbiased estimate as $N \to \infty$. Especially, the unbiasedness is achieved even if T_k not return independent samples. However, in practice, the variance of AIS estimates can be quite large depending on several factors. First, as suggests, poor mixing of T_k can damage the estimation accuracy. Recently, introduced Hamiltonian dynamics for sampling visible units to ease this problem. Second, the choice of the annealing path can have a great impact on the estimation accuracy. A typical choice for the annealing path is the following *geometric path*:

$$p_k(\mathbf{v}) = p_{\beta=\beta_k}(\mathbf{v}) \propto p_A(\mathbf{v})^{1-\beta_k}p_B(\mathbf{v})^{\beta_k} \tag{5}$$

where $\{\beta_k\}$ is a sequence of real numbers s.t. $0 = \beta_0 < \beta_1 < \cdots < \beta_K = 1$.[4,7] Note that we introduced a notation $p_{\beta=\beta_k}(\cdot)$ equivalent to $p_k(\cdot)$ for convenience. Although the geometric path is suboptimal in estimation accuracy,[1] this path is useful and widely implemented.[4,10] Grosse et al.[5] recently developed an alternative method for constructing a path to improve performance.

To determine the reliability of an AIS estimate, we can use a statistic called the effective sample size (ESS),[5,7] which

can be computed as

$$\mathrm{ESS} = \frac{N}{1 + s^2(w_*^{(i)})} \tag{6}$$

where $s^2(w_*^{(i)})$ is the sample variance of the normalized AIS weights $w_*^{(i)} = Nw^{(i)} / \sum_{i=1}^{N} w^{(i)}$. The ESS roughly measures the number of effective AIS samples with large AIS weights. Because the variance of estimates is effectively dominated by such samples, the variance is approximately proportional to ESS^{-1}. Note that caution should be exercised when using the ESS because it can be misleading when AIS samples fail to find important modes of the target distribution.

AIS for GRBMs

Suppose that we estimate the partition function of a GRBM with parameters $\theta^B = \{W^B, \mathbf{a}^B, \mathbf{b}^B, \sigma^B\}$ via AIS by using another GRBM with parameters $\theta^A = \{W^A, \mathbf{a}^A, \mathbf{b}^A, \sigma^A\}$ as an initial distribution. Because the MCMC operators for intermediate distributions of RBMs along the geometric path are not efficient, Salakhutdinov and Murray[4] instead proposed the use of the geometric path between the joint distributions of RBMs. The energy of intermediate distributions becomes

$$E_\beta(\mathbf{h}, \mathbf{v}) = \beta E(\mathbf{h}, \mathbf{v}; \theta^B) + (1-\beta) \sum_{j=1}^{D} \frac{(v_j - b_j^A)^2}{2\sigma_j^{A^2}} \tag{7}$$

where we assumed a convention that an initial distribution has zero weight, i.e., $W_{ij}^A = 0$. This quite prevalent assumption allows us to easily compose a tractable initial distribution. Although there is an alternative way to compose a tractable RBM with non-zero weight connections (e.g., by limiting the number of hidden units), no extensive research has been made on this point. Therefore, almost all application studies rely on the zero weight convention. Because the intermediate distributions defined as Eq **??**, are also GRBMs, we can easily evaluate the logarithm of the unnormalized density as follows:

$$\log p_\beta^*(\mathbf{v}) = -\beta \sum_{j=1}^{D} \frac{(v_j - b_j^B)^2}{2\sigma_j^{B^2}} + \sum_{i=1}^{M} \log \left\{ 1 + e^{\beta\left(\sum_{j=1}^{D} w_{ij}^B \frac{v_j}{\sigma_j^B} + a_i \right)} \right\} - (1-\beta) \sum_{j=1}^{D} \frac{(v_j - b_j^A)^2}{2\sigma_j^{A^2}} \tag{8}$$

The MCMC transition operators that render $p_\beta(\mathbf{v})$ invariant are also easily obtained as

$$p_\beta(h_i = 1 | \mathbf{v}) = \mathrm{sigm}\left(\beta \left(\sum_{j=1}^{D} W_{ij}^B \frac{v_j}{\sigma_j^B} + a_i^B \right) \right), p_\beta(v_j | \mathbf{h}) = N\left(v_j | m_j(\beta, \mathbf{h}), \sigma_j^2(\beta)\right) \tag{9}$$

where $\sigma_j^2(\beta) = \left\{ \frac{\beta}{\sigma_j^{A^2}} + \frac{1-\beta}{\sigma_j^{A^2}} \right\}^{-1}$, $m_j(\beta, \mathbf{h}) = \sigma_j^2(\beta) \left\{ \frac{\beta}{\sigma_j^{B^2}} (\sigma_j^B \sum_{i=1}^{M} W_{ij}^B h_i + b_j^B) + \frac{(1-\beta)}{\sigma_j^{A^2}} b_j^A \right\}$, $\mathrm{sigm}(x) = 1/(1 + \exp(-x)))$, and $\mathcal{N}(\cdot | \mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2.

Model parameters that enable fast mixing, i.e., *hot distributions*, are suitable for initial distributions.[5] Therefore, the weight matrix of an initial distribution is usually set to zero so that it has only a single mode. Another commonly adopted technique is to choose an initial distribution that approximates the target distribution in terms of its first and second moments. Assuming the weight matrix is zero, this can be performed by setting b_j^A and $\sigma_j^{A^2}$ to the estimated means and variances of the target distribution. The estimation can be carried out by using data that are used to train the target distribution[4] or by using MCMC approximation with the target distribution.

4 Initial distributions with nondiagonal covariance matrices

Initial distributions given in the previous section do not model any dependency between variables; the covariance matrix of an initial distribution is restricted to be diagonal. However, target distributions can generally have a non-zero covariance between variables. This observation motivates us to develop a method that manages initial distributions with any covariance matrix.

We propose the following energy function for intermediate distributions instead of Eq.7:

$$E_\beta(\mathbf{h}, \mathbf{x}, \mathbf{v}) = \beta E(h, v; \theta^\beta) + (1-\beta) \left\{ \sum_{j=1}^{D} \frac{(v_j - x_j)^2}{2\sigma_j^{B^2}} + \frac{1}{2} (\mathbf{x} - \mathbf{b}^A)^T \Lambda (mathbf{x} - \mathbf{b}^A) \right\} \tag{10}$$

where $\mathbf{x} \in \mathbf{R}^D$ are newly introduced hidden variables that obey a multivariate normal distribution with a covariance matrix $\Lambda^{-1}/(1-\beta)$ and means $\mathbf{b}^A$. At the point $\beta = 0$ (i.e., the initial distribution), the variables control the means of visible variables that are also normally distributed, given $\mathbf{x}$. It is easy to obtain the marginal for $\mathbf{v}$, which is also normally distributed. Based on these observations, we can gain the logarithm of the unnormalized density of intermediate distributions as follows:

$$\log p^*_\beta(\mathbf{v}) = -\sum_{j=1}^{D} \beta \frac{(v_j - b_j^B)^2}{2\sigma_j^{B^2}} + \sum_{i=1}^{M} \log \left\{ 1 + e^{\beta(\sum_{j=1}^{D} W_{ij}^B) \frac{v_j}{\sigma_j^B} + a_i} \right\} - (1-\beta)\frac{1}{2}(\mathbf{v}-\mathbf{b}^A)^T(\Lambda^{-1}+\Lambda^{B^{-1}})^{-1}(\mathbf{v}-\mathbf{b}^A) \tag{11}$$

where we define $\Lambda^B = diag(\sigma_1^{B^{-2}}, \cdots, sigma_D^{B^{-2}})$. This joint distribution is illustrated as an undirected graph in Figure 1. Note that we can obtain i.i.d. samples from the initial distribution because it is an unimodal normal distribution.

Equation **??** shows that the covariance matrix and means of the initial distribution are $\sum^A = \Lambda^{-1} + \Lambda^{B^{-1}}$ and $\mathbf{b}^A$. Conversely, we can design the initial distribution to have a covariance matrix $\sum^A$ by setting $\Lambda = (\sum^A - \Lambda^{B^{-1}})^{-1}$.

Because of the conditional independence $p(\mathbf{h}, \mathbf{x}|\mathbf{v}) = p(\mathbf{h}|\mathbf{v})p(\mathbf{x}|\mathbf{v})$, the MCMC transition operators can be defined as follows:

$$p_\beta(h_i = 1|\mathbf{v}) = \text{Sigm}\left(\beta(\sum_{j=1}^{D} W_{ij}^B \frac{v_j}{\sigma_j^B} + a_i^B)\right) \tag{12}$$

$$p_\beta(\mathbf{x}|\mathbf{v}) = N(\mathbf{x}|(\Lambda + \Lambda^B)^{-1}(\Lambda\mathbf{b}^A + \Lambda^B\mathbf{v}), (\Lambda + \Lambda^B)^{-1}/(1-\beta)) \tag{13}$$

and

$$p_\beta(v_j|\mathbf{x}, \mathbf{h}) = N\left(v_j|\beta(\sigma_j^B \sum_{i=1}^{M} W_{ij}^B h_i + b_j^B) + (1-\beta)x_j, \sigma_j^{B^2}\right) \tag{14}$$

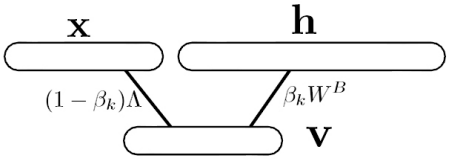

Figure 1: Proposed model

The conditional distributions over $\mathbf{x}$ given $\mathbf{v}$ are multivariate normal distributions with covariance matrices $(\Lambda + \Lambda^B)^{-1}/(1-\beta)$. Sampling from these distributions can be efficiently performed because the covariance matrices are merely scaled on the annealing path; few matrix operations are required for each MCMC transition once the eigenvectors and eigenvalues of $\Lambda + \Lambda^B$ are computed.

Figures 2 and 3 show the evolution of $p_k(\mathbf{v})$ along the annealing path with the conventional and the proposed method. A target distribution is a GRBM with $M = 2$ and $D = 2$ that has four modes: two large and two small. Initial distributions have the same moments (up to the second order) as the target distribution. We can observe that the conventional method assigns more sampling points to small modes at the end of the annealing path ($\beta = 1.0$) than the proposed method. Therefore, the proposed method should produce more accurate estimates than the conventional method.

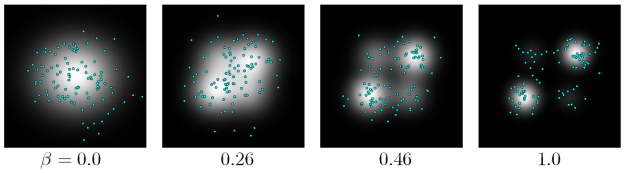

Figure 2: Heatmaps (white denotes large) of annealed distributions p_β by the **conventional** method (labeled as **AIS**). Corresponding values of beta are shown below. Points (best viewed in color) are the sample points of AIS.

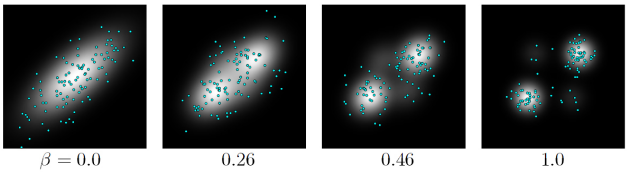

Figure 3: Heatmaps as Fig.2 by the **proposed** method (**AIS_COV**).

4.1 Remarks

Gelman and Meng[1] showed that the variance of estimates based on Eq.[4] can be decomposed into two factors: one comes from the difference of the partition functions $Z(\theta^B)$ and $Z(\theta^A)$, and one comes from the difference of the shapes of the distributions $p_B(\mathbf{v})$ and $p_A(\mathbf{v})$. By selecting the moments of an initial distribution, we can minimize the second factor. Gelman and Meng[1] derived a lower bound for this factor and showed that the optimal initial distribution minimizes the Hellinger distance between $p_B(\mathbf{v})$ and $p_A(\mathbf{v})$. However, our current strategy of matching the moments between $p_A(\mathbf{v})$ and $p_B(\mathbf{v})$ corresponds to minimization of the KL divergence $D_{\mathrm{KL}}(p_B||p_A)$. Therefore, this strategy is suboptimal.

It is convenient to consider the α-divergence[11] to see the relationship between our strategy and the optimal one. The α-divergence forms a family of divergences parameterized by a scalar parameter α, and includes both the KL divergence and the Hellinger distance as its instances. The Hellinger distance corresponds to $\alpha = 0.5$ and the KL divergence corresponds to $\alpha = 1$. This suggests that our strategy approximates the optimal strategy by minimizing the α-divergence of $\alpha = 1$ instead of the optimal value $\alpha = 0.5$.

4.2 Related methods

Jascha Sohl-Dickstein et al.[16] recently proposed to use Hamiltonian dynamics for AIS to accelerate the mixing of Markov chains and demonstrated the efficacy on mean-covariance RBMs. Because MCMC transition operators and an initial distribution can be defined independently, our method would be extended to use Hamiltonian dynamics with little effort.

Roger Grosse[11] recently reported that annealing along the *moment averaging path* empirically achieves better estimation accuracy than along the geometric path. Because the choice of the annealing path is independent of the choice of an initial distribution, our method can readily be used with the moment averaging path.

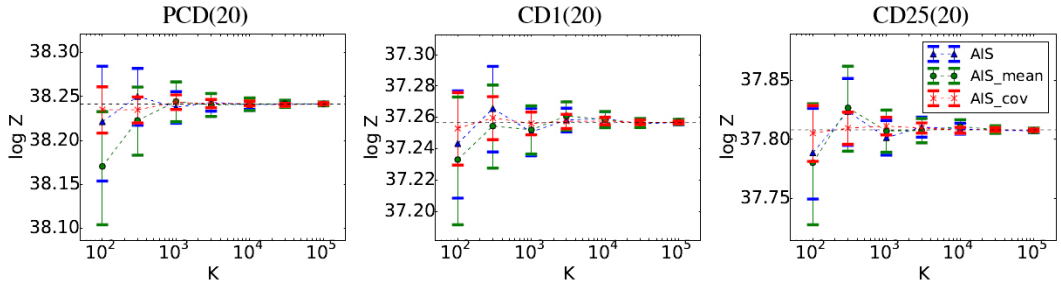

Figure 4: Estimated $\log Z(\theta^B)$ for tractable **GRBMs**. Error bars show $\pm 3\theta$ intervals where σ^2 is the sample variance of the AIS estimate.

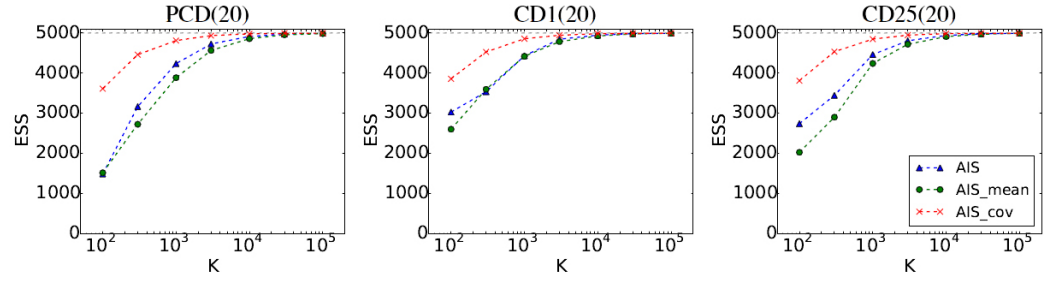

Figure 5: ESS (the larger the better) for tractable GRBMs

5 Experiments

In our experiments, we compared the following methods for designing initial distributions: (AIS) the conventional method with Eq.7 where $\mathbf{b}^A$ and σ^A are identically chosen as the target distribution; (AISM) another baseline method based on Eq. 7 where $\mathbf{b}^A$ is determined from the target distribution but σ^A is simply determined as $\sigma^A = \sigma^B$; and (AISC) the proposed method with Eq. 10 where $\mathbf{b}^A$ and Λ are chosen according to the target distribution statistics. Note that we examine **AISMMean** as well as **AIS** and **AISC** to better illustrate the impacts that the covariance matrix of initial distributions can have on the estimation accuracy. For all three methods, we approximated the first and second order moments of target distributions by using samples drawn from 100 independent Markov chains of length 5000 where the initial 100 samples were discarded as burn-in samples. Estimation was made with a various number of intermediate distributions K. The annealing schedule $\{\beta_k\}$ was di-

vided into four periods: β_k uniformly spaced from 0 to 0.1, β_k uniformly spaced from 0.1 to 0.25, β_k uniformly spaced from 0.25 to 0.5, and β_k uniformly spaced from 0.5 to 1; $K/4$ intermediate distributions were assigned to all of these periods. The numbers of AIS runs were identically set to 5000.

For each combination of a method and a target distribution, we report two kinds of results. First, we report estimated $\log Z(\theta^B)$ as a function of K to see a trade-off between computational burden and estimation accuracy. Second, we show the ESS as a function of K to compare the reliability of estimates. As mentioned earlier, the ESS can be misleading when AIS fails to allocate samples to large modes of a target distribution. Nevertheless, we believe that this statistic is reliable in almost all cases because the estimated partition functions are near to the true value (for tractable GRBMs) or roughly coincide with each other in many cases. Even when estimates seem to be unreliable (e.g., results of **AISMean** for CD1(200) for small K), the corresponding ESSs are small and thus consistent to the estimation reliability.

We used GRBMs trained on 100,000 of 6×6 color (i.e., 3 channels) image patches extracted from the CIFAR-10 dataset.[12] Thus the number of visible variables was 108 for all GRBMs. The image patches were contrast-normalized and whitened before training. GRBMs were trained through 80,000 parameter updates with three methods: (CD1) contrastive divergence (CD)[6] with one transition; (CD25) CD with 25 steps of transitions; and (PCD) persistent contrastive divergence.[13]

We first evaluated the three methods for designing initial distributions on GRBMs with only 20 hidden units. The partition functions of such GRBMs can be exactly computed by exhaustive summation over all 2^{20} hidden configurations.

The results are shown in Figures 4 and 5. While none of the three methods severely underestimated/overestimated the log partition functions, the choice of method critically affected the variances of estimates. AISM showed greater variances than the other two methods in almost all conditions. AIS showed the same or slightly smaller variances than AISM. AISC clearly showed smaller variances than these two conventional methods and returned more accurate estimates. The plots of ESS were consistent to these observations: AISC achieved greater ESS than AIS and AISM.

Full-size, intractable GRBMs: We next evaluated the methods on intractable GRBMs with 200 hidden units. The results are shown in Figures 6 and 7. The largest error was given by AISM for CD1(200) for $K = 100$, which underestimated the best estimate (given by AISC for large K) by nearly 1 nat. This estimate and the one by AISM for CD1(200) for $K = 300$ exhibited especially great variances in AIS weights, which caused the lower bounds of the estimates to be negative. Under almost all the conditions, AISM showed greater variances than AIS and AISC. Like the results for the tractable GRBMs, AISC produced smaller estimation variances than AIS in most conditions. The plots of ESS were consistent to the variances as well. We therefore conclude that (1) the variances or covariances of initial distributions critically dominate the estimation accuracy of AIS, and (2) AIS starting from initial distributions that approximate the covariances of target distributions gives more accurate estimates than conventional approaches.

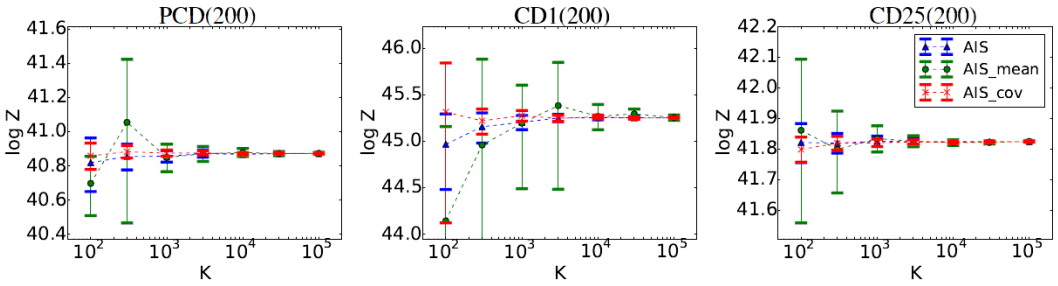

Figure 6: Estimated $\log Z(\theta^B)$ for tractable **GRBMs**. Error bars show $\pm 3\theta$ intervals as in Figure 4.

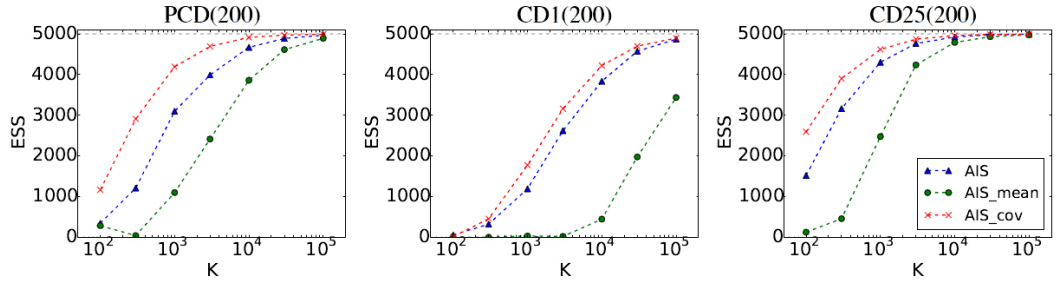

Figure 7: ESS for tractable GRBMs

6 Conclusion

We have proposed an algorithm for designing the covariance matrix of an initial distribution for estimating a GRBM partition function via AIS. We have empirically evaluated the estimation accuracy for tractable and intractable GRBMs and compared the proposed method with conventional ones. We have observed that the covariances of initial distributions have a significant impact on the estimation accuracy, and our proposed method outperformed the conventional methods under almost all the conditions in our experiments.

Acknowledgements

This research is supported by JSPS Grant-in-Aid for JSPS Fellows (145500000159). We thank Motoki Nagata for helpful comments.

References

[1] A Gelman, X L Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. Statistical science. 1998 May; 13(2): 163–185. http://dx.doi.org/10.1214/ss/1028905934

[2] Radford M Neal. Annealed Importance Sampling. Statistics and Computing. 2001; 11: 125–139. http://dx.doi.org/10.1023/A:1008923215028

[3] Jonathan S Yedidia, William T Freeman, and YairWeiss. Constructing free-energy approximations and generalized belief propagation algorithms. Information Theory, IEEE Transactions on. 2005; 51(7): 2282–2312.

[4] Geoffrey E Hinton, Ruslan Salakhutdinov. Replicated softmax: an undirected topic model. In Advances in Neural Information Processing Systems. 2009; 22: 1607–1614.

[5] Graham W Taylor, Geoffrey E Hinton. Products of hidden Markov models: It takes 1 to tango. In Proceedings of the 25th conference on Uncertainty in artificial intelligence, pages 522–529. AUAI Press, 2009.

[6] Guillaume Desjardins, Aaron Courville, Yoshua Bengio, Pascal Vincent, Olivier Delalleau. Parallel Tempering for Training of Restricted Boltzmann Machines. In AISTATs '10, 2010.

[7] Ruslan Salakhutdinov, Iain Murray. On the quantitative analysis of deep belief networks. In Proceedings of the 25th International Conference on Machine Learning. ACM, July 2008.

[8] Ruslan Salakhutdinov, Hugo Larochelle. Efficient learning of deep Boltzmann machines. In AISTATs '10, pages 693–700, 2010.

[9] Lucas Theis, Sebastian Gerwinn, Fabian Sinz, Matthias Bethge. In all likelihood, deep belief is not enough. The Journal of Machine Learning Research. 2011; 12: 3071–3096.

[10] Guillaume Desjardins, Razvan Pascanu, Aaron Courville, Yoshua Bengio. Metric-Free Natural Gradient for Joint-Training of Boltzmann Machines. In Proceedings of the 1st International Conference on Learning Representations, January 2013.

[11] Roger Grosse, Chris Maddison, Ruslan Salakhutdinov. Annealing Between Distributions by Averaging Moments. In Advances in Neural Information Processing Systems 26, 2013.

[12] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. Neural Computation. 2002 August; 14(8): 1771–1800. http://dx.doi.org/10.1162/089976602760128018

[13] Radford M Neal. Annealed importance sampling. Statistics and Computing. 2001; 11(2): 125–139. http://dx.doi.org/10.1023/A:1008923215028

[14] Yosihiko Ogata. A Monte Carlo method for high dimensional integration. Numerische Mathematik. 1989; 55(2): 137–157. http://dx.doi.org/10.1007/BF01406511

[15] Radford M Neal. Sampling from multimodal distributions using tempered transitions. Statistics and Computing. 1996; 6(4): 353–366.

[16] Jascha Sohl-Dickstein, Benjamin J Culpepper. Hamiltonian Annealed Importance Sampling for partition function estimation. Technical Report, Redwood Center, UC Berkeley, 2012.

[17] Ruslan Salakhutdinov, Geoffrey Hinton. Deep Boltzmann machines. In Proceedings of the 12th International Conference on Artificial Intelligence and Statistics. 2009: 448–455.

[18] Tom Minka. Divergence measures and message passing. Technical Report TR-2005-173, Microsoft Research, 2005.

[19] Marc'Aurelio Ranzato, Geoffrey E Hinton. Modeling pixel means and covariances using factorized third-order Boltzmann machines. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pages 2551–2558. IEEE, 2010.

[20] Alex Krizhevsky. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, 2009.

[21] Tijmen Tieleman. Training restricted Boltzmann machines using approximations to the likelihood gradient. In Proceedings of the 25th International Conference on Machine Learning, pages 1064–1071. ACM, July 2008.