On Importance of Base Model Covariance for Annealing Gaussian RBMs

Taichi Kiwaki
Graduate School of Engineering
The University of Tokyo
kiwaki@sat.t.u-tokyo.ac.jp

Kazuyuki Aihara
Institute of Industrial Science
The University of Tokyo
aihara@sat.t.u-tokyo.ac.jp

Abstract

In this paper, we investigate the effects that the covariance of a base distribution has on the accuracy of annealed importance sampling (AIS) estimates for Gaussian restricted Boltzmann machines (RBMs). A common choice for an AIS base model is a Gaussian RBM (GRBM) with zero weight connections. Such a base model 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 a base distribution for GRBMs. We empirically analyze the effect of the base model covariance on the estimation accuracy of AIS. The proposed method for designing base models outperforms conventional methods under various conditions.

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, 2, 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 base distribution to estimate the statistics of the intractable target distribution. For restricted Boltzmann machines (RBMs), a common choice for the base distribution is an RBM with zero weight connections [4, 5]. Particularly for Gaussian RBMs (GRBMs), the base 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 a base model.

In this paper, we propose an AIS algorithm for GRBMs in which a base 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 the covariance matrix of a base 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 $\mathbf{v} \in \mathbb{R}^D$ and binary hidden variables $\mathbf{h} \in \{0,1\}^M$. 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 $\theta = \{W, \mathbf{a}, \mathbf{b}, \boldsymbol{\sigma}\}$ 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

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

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

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

3 Annealed Importance Sampling

```
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 $\hat{Z}(\theta^B) = Z(\theta^A) \sum_i^N w^{(i)}/N$
```

Algorithm 1: AIS

Suppose that we have a distribution defined on some space $\mathcal{V}$ 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 $\mathbf{v} \in \mathcal{V}$, 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 annealed importance sampling (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 $k \in \{0, \cdots, K\}$, where the starting point of the path $p_0(\mathbf{v}) = p_A(\mathbf{v})$ is the tractable base 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.

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

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

where $p_\beta(\cdot)$ is a continuously parameterized probability mass or density function by $\beta \in [0,1]$ 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. 4 where the interval $[0, 1]$ is partitioned by a monotonic sequence $\{\beta_k\}$ $(k = 0, \cdots, K)$ and $p_{\beta_k}(\cdot)$ is substituted by $p_k(\cdot)$. Although the equality of Eq. 4 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 do not return independent samples [7]. However, in practice, the variance of AIS estimates can be quite large depending on several factors. First, as Neal [7] suggests, poor mixing of T_k can damage the estimation accuracy. Recently, Sohl-Dickstein and Culpepper [10] 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 *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 is widely implemented [4, 11]. 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 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 $w_*^{(i)} = N w^{(i)} / \sum_{i=1}^N w^{(i)}$. 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.

3.1 AIS for GRBMs

Suppose we are to estimate the partition function of a GRBM with parameters $\theta^B = \{W^B, \mathbf{a}^B, \mathbf{b}^B, \boldsymbol{\sigma}^B\}$ via AIS by using another GRBM with parameters $\theta^A = \{W^A, \mathbf{a}^A, \mathbf{b}^A, \boldsymbol{\sigma}^A\}$ as a base model. Because the MCMC operators for intermediate distributions for 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 a base distribution has zero weight, i.e., $W_{ij}^A = 0$. Because the intermediate distributions defined as Eq. 7 are also GRBMs, we can easily evaluate the log of the unnormalized density as:

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

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

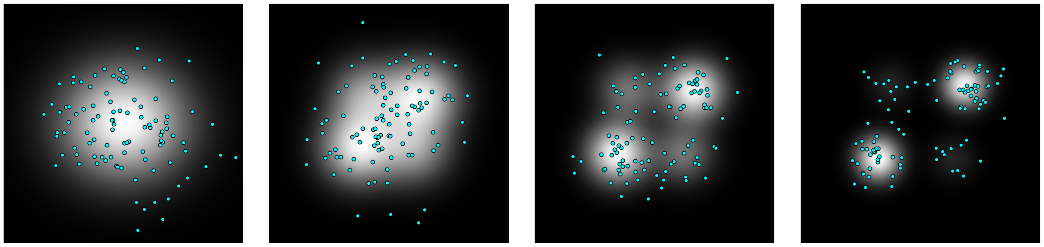
$$p_\beta(h_i = 1|\mathbf{v}) = \mathrm{Sigm}\left(\beta(\sum_{j=1}^{D} W_{ij}^B \frac{v_j}{\sigma_j^B} + a_i^B) \right), \quad p_\beta(v_j|\mathbf{h}) \ = \mathcal{N}\left(v_j \middle| m_j(\beta, \mathbf{h}),\ \sigma_j^2(\beta) \right), \tag{9}$$

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 base distributions [5]. Therefore, the weight matrix of a base distribution is usually set to zero so that it has only a single mode. Another commonly adopted technique is to choose a base model 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.

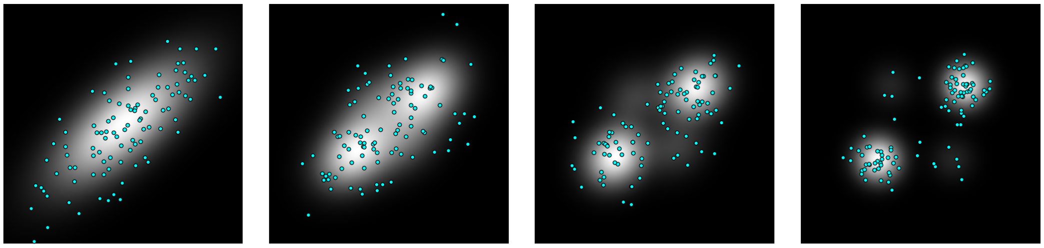
4 Multivariate Normal Distributions for Base Distributions

Base models given in the previous section do not model any dependency between variables; the covariance matrix of a base 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 base distributions with any covariance matrix.



$\beta = 0.0$ 0.26 0.46 1.0

Figure 1: Heatmaps (white denotes large) of annealed distributions p_β by the **conventional** method (labeled as **AIS**). Corresponding values of beta are printed below. Points (best viewed in color) are the sample points of AIS.



$\beta = 0.0$ 0.26 0.46 1.0

Figure 2: Heatmaps as Fig. 1 by the **proposed** method (**AIS_COV**).

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

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

where $\mathbf{x} \in \mathbb{R}^D$ 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 base model), 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 are also normally distributed. Based on these observations, we can gain the log of the unnormalized density of intermediate distributions as:

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

where we define $\Lambda^B = \mathrm{diag}({\sigma_1^B}^{-2}, \cdots, {\sigma_D^B}^{-2})$. Equation 11 shows that the covariance matrix and means of the base model are $\Sigma^A = \Lambda^{-1} + {\Lambda^B}^{-1}$ and $\mathbf{b}^A$. Conversely, we can design the base model to have a covariance matrix Σ^A by setting $\Lambda = (\Sigma^A - {\Lambda^B}^{-1})^{-1}$.

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:

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

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

and

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

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 1 and 2 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. Base models have the same moments (up to 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.

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 a base distribution, we can minimize the second factor. Gelman and Meng [1] derived a lower bound for this factor and showed that the optimal base 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 α-divergence to see the relationship between our strategy and the optimal one [12]. α-divergence forms a family of divergences parameterized by a scalar parameter α, and includes both KL divergence and Hellinger distance as its instances. Hellinger distance corresponds to $\alpha = 0.5$ and KL divergence corresponds to $\alpha = 1$. This suggests that our strategy approximates the optimal strategy by minimizing α-divergence of $\alpha = 1$ instead of the optimal value $\alpha = 0.5$.

5 Experiments

In our experiments, we compare the following methods for designing base distributions: (**AIS**), the conventional method with Eq. 7 where $\mathbf{b}^A$ and $\boldsymbol{\sigma}^A$ are identically chosen as the target distribution; (**AIS_MEAN**), another baseline method based on Eq. 7 where $\mathbf{b}^A$ is determined from the target distribution but $\boldsymbol{\sigma}^A$ is simply determined as $\boldsymbol{\sigma}^A = \boldsymbol{\sigma}^B$; and (**AIS_COV**), the proposed method with Eq. 10 where $\mathbf{b}^A$ and Λ are chosen according to the target distribution statistics[1]. For all three methods, we approximate 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 are discarded as burn-in samples. Estimation is made with a various number of intermediate distributions K. The annealing schedule $\{\beta_k\}$ is divided into four periods: β_k uniformly spaced from 0 to 0.1, β_k

[1] Note that we examine **AIS_MEAN** as well as **AIS** and **AIS_COV** to better illustrate the impacts that the covariance matrix of the base distribution can exert on the estimation accuracy.

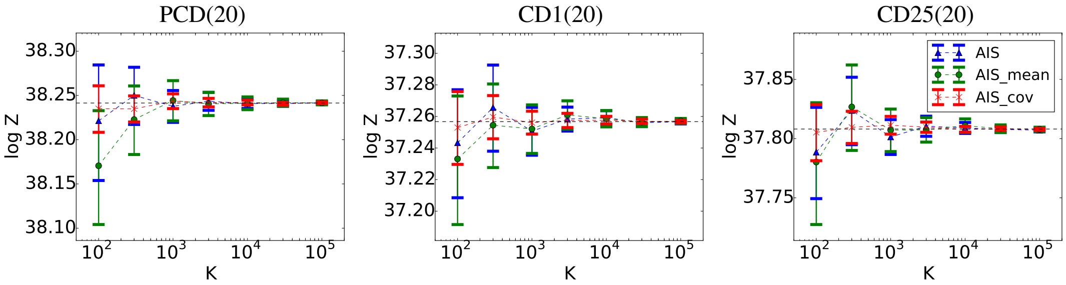


Figure 3: Estimated $\log Z(\theta^B)$ for tractable GRBMs. Error bars show $\pm 3\sigma$ intervals where σ^2 is the sample variance of the AIS estimate.

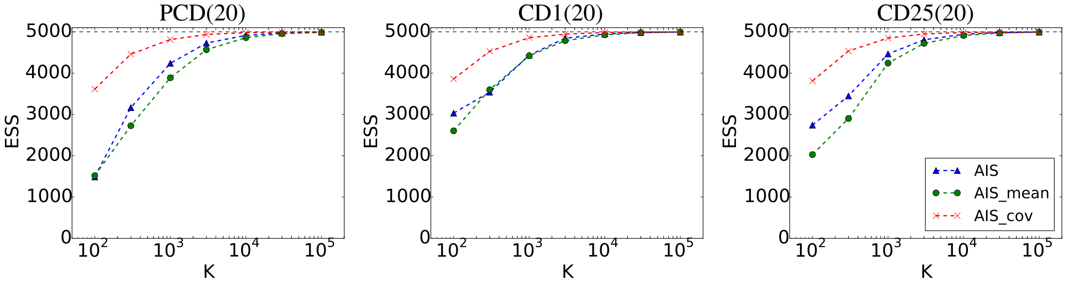


Figure 4: ESS (the larger the better) for tractable GRBMs.

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 are assigned to all of these periods. The numbers of AIS runs are identically set to 5000.

For each combination of a method and a target distribution, we show 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 compute 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 **AIS_MEAN** for CD1(200) for small K), corresponding the 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 [13]. 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 [14].

Small, tractable GRBMs: We first evaluated the three methods for designing base 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 Figs. 3 and 4. While none of the three methods severely underestimated/overestimated the log partition functions, the choice of method critically affected the variances of estimates. **AIS_MEAN** showed greater variances than the other two methods in almost all conditions. **AIS** showed the same or slightly smaller variances than **AIS_MEAN**. **AIS_COV** clearly showed smaller variances than these two conventional methods and returned more accurate estimates. The plots of ESS were consistent to these observations: **AIS_COV** achieved greater ESS than **AIS** and **AIS_MEAN**.

Full-size, intractable GRBMs: We next evaluated the methods on intractable GRBMs with 200 hidden units. The results are shown in Figs. 5 and 6. The largest error was given by **AIS_MEAN**

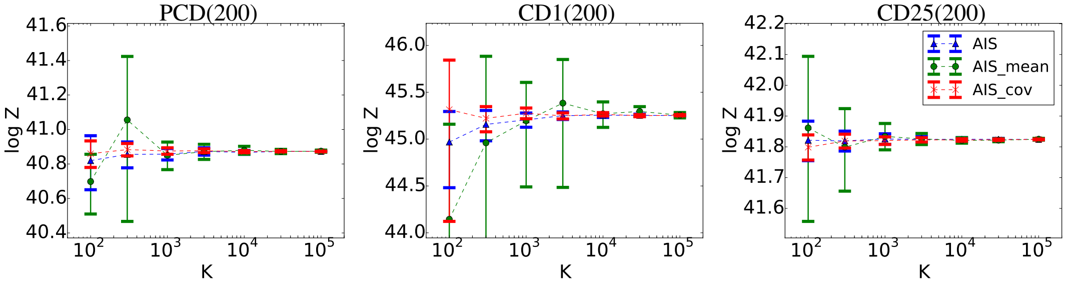


Figure 5: Estimated $\log Z(\theta^B)$ for intractable GRBMs. Error bars show $\pm 3\sigma$ intervals as in Fig. 3

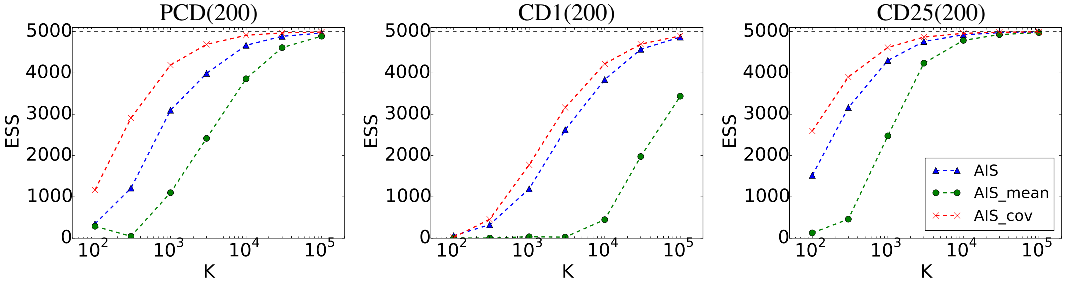


Figure 6: ESS for intractable GRBMs.

for CD1(200) for $K = 100$, which underestimated the best estimate (given by **AIS_COV** for large K) by nearly 1 nat. This estimate and the one by **AIS_MEAN** 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 conditions, **AIS_MEAN** showed greater variances than **AIS** and **AIS_COV**. Like the results for tractable GRBMs, **AIS_COV** 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 base models critically dominate the estimation accuracy of AIS, and (2) AIS starting from the base models that model covariances of target distributions gives more accurate estimates than conventional approaches.

6 Conclusion

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

References

[1] A Gelman and X L Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, 13(2):163–185, May 1998.

[2] Radford M Neal. Annealed Importance Sampling. *Statistics and Computing*, 11:125–139, 2001.

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

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

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

[6] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, August 2002.

[7] Radford M Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

[8] Yosihiko Ogata. A Monte Carlo method for high dimensional integration. *Numerische Mathematik*, 55(2):137–157, 1989.

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

[10] Jascha Sohl-Dickstein and Benjamin J Culpepper. Hamiltonian Annealed Importance Sampling for partition function estimation. *arXiv.org*, May 2012.

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

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

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

[14] 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.