

An Analysis of Contrastive Divergence Learning in Gaussian Boltzmann Machines

Christopher K. I. Williams and Felix V. Agakov

Division of Informatics, University of Edinburgh, Edinburgh EH1 2QL, UK

c.k.i.williams@ed.ac.uk, felixa@dai.ed.ac.uk

http://anc.ed.ac.uk

May 17, 2002

Abstract

The Boltzmann machine (BM) learning rule for random field models with latent variables can be problematic to use in practice. These problems have (at least partially) been attributed to the negative phase in BM learning where a Gibbs sampling chain should be run to equilibrium. Hinton (1999, 2000) has introduced an alternative called contrastive divergence (CD) learning where the chain is run for only 1 step. In this paper we analyse the mean and variance of the parameter update obtained after i steps of Gibbs sampling for a simple Gaussian BM. For this model our analysis shows that CD learning produces (as expected) a biased estimate of the true parameter update. We also show that the variance does usually increase with i and quantify this behaviour.

Recently Hinton (1999, 2000) has introduced the *contrastive divergence* (CD) learning rule. This was introduced in the context of Products of Experts architectures, although it is a general learning algorithm for random field models. The idea is that instead of using the negative phase of Boltzmann machine (BM) learning (which in theory requires running a Gibbs sampler to equilibrium), a smaller number of Gibbs sampling iterations should be used (e.g. 1). The contribution of this paper is to analyse the CD learning rule for an arbitrary number $i > 0$ of Gibbs sampling iterations for a simple Gaussian Boltzmann machine. This allows us to compare the mean and variance of the $CD(i)$ update with the BM update. In a nutshell, we find that the bias of the $CD(i)$ update decreases with i , while the variance of the update increases with i (although this latter conclusion depends on exactly how the learning rule is implemented).

The structure of the paper is as follows: in section 1 we introduce binary and Gaussian Boltzmann machines, and the BM and CD learning rules. In section 2 we first introduce a simple Gaussian BM and then calculate the mean and variance of the parameter update as a function of i , the number of Gibbs sampling iterations. Finally, in section 3 we briefly describe extension of the results to the case of multivariate Gaussian Boltzmann machines.

In this section we analyse in detail the properties of the CD(i) learning rule for the 1-hidden 1-visible Gaussian Boltzmann machine. In section 3 we consider the general multivariate case, where we obtain more general but less strong results than in the specific case.

2.1 Gibbs Sampling for a Gaussian Boltzmann Machine

Let \mathbf{y} be partitioned as $\mathbf{y}^T = (\mathbf{y}_1^T, \mathbf{y}_2^T)$, and the corresponding partition of the covariance matrix \mathbf{C} of the joint Gaussian be

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix}. \quad (8)$$

The conditional $p(\mathbf{y}_1|\mathbf{y}_2)$ can be expressed as

schur complement

$$p(\mathbf{y}_1|\mathbf{y}_2) \sim N(\boldsymbol{\mu}_1 + \mathbf{C}_{12}\mathbf{C}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \mathbf{C}_{11} - \mathbf{C}_{12}\mathbf{C}_{22}^{-1}\mathbf{C}_{21}) \quad (9)$$

where $(\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T)^T$ is the mean vector of the Gaussian (see e.g. pp. 226-228 of Mises 1964). From (7) and (9) it is easy to show that for the 1-visible 1-hidden variable model

$$p(x|z) \sim N(\sigma z, \tau^2), \quad p(z|x) \sim N(\sigma x, \tau^2), \quad (10)$$

with the auxiliary parameters given by $\sigma \stackrel{\text{def}}{=} w/a$ and $\tau^2 \stackrel{\text{def}}{=} (1 - \sigma^2)a$. For the covariance of the complete model to be positive-definite, σ should belong to the open interval $(-1, 1)$.

Let $v_1, v_2, \dots, u_0, u_1, \dots$ correspond to mutually independent $N(0, 1)$ random variables, i.e. $\langle u_i v_k \rangle = 0$, $\langle u_i u_k \rangle = \langle v_i v_k \rangle = \delta_{ik}$, where δ_{ik} is the Kronecker delta. From (10) we have

$$z_i = \sigma x_i + \tau u_i, \quad i \geq 0, \quad x_j = \sigma z_{j-1} + \tau v_j, \quad j \geq 1. \quad (11)$$

We are interested in the CD(i) parameter update which is proportional to

$$\Delta_{0i} = x_0 z_0 - x_i z_i. \quad (12)$$

This is a random quantity, which depends not only on the u 's and v 's in the Gibbs sampling chain, but also on the random choice of x_0 . Below we calculate the mean conditional on x_0 i.e. $\langle \Delta_{0i} | x_0 \rangle$ and the unconditional mean $\langle \Delta_{0i} \rangle = \langle \langle \Delta_{0i} | x_0 \rangle \rangle_{x_0}$, where $\langle \dots \rangle_{x_0}$ denotes expectation over the data distribution $p(x_0)$. We also calculate the conditional variance $\text{var}(\Delta_{0i} | x_0)$ and the unconditional variance $\text{var}(\Delta_{0i})$.

Of course for a Gaussian Boltzmann machine it is not necessary to use the Boltzmann machine or CD(i) learning rules to adapt the parameters W , one can simply use matrix inversion and analytic derivatives of the likelihood. However, our aim is to investigate these learning rules and the Gaussian model is an interesting one in which exact analysis can be carried out.

2.2 Calculation of the Mean $\langle \Delta_{0i} \rangle$

Expression (11) leads to

$$x_i z_i = \sigma x_i^2 + \tau u_i x_i \Rightarrow \langle x_i z_i | x_0 \rangle = \sigma \langle x_i^2 | x_0 \rangle. \quad (13)$$

It can be shown (see Appendix A) that $\langle x_i^2 | x_0 \rangle = \sigma^{4i}(x_0^2 - a) + a$, thus

$$\langle x_i z_i | x_0 \rangle = \sigma^{4i+1}(x_0^2 - a) + a\sigma. \quad (14)$$

Therefore

$$\langle \Delta_{0i} | x_0 \rangle = \sigma(x_0^2 - a)(1 - \sigma^{4i}) \quad (15)$$

and

$$\langle \Delta_{0i} \rangle = \langle \langle \Delta_{0i} | x_0 \rangle \rangle_{x_0} = \sigma(a_t - a)(1 - \sigma^{4i}), \quad (16)$$

where a_t is the variance of the data.

2.3 Comparison with Boltzmann Learning

From (16) we can compute the average weight update of Boltzmann learning $\langle \Delta^{BM} \rangle$:

$$\langle \Delta^{BM} \rangle = \langle \langle x_0 z_0 | x_0 \rangle - \langle x_\infty z_\infty \rangle \rangle_{x_0} = \sigma(a_t - a). \quad (17)$$

We can further notice that since $|\sigma| < 1$ then

$$\lim_{i \rightarrow \infty} |\langle \Delta_{0i} \rangle| = \lim_{i \rightarrow \infty} |\sigma(a_t - a)|(1 - \sigma^{4i}) = |\sigma(a_t - a)|. \quad (18)$$

Thus, in the considered model the gradient of the log-likelihood in i -step contrastive divergence learning $|\partial \mathcal{L}^{(i)} / \partial \omega|$ underestimates the absolute value of the gradient of the log-likelihood of the Boltzmann learning rule $|\partial \mathcal{L}^{BM} / \partial \omega|$, but asymptotically approaches to it as the number of Gibbs sampling iterations i increases, see Figure 2. Moreover, it is easy to see from (16) and (17) that for both BM and CD(i) learning, the optimal choice of ω leads to $a = a_t$. This is an expected result, since the Gaussian Boltzmann machine can perfectly fit the training distribution $N(0, a_t)$. Note also that $\langle \Delta_{0i} \rangle$ has the same sign as $\langle \Delta^{BM} \rangle$.

2.4 Calculation of the Variance $var(\Delta_{0i})$

We first calculate the conditional variance $var(\Delta_{0i} | x_0)$ due to stochasticity of the Gibbs sampling and then calculate the unconditional variance $var(\Delta_{0i})$.

There are two different situations that we can analyse, depending on whether or not two different chains are run to calculate equation 12, i.e. that the sample z_0 used in the negative phase of the learning rule is distinct from the sample used in the positive phase for calculating $x_0 z_0$. For the case that two chains are used (call this case I, where I stands for independent), we have

$$var(\Delta_{0i} | x_0) = var(x_0 z_0 | x_0) + var(x_i z_i | x_0). \quad (19)$$

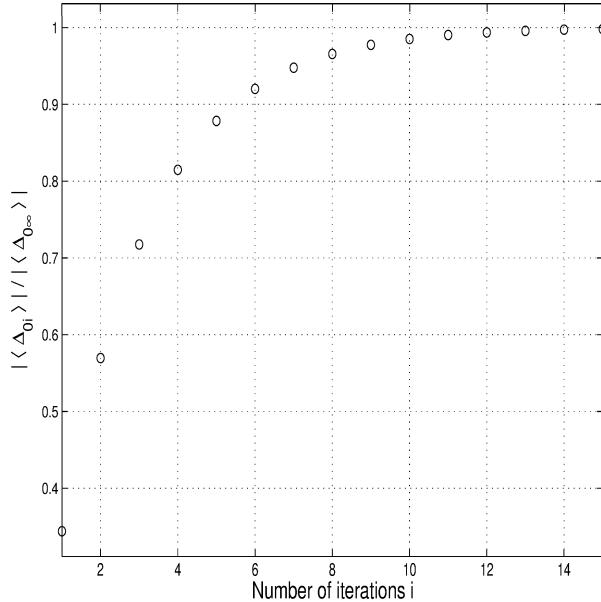


Figure 2: Plot of $|\langle \Delta_{0i} \rangle| / |\langle \Delta_{0\infty} \rangle|$ against iteration number i for $\sigma = 0.9$.

It is easy to show that $\text{var}(x_0 z_0 | x_0) = \tau^2 x_0^2 = (1 - \sigma^2) a x_0^2$ and thus

$$\text{var}(\Delta_{0i} | x_0) = (1 - \sigma^2) a x_0^2 + \langle (x_i z_i)^2 | x_0 \rangle - (\langle x_i z_i | x_0 \rangle)^2.$$

If only one chain is used (call this case D, D for dependent) then (19) must be corrected by a term $-2\text{cov}(x_0 z_0, x_i z_i | x_0)$. Analysis shows that $\text{cov}(x_0 z_0, x_i z_i | x_0) = 2x_0^2 a(1 - \sigma^2) \sigma^{4i}$. In the remainder of this section our derivations are made for case I; expressions for case D can be obtained by including this extra term. The expression for $\langle x_i z_i | x_0 \rangle$ in the r.h.s. of (20) has been derived in (14). Expanding $\langle (x_i z_i)^2 | x_0 \rangle$, we obtain

$$\begin{aligned} \langle (x_i z_i)^2 | x_0 \rangle &= \sigma^2 \langle x_i^4 | x_0 \rangle + 2\sigma\tau \langle x_i^3 u_i | x_0 \rangle \\ &\quad + \tau^2 \langle u_i^2 x_i^2 | x_0 \rangle \\ &= \sigma^2 \langle x_i^4 | x_0 \rangle + \tau^2 \langle x_i^2 | x_0 \rangle. \end{aligned} \quad (20)$$

After some simplifications shown in Appendix A, (19) can be re-written as

$$\begin{aligned} \text{var}(\Delta_{0i} | x_0) &= 2a\sigma^2(a - 2x_0^2)k^2 - 3\sigma^2 a(a - x_0^2)k \\ &\quad - a(a - x_0^2)k + a^2(1 + \sigma^2) + (1 - \sigma^2)a x_0^2, \end{aligned} \quad (21)$$

where $k = \sigma^{4i} \subset (0, 1]$. The unconditional variance $\text{var}(\Delta_{0i})$ may be expressed as

$$\int [(\langle \Delta_{0i} | x_0 \rangle - \langle \Delta_{0i} \rangle)^2 + \text{var}(\Delta_{0i} | x_0)] p(x_0) dx_0. \quad (22)$$

By applying (22) to (15) and (16) and performing some manipulations, we can express the unconditional variance of the parameter update as

$$\text{var}(\Delta_{0i}) = \langle \text{var}(\Delta_{0i} | x_0) \rangle_{x_0} + \sigma^2 (\langle x_0^4 \rangle - a_t^2)(1 - k)^2. \quad (23)$$

By averaging (21) over $p(x_0)$ and using $\langle x_0^4 \rangle = 3a_t^2$ for a Gaussian target distribution we obtain

$$\begin{aligned} \text{var}(\Delta_{0i}) &= 2\sigma^2(a - a_t)^2 k^2 - [a(a - a_t)(1 + 3\sigma^2) + 4\sigma^2 a_t^2]k \\ &\quad + a^2(1 + \sigma^2) + (1 - \sigma^2)a a_t + 2\sigma^2 a_t^2. \end{aligned} \quad (24)$$

2.5 Behaviour of $\text{var}(\Delta_{0i})$ as a function of i

Here we investigate the behaviour of the variance of the $\text{CD}(i)$ update term for the given model as a function of the the number of Gibbs sampling iterations.

Note that (24) is a quadratic in k , say $Ak^2 + Bk + C$. As $k = \sigma^{4i}$ we note that as i increases from 1, k will vary from σ^4 towards 0 (as $|\sigma| < 1$). Hence the behaviour of the variance as a function of i depends on the parameters A and B in the quadratic. Note that $A > 0$ and thus that we have a quadratic bowl whose minimum falls at $k^* = -B/2A$. If $k^* \geq \sigma^4$ then the variance will increase monotonically with i . Conversely if $k^* \leq 0$ the variance will decrease monotonically with i , but if $0 < k^* < \sigma^4$ then there can be non-monotonic behaviour, with the variance first rising then falling. Which behaviour is obtained will depend on the values of the parameters a_t , a and σ .

There are many quantities that we might examine, e.g. the conditional and unconditional variances as a function of i , for both cases I and D. We first focus on the unconditional variance for case I. We can show that $a \geq a_t$ is a sufficient (although not necessary) condition for this quantity to increase monotonically. Consider the specific case of $a_t = 1$ and $\alpha = 2$; here increasing $|\omega|$ away from 0 causes a to increase. For $|\omega| \lesssim 0.5524$ the variance decreases as a function of i , although this decrease is very small and $\text{var}(\Delta_{0i})$ is in fact almost constant. (For example, for $a_t = 1$, $\alpha = 2$, and $|\omega| = 0.5$, the drop is from ≈ 0.92740 on iteration 1 to ≈ 0.92722 for subsequent iterations.) For $|\omega| \gtrsim 0.5528$ the variance increases monotonically with i . In the intermediate region the behaviour is non-monotonic (although almost constant). For $|\omega| = \sqrt{2}$ (corresponding to $a = a_t = 1$) the increase in variance with iteration number i is plotted in Figure 3(a). Note that for small $|\omega|$ there is weak coupling between the hidden and visible variables which explains the almost-constant behaviour of the variance. For reasonably large $|\omega|$ the variance $\text{var}(\Delta_{0i})$ increases significantly with i .

For case D, i.e. when a single chain is used for both positive and negative stages of learning, it can be shown that the unconditional variance increases monotonically as a function of i for all attainable values of the parameters a_t , a and σ . It is also worth noting that for both cases I and D, $\langle \text{var}(\Delta_{0i}|x_0) \rangle_{x_0}$ can display non-monotonic or decreasing behaviour of relatively large magnitude (see Figure 3(b) and Figure 3(c)). This suggests that variation of the variance of the parameter update with the number of iterations is strongly influenced by the exact learning rule used.

In all cases these analytical results have been confirmed by experiments using many Gibbs sampling runs.

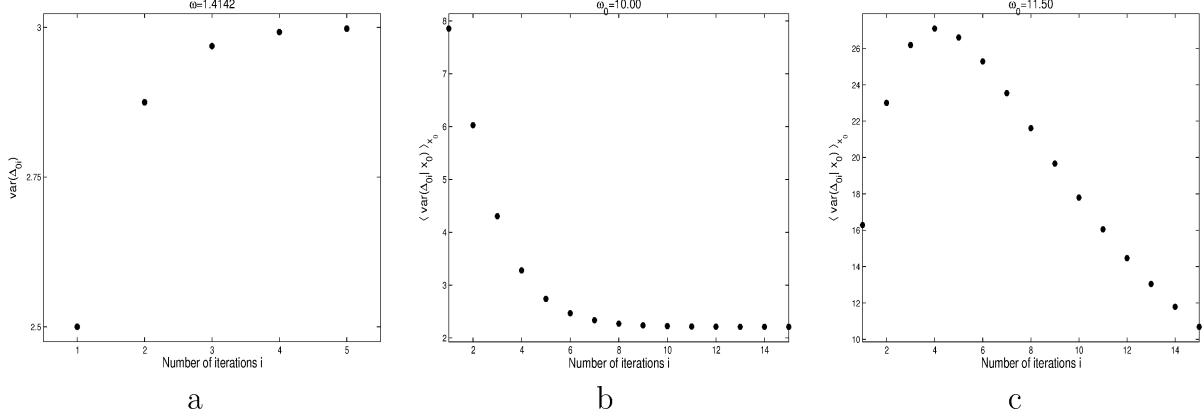


Figure 3: (a) Plot of $\text{var}(\Delta_{0i})$ for case I as a function of i for $a_t = 1$, $\alpha = 2$ and $\omega = \sqrt{2}$. (b) Plot of the conditional variance $\langle \text{var}(\Delta_{0i}|x_0) \rangle_{x_0}$ for case I as a function of i for $a_t = 25$, $\alpha = 12$ and $\omega = 10$. (c) Plot of the conditional variance $\langle \text{var}(\Delta_{0i}|x_0) \rangle_{x_0}$ for case I as a function of i for $a_t = 25$, $\alpha = 12$ and $\omega = 11.5$.

2.6 Quantification of the Contrastive Divergence Approximation

The CD(i) learning rule discards a term in the expression for the gradient of the log-likelihood. Here we quantify the CD(i) approximation of the gradient of the log-likelihood for the case of a simple Gaussian Boltzmann machine defined above.

Let $Q_0(\mathbf{x}) \stackrel{\text{def}}{=} p(\mathbf{x}_0)$ and $Q_i(\mathbf{x}) \stackrel{\text{def}}{=} p(\mathbf{x}_i)$ be the data distribution and the distribution of the visible variables after their i -step reconstruction. It is easy to see that maximization of the likelihood $Q_\infty(\mathbf{x}) \stackrel{\text{def}}{=} p(\mathbf{x})$ under the model is equivalent to minimization of the KL divergence $KL(Q_0(\mathbf{x})\|Q_\infty(\mathbf{x}))$ between the data and the model as

$$KL(Q_0\|Q_\infty) = -H(Q_0) - \langle \log(Q_\infty) \rangle_{Q_0}, \quad (25)$$

where $H(Q_0)$ is the empirical entropy. Clearly the free energy term is in general difficult to compute [see expressions (1) and (2)].

Let $\mathcal{L}^{(i)}$ be the i -step estimate of the log-likelihood, defined as

$$\mathcal{L}^{(i)} = KL(Q_i\|Q_\infty) - KL(Q_0\|Q_\infty) - H(Q_0) \quad (26)$$

$$\begin{aligned} &= \int Q_0(\mathbf{x}) \log Q_\infty(\mathbf{x}) d\mathbf{x} - H(Q_i) \\ &\quad - \int Q_i(\mathbf{x}) \log Q_\infty(\mathbf{x}) d\mathbf{x}. \end{aligned} \quad (27)$$

Notice that $\lim_{i \rightarrow \infty} \mathcal{L}^{(i)} = \mathcal{L}$. The gradient of $\mathcal{L}^{(i)}$ is given by

$$\nabla_\omega \mathcal{L}^{(i)} = \langle \Delta_{0i} \rangle - \frac{\partial H(Q_i(\mathbf{x}))}{\partial \omega} - \int \frac{\partial Q_i(\mathbf{x})}{\partial \omega} \log Q_\infty(\mathbf{x}) d\mathbf{x}. \quad (28)$$

Here $\langle \Delta_{0i} \rangle$ is the CD(i) parameter update given by

$$\langle \Delta_{0i} \rangle = \left\langle \frac{\partial \log \mathcal{L}}{\partial \omega} \right\rangle_{Q_0} - \left\langle \frac{\partial \log \mathcal{L}}{\partial \omega} \right\rangle_{Q_i}. \quad (29)$$

For the Gaussian Boltzmann machine considered in section 2.1 $Q_\infty(x) \sim N(0, a)$ and $Q_i \sim N(0, \sigma_i^2)$, where $\sigma_i^2 = \sigma^{4i}(a_t - a) + a$. The variance a of the data under the model changes over time as the CD(i) updates are performed and approaches a_t if the learning rule is set up correctly.

Let $\epsilon_i \stackrel{\text{def}}{=} \nabla_\omega \mathcal{L}^{(i)} - \langle \Delta_{0i} \rangle$, the term discarded from the gradient (28) under the CD(i) learning. From (28) we obtain

$$\begin{aligned} \epsilon_i = & \frac{-\partial H(Q_i(x))}{\partial \omega} + \frac{1}{\sqrt{2\pi}\sigma_i^2} \frac{\partial \sigma_i}{\partial \omega} \int \exp \left\{ -\frac{x^2}{2\sigma_i^2} \right\} \times \\ & \left[1 - \frac{x^2}{\sigma_i^2} \right] \left(-\frac{x^2}{2a} - \frac{\log 2\pi a}{2} \right) dx. \end{aligned} \quad (30)$$

Analytic expressions for the Gaussian integrals in the r.h.s. of equation (30) are well known: if

$$I_n \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} \exp \left\{ -\frac{x^2}{2\sigma_i^2} \right\} x^n dx \quad (31)$$

then $I_0 = \sigma_i \sqrt{2\pi}$, $I_2 = \sigma_i^3 \sqrt{2\pi}$, $I_4 = 3\sigma_i^5 \sqrt{2\pi}$. Also, the entropy of a Gaussian $\sim N(0, \sigma_i^2)$ is $\frac{1}{2} \log(2\pi e \sigma_i^2)$. Substituting these expressions into (30) and performing some manipulations we get

$$\epsilon_i = \left(-\frac{1}{\sigma_i} + \frac{\sigma_i}{a} \right) \frac{\partial \sigma_i}{\partial \omega}, \quad (32)$$

$$\frac{\partial \sigma_i}{\partial \omega} = \frac{1}{2\sigma_i} \left[((a_t - a)(4i/\omega) + 2a^2\sigma) \sigma^{4i} - 2a^2\sigma \right]. \quad (33)$$

Note that from (32) and the fact that $\lim_{i \rightarrow \infty} \sigma_i^2 = a$ we obtain $\lim_{i \rightarrow \infty} \epsilon_i = 0$ as expected.

In order to analyze importance of the discarded term ϵ_i for evaluation of the gradient $\nabla_\omega \mathcal{L}^{(i)}$ we can consider the ratio between ϵ_i and the mean parameter update of the CD(i) learning. From equations (16) and (32) it follows that

$$\left| \frac{\epsilon_i}{\langle \Delta_{0i} \rangle} \right| = \frac{\sigma^{4i}}{a\sigma_i(1 - \sigma^{4i})} \left| \frac{1}{\sigma} \frac{\partial \sigma_i}{\partial \omega} \right|. \quad (34)$$

As we see from Figure 4, there exist parameter settings such that $|\epsilon_i|$ yields a large contribution to $\nabla_\omega \mathcal{L}^{(1)}$. This is consistent with the experimental results in Hinton (2000, section 10) where quite large deviations can be observed for individual parameters. However, Hinton notes that for networks with several units, the vector $\langle \Delta_{0i} \rangle$ is almost certain to have a positive cosine with $\langle \Delta_{0\infty} \rangle$. Figure 4 also shows that $\lim_{i \rightarrow \infty} |\epsilon_i / \langle \Delta_{0i} \rangle| = 0$ as we would expect.

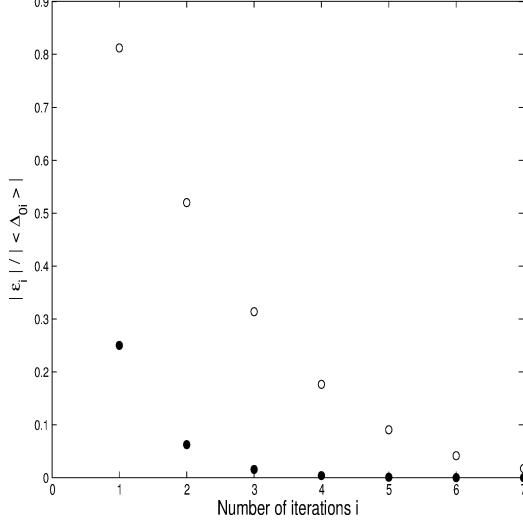


Figure 4: Plot of $|\epsilon_i| / |\langle \Delta_{0i} \rangle|$ as a function of i for $a_t = 25$, $\alpha = 12$, $\omega = 10$ (empty circles) and $a_t = 1$, $\alpha = 2$ and $\omega = \sqrt{2}$ (filled circles). Notice that in the latter case $a = a_t$.

3 Extension To Multivariate Gaussian Boltzmann Machines

In this section we describe general properties of the $CD(i)$ learning for a multivariate Gaussian Boltzmann machine. We give an upper bound on the geometric rates of convergence of the mean and the variance of the parameter update and discuss how the exact convergence rate for the mean can be found.

3.1 Gibbs Sampling for a Multivariate Gaussian Boltzmann Machine

Let Σ and W be the covariance and the inverse covariance (weight) matrix of a Gaussian Boltzmann machine with $|x|$ visible and $|z|$ hidden variables, such that

$$W = \begin{bmatrix} W_{zz} & W_{zx} \\ W_{xz} & W_{xx} \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{zz} & \Sigma_{zx} \\ \Sigma_{xz} & \Sigma_{xx} \end{bmatrix}, \quad W = \Sigma^{-1}. \quad (35)$$

Since both W , $\Sigma \in \mathbb{R}^{(|z|+|x|) \times (|z|+|x|)}$ are symmetric, $W_{zx} = W_{xz}^T$ and $\Sigma_{zx} = \Sigma_{xz}^T$. As in the simple case considered above, learning in a multivariate Gaussian Boltzmann machine assumes adapting the weights W_{zx} between hidden and visible variables so that the covariance matrix Σ_{xx} of the visible variables under the model matches the covariance of the data (as before, we assume that the data is centered at the origin).

Representing the conditional distributions (9) in terms of W (using the partitioned matrix inverse equations, see e.g. Press et al. (1992)[p 77]) we obtain

$$p(z|x) \sim N(-W_{zz}^{-1}W_{zx}x, W_{zz}^{-1}) \quad (36)$$

$$p(x|z) \sim N(-W_{xx}^{-1}W_{xz}z, W_{xx}^{-1}). \quad (37)$$

Equivalently, by analogy with the 1-hidden 1-visible variable case we can expand the Gibbs sampling chain as

$$\mathbf{z}_i = \mathbf{S}\mathbf{x}_i + \mathbf{u}_i \in \mathbb{R}^{|\mathbf{z}|}, i \geq 0 \quad (38)$$

$$\mathbf{x}_j = \mathbf{T}\mathbf{z}_{j-1} + \mathbf{v}_j \in \mathbb{R}^{|\mathbf{x}|}, j \geq 1. \quad (39)$$

Here $\mathbf{S} = -\mathbf{W}_{zz}^{-1}\mathbf{W}_{zx} \in \mathbb{R}^{|\mathbf{z}|\times|\mathbf{x}|}$, $\mathbf{T} = -\mathbf{W}_{xx}^{-1}\mathbf{W}_{xz} \in \mathbb{R}^{|\mathbf{x}|\times|\mathbf{z}|}$, and $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{u}_0, \mathbf{u}_1, \dots$ are mutually independent random variables such that

$$\mathbf{u}_i \sim N(0, \mathbf{W}_{zz}^{-1}), \quad \mathbf{v}_j \sim N(0, \mathbf{W}_{xx}^{-1}). \quad (40)$$

For the i -step learning rule the parameter update is given by

$$\Delta_{0i} = \mathbf{z}_0\mathbf{x}_0^T - \mathbf{z}_i\mathbf{x}_i^T \in \mathbb{R}^{|\mathbf{z}|\times|\mathbf{x}|}, \quad (41)$$

where a (k, j) element Δ_{0i}^{kj} of Δ_{0i} corresponds to the weight update between the k^{th} hidden and the j^{th} visible unit.

3.2 Geometric Convergence of $\langle \Delta_{0i} \rangle$ and $var(\Delta_{0i})$

Let $\langle \Delta_{0i} \rangle = \left\{ \langle \Delta_{0i}^{kj} \rangle \right\}$ and $var(\Delta_{0i}) = \left\{ var(\Delta_{0i}^{kj}) \right\}$ for $k = 1 \dots |h|$, $j = 1 \dots |v|$. Note that each element Δ_{0i}^{kj} is a function of the chain variables \mathbf{x}_i and \mathbf{z}_i . We may hope to understand dependence of $\langle \Delta_{0i} \rangle$ and $var(\Delta_{0i})$ on i if we are able to estimate the rate of convergence for arbitrary functions defined on the induced Markov chain.

Suppose that $\{\mathbf{y}_0, \mathbf{y}_1, \dots\}$ is a Markov chain with the target density $p^*(\mathbf{y})$, $f(\mathbf{y})$ is some p^* -integrable function, and

$$p^*(f) = \int f(\mathbf{y}) p^*(\mathbf{y}) d\mathbf{y} \quad (42)$$

is the expectation of function f under the stationary density. The rate of geometric convergence of function f on the chain $\{\mathbf{y}\}$ may be defined as the minimum number $\rho(f)$ such that for all $r > \rho(f)$

$$\lim_{i \rightarrow \infty} \frac{1}{r^i} \int \left(\int f(\mathbf{y}_i) p(\mathbf{y}_i | \mathbf{y}_0) d\mathbf{y}_i - p^*(f) \right)^2 p(\mathbf{y}_0) d\mathbf{y}_0 = 0. \quad (43)$$

Roberts and Sahu (1997) investigate properties of geometric convergence for functions of Markov chains when the target density is a Gaussian. They show that under the deterministic updating strategy the convergence rate $\rho(f)$ of *any* function $f(\mathbf{y})$ is bounded above by the spectral radius ρ (maximum modulus eigenvalue) of a matrix \mathbf{B} formed from elements of the inverse covariance \mathbf{W} . For the case of the Gaussian Boltzmann machine described in section 3.1 the chain is given by $\{\mathbf{y}\} = \{[\mathbf{z}_0^T \mathbf{x}_0^T]^T, [\mathbf{z}_1^T \mathbf{x}_1^T]^T, \dots\}$ and \mathbf{B} is given by¹

$$\mathbf{B} = \begin{bmatrix} \mathbf{0} & -\mathbf{W}_{zz}^{-1}\mathbf{W}_{zx} \\ \mathbf{0} & \mathbf{W}_{xx}^{-1}\mathbf{W}_{xz}\mathbf{W}_{zz}^{-1}\mathbf{W}_{zx} \end{bmatrix}. \quad (44)$$

¹Note that since the leftmost blocks of \mathbf{B} are zeros $\rho(\mathbf{B}) = \rho(\mathbf{W}_{xx}^{-1}\mathbf{W}_{xz}\mathbf{W}_{zz}^{-1}\mathbf{W}_{zx})$.

From expression (41) we see that $\langle \Delta_{0i} \rangle$ is a function of $\{y\}$. Therefore, $\rho(B)$ gives an upper bound on the rate of geometric convergence of $\langle \Delta_{0i} \rangle$ to its expectation $\langle \Delta^{BM} \rangle$ under the stationary density. Analogously, $\rho(B)$ is an upper bound on the rate of geometric convergence for the variances $var(\Delta_{0i})$.

If we apply this bound to the 1-hidden 1-visible BM analyzed in section 2 we obtain a loose bound on the true rate of convergence. However, this is not very surprising as the spectral radius bound must apply for any function f . A specific analysis for $\langle \Delta_{0i} \rangle$ is given in section 3.3.

3.3 Analysis of $\langle \Delta_{0i} \rangle$

Consider the Markov chain for the evolution of x_i . This has Gaussian dynamics, so that $x_i = Fx_{i-1} + n_i$ for some state transition matrix $F = TS$ and some zero-mean Gaussian noise vector n_i with covariance $Q \stackrel{\text{def}}{=} T\text{cov}(u_i)T^T + \text{cov}(v_i)$. As $z_i = Sx_i + u_i$, we obtain

$$\langle \Delta_{0i} \rangle = S \langle x_0 x_0^T \rangle - S \langle x_i x_i^T \rangle. \quad (45)$$

Of course we have the decomposition

$$\langle x_i x_i^T \rangle = \langle x_i \rangle \langle x_i \rangle^T + cov(x_i). \quad (46)$$

Assuming that $x_0 \sim N(0, \Sigma_t)$ (the target density), then $\langle x_i \rangle = 0$. Let P_i denote $cov(x_i)$; clearly $P_i = FP_{i-1}F^T + Q$. Applying this recursively we can build up the expression $P_i = F^i P_0 (F^T)^i + \sum_{k=0}^{i-1} F^k Q (F^T)^k$ but this does not give a clear view of the convergence behaviour. However, we can carry out an analysis by viewing the Markov chain for x_i as a Kalman Filter with no observations, and solving the discrete-time matrix Riccati equation (see e.g. Grewal and Andrews (1993), section 4.9) with a zero state-to-observation mapping.

We represent P_i as $P_i = A_i B_i^{-1}$. It can then be shown that the equation for the update of P_i is equivalent to

$$\begin{bmatrix} A_i \\ B_i \end{bmatrix} = \begin{bmatrix} F & QF^{-T} \\ 0 & F^{-T} \end{bmatrix} \begin{bmatrix} A_{i-1} \\ B_{i-1} \end{bmatrix}. \quad (47)$$

This is initialized with $A_0 = \Sigma_t$ and $B_0 = I$. The $2|x| \times 2|x|$ matrix in equation (47) is known as the Hamiltonian matrix; let it have an eigendecomposition $V \Lambda V^{-1}$, where Λ is a diagonal matrix. Then we obtain

$$\begin{bmatrix} A_i \\ B_i \end{bmatrix} = V \Lambda^i V^{-1} \begin{bmatrix} \Sigma_t \\ I \end{bmatrix}. \quad (48)$$

Clearly the convergence of both A_i and B_i can be analyzed in terms of the eigenspectrum $\text{diag}(\Lambda)$, but as $P_i = A_i B_i^{-1}$ an exact analysis of the convergence of P_i is more taxing.

We note that as x_i is a Gaussian random variable, the fourth-order moments needed to analyze $var(\Delta_{0i})$ can be expressed in terms of the second order moments, although the analysis will be quite messy.