An Analysis of Contrastive Divergence Learning in GBMs

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 GS iterations for a simple GBM. 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.

1 BMs and Gaussian BMs

A BM (Ackley, Hinton and Sejnowski, 1985) is a lvm used for modelling data. Let x and z be stochastic visible and hidden variables of the BM, and let $\mathsf{y} = (\mathsf{x}^T, \mathsf{z}^T)^T$. There is a weight matrix W so that the energy of configuration y is $E(\mathsf{y}) = \frac{1}{2}\mathsf{y}^T W \mathsf{y}$. The usual BM has binary variables . The probability distribution $p(\mathsf{y}) \propto \exp -E(\mathsf{y})$.

In the case that x and z are real-valued we can still maintain the BM formalism, although W must be spd in order for the distribution to be proper (Williams (1993)).

1.1 BM Learning and CD Learning

The BM:

$$p(\mathsf{x}) = \frac{1}{Z} \int \exp -E(\mathsf{x}, \mathsf{z})\ d\mathsf{z}, \tag{1}$$

$$Z = \int \int \exp -E(\mathsf{x}, \mathsf{z})\ d\mathsf{z}\ d\mathsf{x}. \tag{2}$$

Let X denote a sample drawn iid from a target distribution over the visible variables . The log likelihood of X under the Boltzmann machine model is $\mathcal{L}(W) = \sum_{\mathsf{x} \in \mathsf{X}} \log p(\mathsf{x})$. Let us consider a weight w_{ij} which connects a visible unit x^i with a hidden unit z^j. ==>

$$\frac{\partial \mathcal{L}}{\partial w_{ij}} = \sum_{\mathsf{x} \in \mathsf{X}} [\langle x^i z^j \rangle_+ - \langle x^i z^j \rangle_-] \tag{3}$$

where

$$\langle x^i z^j \rangle_+ \stackrel{\text{def}}{=} \int x^i z^j p(\mathsf{z}|\mathsf{x}) d\mathsf{z}, \tag{4}$$

$$\langle x^i z^j \rangle_- \stackrel{\text{def}}{=} \int x^i z^j p(\mathsf{x}, \mathsf{z}) d\mathsf{x} d\mathsf{z}. \tag{5}$$

Thus the learning rule consists of the positive phase, where z is sampled given the presented x pattern, and the negative phase, where a sample is drawn from the joint distribution of x and z. In fact the exact averages shown in (4) and (5) are usually intractable and replaced by a sample drawn from the correct distribution. This is easily achieved by GS for the positive phase , but for the negative phase a MCMC method which alternates sampling from the hidden and visible units is required . This chain is illustrated in Figure 1. Starting at the clamped data vector x_0, we sample first from $p(\mathsf{z}|\mathsf{x}_0)$ to obtain z_0, then from $p(\mathsf{x}|\mathsf{z}_0)$ to obtain x_1 and so on. Under general conditions the MCMC method is guaranteed to draw from the correct distribution $p(\mathsf{x}, \mathsf{z})$ as the number of iterations tends to infinity.

However, MCMC approximation of the negative phase may be unacceptably slow in practice and give rise to samples with high variance. The idea of CD learning (Hinton, 1999, 2000) is to replace the negative phase of BM learning with $\langle x^i z^j \rangle_{p(\mathsf{x}_1, \mathsf{z}_1)}$, where $p(\mathsf{x}_1, \mathsf{z}_1)$ denotes the distribution of the GS variables as illustrated in Figure 1. We denote this as the CD(1) learning rule. In this notation the original negative phase is denoted $\langle x^i z^j \rangle_{p(\mathsf{x}_\infty, \mathsf{z}_\infty)}$. In general we can consider a CD(i) learning rule, that replaces the negative phase of the BM with $\langle x^i z^j \rangle_{p(\mathsf{x}_i, \mathsf{z}_i)}$.

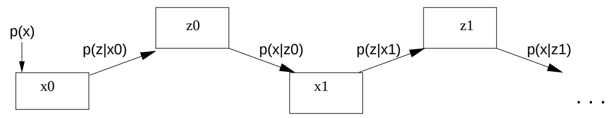


Figure 1: Markov chain used in GS for the x and z variables, starting with the data vector x_0.

The advantage of the CD(1) method is that it reduces the computational burden of the update. Also, it can be shown (Hinton, 2002, personal communication) that if the model can perfectly represent the data distribution then the stationary points of the CD(1) objective function are also stationary points of the CD(∞) or Boltzmann machine learning rule. We would expect that the gradient of the CD(1) objective function would be a biased estimate of the CD(∞) gradient, but that it would have smaller variance. These issues are explored below for a particular GBM architecture.

2 Case Study: CD(i) Learning in a Simple GBM

We consider a simple 1-hidden variable, 1-visible variable GBM, with x and z denoting the visible and hidden variables respectively. Let

$$W = \begin{bmatrix} \alpha & \omega \\ \omega & \alpha \end{bmatrix}. \tag{6}$$

Inverting this matrix we find that the covariance matrix:

$$\mathsf{C} = \frac{1}{\alpha^2 - \omega^2} \begin{bmatrix} \alpha & -\omega \\ -\omega & \alpha \end{bmatrix} \stackrel{\text{def}}{=} \begin{bmatrix} a & w \\ w & a \end{bmatrix}, \tag{7}$$

where $a \stackrel{\text{def}}{=} \alpha/(\alpha^2 - \omega^2)$, $w \stackrel{\text{def}}{=} -\omega/(\alpha^2 - \omega^2)$, $|w| < a$. We consider α to be fixed and are interested in the distribution of x as ω is varied. In fact $x \sim N(0, a)$. Thus we seek to adapt ω so that the resulting variance a of the visible variable matches the variance of the data. Let this *target* variance be denoted $a_t \stackrel{\text{def}}{=} var(\mathrm{p})$. We assume that the data is centered so that $E[x] = 0$.

In this section we analyse in detail the properties of the CD(i) learning rule for the 1-hidden 1-visible GBM. 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 GS for a GBM

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

$$\mathsf{C} = \begin{bmatrix} \mathsf{C}_{11} & \mathsf{C}_{12} \\ \mathsf{C}_{21} & \mathsf{C}_{22} \end{bmatrix}. \tag{8}$$

The conditional

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

where $(\boldsymbol{\mu}^T{}_1, \boldsymbol{\mu}^T{}_2)^T$ is the mean vector of the Gaussian (pp. 226-228 of Mises 1964).

From (7) and (9) it is easy to show that for the 1-visible 1-hidden variable model

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

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 $v_1, v_2, \ldots, u_0, u_1, \ldots$ correspond to mutually independent $N(0,1)$ rvs, 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

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

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

$$\Delta_{0i} = x_0 z_0 - x_i z_i. \tag{12}$$

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 $\langle \ldots \rangle_{x_0}$ denotes expectation over the data distribution $p(x_0)$. We also calculate the conditional variance $var(\Delta_{0i}|x_0)$ and the unconditional variance $var(\Delta_{0i})$.

Of course for a GBM it is not necessary to use the BM 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

$$x_i z_i = \sigma x_i^2 + \tau u_i x_i \;\Rightarrow\; \langle x_i z_i|x_0\rangle = \sigma\langle x_i^2|x_0\rangle. \tag{13}$$

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

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

Therefore

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

and

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

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 $\langle\Delta^{BM}\rangle$:

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

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

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

Thus, the gradient of the log-likelihood in i-step CD 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 GS 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 GBM 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

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

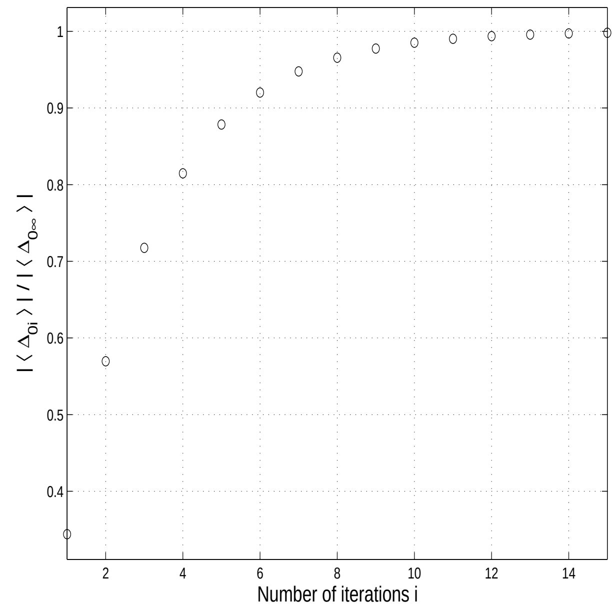


Figure 2: Plot of $|\langle \Delta_{0i} \rangle| / |\langle \Delta_{0\infty} \rangle|$ against iteration number i for $\sigma = 0.9$.

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

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

If only one chain is used (call this case D, D for dependent) then (19) must be corrected by a term $-2cov(x_0 z_0, x_i z_i | x_0)$. Analysis shows that $cov(x_0 z_0, x_i z_i | x_0) = 2 x_0^2 a (1 - \sigma^2) \sigma^{4i}$. 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

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

After some simplifications shown in Appendix A, (19) ==

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

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

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

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

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

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

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

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

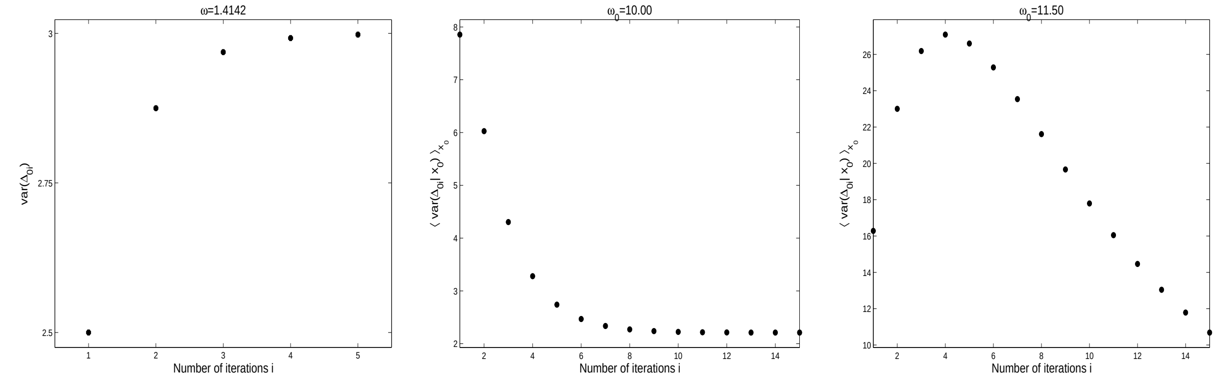
Here we investigate the behaviour of the variance of the CD(i) update term for the given model as a function of the the number of GS iterations.

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 $var(\Delta_{0i})$ 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 $var(\Delta_{0i})$ 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, $\langle var(\Delta_{0i}|x_0)\rangle_{x_0}$ 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 GS runs.



a b c

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

2.6 Quantification of the CD 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 GBM defined above.

Let $Q_0(\mathsf{x}) \stackrel{\text{def}}{=} p(\mathsf{x}_0)$ and $Q_i(\mathsf{x}) \stackrel{\text{def}}{=} p(\mathsf{x}_i)$ 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 $Q_\infty(\mathsf{x}) \stackrel{\text{def}}{=} p(\mathsf{x})$ under the model== minimization of the KL divergence $KL(Q_0(\mathsf{x})\|Q_\infty(\mathsf{x}))$ between the data and the model as

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

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:

$$\begin{aligned} \mathcal{L}^{(i)} &= KL(Q_i\|Q_\infty) - KL(Q_0\|Q_\infty) - H(Q_0) && (26)\\ &= \int Q_0(\mathsf{x}) \log Q_\infty(\mathsf{x}) d\mathsf{x} - H(Q_i) \\ &\quad - \int Q_i(\mathsf{x}) \log Q_\infty(\mathsf{x}) d\mathsf{x}. && (27) \end{aligned}$$

Notice that $\lim_{i\to\infty} \mathcal{L}^{(i)} = \mathcal{L}$. ==>

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

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

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

For the GBM 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

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

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

$$I_n \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} \exp\left\{-\frac{x^2}{2\sigma_i^2}\right\} x^n \mathrm{d}x \tag{31}$$

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

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

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

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

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

As we see from Figure 4, there exist parameter settings such that $|\epsilon_1|$ 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 $\langle\boldsymbol{\Delta_{0i}}\rangle$ is almost certain to have a positive cosine with $\langle\boldsymbol{\Delta_{0\infty}}\rangle$. Figure 4 also shows that $\lim_{i\to\infty} |\epsilon_i/\langle\Delta_{0i}\rangle| = 0$ as we would expect.

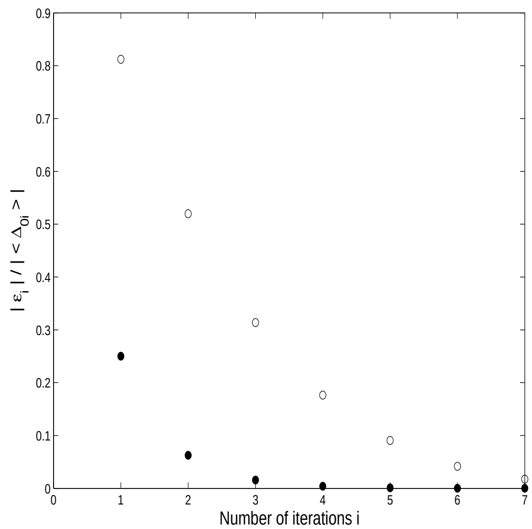


Figure 4: Plot of $|\epsilon_i/\langle\Delta_{0i}\rangle|$ as a function of i for $a_t = 25$, $\alpha = 12$, $\omega = 10$ (empty circles) and $a_t = 1$, $\alpha = 2$ and $\omega = \sqrt{2}$ (filled circles). Notice that in the latter case $a = a_t$.

3 Extension To Multivariate GBMs

In this section we describe general properties of the CD(i) learning for a multivariate GBM. 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 GBM

Let $\mathbf{\Sigma}$ and W be the covariance and the inverse covariance (weight) matrix of a GBM with $|\mathsf{x}|$ visible and $|\mathsf{z}|$ hidden variables, such that

$$\mathsf{W} = \begin{bmatrix} \mathsf{W}_{zz} & \mathsf{W}_{zx} \\ \mathsf{W}_{xz} & \mathsf{W}_{xx} \end{bmatrix}, \quad \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{zz} & \mathbf{\Sigma}_{zx} \\ \mathbf{\Sigma}_{xz} & \mathbf{\Sigma}_{xx} \end{bmatrix}, \quad \mathsf{W} = \mathbf{\Sigma}^{-1}. \tag{35}$$

Since both W, $\mathbf{\Sigma} \in \mathbb{R}^{(|\mathsf{z}|+|\mathsf{x}|)\times(|\mathsf{z}|+|\mathsf{x}|)}$ are symmetric, $\mathsf{W}_{zx} = \mathsf{W}_{xz}^T$ and $\mathbf{\Sigma}_{zx} = \mathbf{\Sigma}_{xz}^T$. As in the simple case considered above, learning in a multivariate GBM assumes adapting the weights W_{zx} between hidden and visible variables so that the covariance matrix $\mathbf{\Sigma}_{xx}$ 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

$$p(\mathsf{z}|\mathsf{x}) \sim N(-\mathsf{W}_{zz}^{-1}\mathsf{W}_{zx}\mathsf{x},\ \mathsf{W}_{zz}^{-1}) \tag{36}$$
$$p(\mathsf{x}|\mathsf{z}) \sim N(-\mathsf{W}_{xx}^{-1}\mathsf{W}_{xz}\mathsf{z},\ \mathsf{W}_{xx}^{-1}). \tag{37}$$

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

$$z_i = Sx_i + u_i \in \mathbb{R}^{|z|},\ i \geq 0 \tag{38}$$
$$x_j = Tz_{j-1} + v_j \in \mathbb{R}^{|x|},\ j \geq 1. \tag{39}$$

Here $S = -W_{zz}^{-1}W_{zx} \in \mathbb{R}^{|z|\times|x|}$, $T = -W_{xx}^{-1}W_{xz} \in \mathbb{R}^{|x|\times|z|}$, and $v_1, v_2, \ldots, u_0, u_1, \ldots$ are mutually independent rvs such that

$$u_i \sim N(0, W_{zz}^{-1}), \quad v_j \sim N(0, W_{xx}^{-1}). \tag{40}$$

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

$$\mathbf{\Delta}_{0i} = z_0 x_0^T - z_i x_i^T \in \mathbb{R}^{|z|\times|x|}, \tag{41}$$

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

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

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

Suppose that $\{y_0, y_1, \ldots\}$ is a Markov chain with the target density $p^\star(y)$, $f(y)$ is some $p^\star$-integrable function, and

$$p^\star(f) = \int f(y)p^\star(y)dy \tag{42}$$

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

$$\lim_{i\to\infty} \frac{1}{r^i} \int \left(\int f(y_i)p(y_i|y_0)dy_i - p^\star(f) \right)^2 p(y_0)dy_0 = 0. \tag{43}$$

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 $f(y)$ is bounded above by the spectral radius ρ (maximum modulus eigenvalue) of a matrix B formed from elements of the inverse covariance W. For the case of the GBM described in section 3.1 the chain is given by $\{y\} = \{[z^T{}_0 x^T{}_0]^T, [z^T{}_1 x^T{}_1]^T, \ldots\}$ and

$$B = \begin{bmatrix} \mathbf{0} & -W_{zz}^{-1}W_{zx} \\ \mathbf{0} & W_{xx}^{-1}W_{xz}W_{zz}^{-1}W_{zx} \end{bmatrix}. \tag{44}$$

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

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

If we apply this bound to the 1-hidden1-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 $\langle \boldsymbol{\Delta}_{0i} \rangle$ is given in section 3.3.

3.3 Analysis of $\langle \boldsymbol{\Delta}_{0i} \rangle$

Consider the Markov chain for the evolution of x_i. This has Gaussian dynamics, so that $\mathsf{x}_i = \mathsf{F}\mathsf{x}_{i-1} + \mathsf{n}_i$ for some state transition matrix $\mathsf{F} = \mathsf{TS}$ and some zero-mean Gaussian noise vector n_i with covariance $\mathsf{Q} \stackrel{\text{def}}{=} \mathsf{T}cov(\mathsf{u}_i)\mathsf{T}^T + cov(\mathsf{v}_i)$. As $\mathsf{z}_i = \mathsf{S}\mathsf{x}_i + \mathsf{u}_i$, we obtain

$$\langle \boldsymbol{\Delta}_{0i} \rangle = \mathsf{S}\langle \mathsf{x}_0 \mathsf{x}_0^T \rangle - \mathsf{S}\langle \mathsf{x}_i \mathsf{x}_i^T \rangle. \tag{45}$$

Of course we have the decomposition

$$\langle \mathsf{x}_i \mathsf{x}_i^T \rangle = \langle \mathsf{x}_i \rangle \langle \mathsf{x}_i \rangle^T + cov(\mathsf{x}_i). \tag{46}$$

Assuming that $\mathsf{x}_0 \sim N(0, \Sigma_t)$ (the target density), then $\langle \mathsf{x}_i \rangle = 0$. Let P_i denote $cov(\mathsf{x}_i)$; clearly $\mathsf{P}_i = \mathsf{F}\mathsf{P}_{i-1}\mathsf{F}^T + \mathsf{Q}$. Applying this recursively we can build up the expression $\mathsf{P}_i = \mathsf{F}^i \mathsf{P}_0 (\mathsf{F}^T)^i + \sum_{k=0}^{i-1} \mathsf{F}^k \mathsf{Q} (\mathsf{F}^T)^k$ 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 $\mathsf{P}_i = \mathsf{A}_i \mathsf{B}_i^{-1}$. It can then be shown that the equation for the update of P_i ==

$$\begin{bmatrix} \mathsf{A}_i \\ \mathsf{B}_i \end{bmatrix} = \begin{bmatrix} \mathsf{F} & \mathsf{Q}\mathsf{F}^{-T} \\ 0 & \mathsf{F}^{-T} \end{bmatrix} \begin{bmatrix} \mathsf{A}_{i-1} \\ \mathsf{B}_{i-1} \end{bmatrix}. \tag{47}$$

This is initialized with $\mathsf{A}_0 = \Sigma_t$ and $\mathsf{B}_0 = I$. The $2|\mathsf{x}| \times 2|\mathsf{x}|$ matrix in equation (47) is known as the Hamiltonian matrix; let it have an eigendecomposition $\mathsf{V}\Lambda\mathsf{V}^{-1}$, where Λ is a diagonal matrix. ==>

$$\begin{bmatrix} \mathsf{A}_i \\ \mathsf{B}_i \end{bmatrix} = \mathsf{V}\Lambda^i\mathsf{V}^{-1} \begin{bmatrix} \Sigma_t \\ \mathsf{I} \end{bmatrix}. \tag{48}$$

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

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

4 Discussion

we have generalized Hinton's one-step GS cd learning rule to the general i-steps case, and analysed its performance as compared to the the BM learning rule on a simple GBM. The CD(i) rule leads to a systematic bias in the calculation of the gradient of the log likelihood, although the stationary points of the CD(i) rule are stationary points of the BM rule. One key reason for the introduction of the CD(i) learning rule was that it was expected to reduce the variance of the parameter update Δ_{0i} (although introducing bias). We have confirmed this effect does indeed occur (for the single GS chain procedure, case D) and have quantified the effect. For case I and certain parameter settings (e.g. small $|\omega|$) $var(\Delta_{0i})$ does not increase monotonically with i, but in these cases it is almost constant. We have also analyzed the error in the CD(i) update rule due to ignoring the ϵ_i term, and have found that there are parameter settings where the relative error is large.

We have also been able to extend the analysis to the multivariate GBM and have shown geometric convergence of $\langle \boldsymbol{\Delta_{0i}} \rangle$ and $var(\boldsymbol{\Delta_{0i}})$ in this case.

A Analysis of CD Learning: Auxiliary Derivations

The appendix offers auxiliary derivations supporting results of sections 2.2 and 2.4 for the second and fourth moments $\langle x_i^2|x_0\rangle$, $\langle x_i^4|x_0\rangle$ of the i^{th} reconstruction of the visible variable conditioned on the initial data point x_0.

From (11) we find that

$$x_i = x_0\sigma^{2i} + \tau \sum_{k=1}^{i} \sigma^{2i-2k}(u_{k-1}\sigma + v_k). \tag{49}$$

Let $C_k \stackrel{\text{def}}{=} \sigma u_{k-1} + v_k$. By squaring (49) we obtain

$$x_i^2 = \sigma^{4i}\left[x_0^2 + 2x_0\tau\sum_{k=1}^{i} C_k\sigma^{-2k} + \tau^2\left(\sum_{k=1}^{i}\sigma^{-2k}C_k\right)^2\right]. \tag{50}$$

Notice that since $u_k,\ v_k \sim N(0,1)$, $\langle C_k\rangle = 0$ and $\langle C_k^2\rangle = 1+\sigma^2$. Thus we obtain

$$\begin{aligned}\langle x_i^2|x_0\rangle &= \sigma^{4i}\left[x_0^2 + (\sigma^2+1)\tau^2\sum_{k=1}^{i}\sigma^{-4k}\right]\\ &= \sigma^{4i}(x_0^2 - a) + a.\end{aligned} \tag{51}$$

Here we used the definition of $\tau^2 \stackrel{\text{def}}{=} a(1-\sigma^2)$ and the known form for the sum of geometric series.

Note that $x_i|x_0$ is a Gaussian rv with mean $\langle x_i|x_0\rangle = x_0\sigma^{2i}$ and variance $\langle {x_i}^2|x_0\rangle - \langle x_i |x_0\rangle^2$. It is well known that for a Gaussian RV $\zeta \sim N(m,v)$ $\langle\zeta^4\rangle = 3v^2 + 6vm^2 + m^4$. Using this and (51) above we obtain $\langle x_i^4|x_0\rangle = 3(\langle x_i^2|x_0\rangle)^2 - 2x_0^4\sigma^{8i}$.

References

Ackley, Hinton and Sejnowski (1985). A learning algorithm for boltzmann machines. *Cognitive Science*, 9:147 – 169.

Grewal, M. and Andrews, A. (1993). *Kalman Filtering: Theory and Practice.* Prentice Hall.

Hinton, G. (2000). Training products of experts by minimizing contrastive divergence. GCNU TR 2000-004, Gatsby Computational Neuroscience Unit, UCL.

Hinton, G. E. (1999). Products of experts. In *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN 99)*, pages 1–6.

Mises, R. V. (1964). *Mathematical Theory of Probability and Statistics.* New York: Academic Press.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C.* Cambridge University Press, Second edition.

Roberts, G. and Sahu, S. (1997). Updating Schemes, Correlation Structure, Blocking and Parametrization for the Gibbs Sampler. *Journal of Royal Statistical Society B*, 59:291 – 317.

Williams, C. K. I. (1993). *Continuous-valued Boltzmann Machines.* Unpublished manuscript.