Justifying and Generalizing Contrastive Divergence

Yoshua Bengio and Olivier Delalleau

Technical Report 1311, Dept. IRO, Université de Montréal

November 25, 2007

Abstract

We study an expansion of the log-likelihood in undirected graphical models such as the Restricted Boltzmann Machine (RBM), where each term in the expansion is associated with a sample in a Gibbs chain alternating between two random variables (the visible vector and the hidden vector, in RBMs). We are particularly interested in estimators of the gradient of the log-likelihood obtained through this expansion. We show that its terms converge to zero, justifying the use of a truncation, i.e. running only a short Gibbs chain, which is the main idea behind the Contrastive Divergence approximation of the log-likelihood gradient. By truncating even more, we obtain a stochastic reconstruction error, related through a mean-field approximation to the reconstruction error often used to train autoassociators and stacked auto-associators. The derivation is not specific to the particular parametric forms used in RBMs, and only requires convergence of the Gibbs chain.

1 Introduction

Motivated by the theoretical limitations of a large class of non-parametric learning algorithms (Bengio & Le Cun, 2007), recent research has focussed on learning algorithms for so-called **deep architectures** (Hinton, Osindero, & Teh, 2006; Hinton & Salakhutdinov, 2006; Bengio, Lamblin, Popovici, & Larochelle, 2007; Salakhutdinov & Hinton, 2007; Ranzato, Poultney, Chopra, & LeCun, 2007; Larochelle, Erhan, Courville, Bergstra, & Bengio, 2007). These represent the learned function through many levels of composition of elements taken in a small or parametric set. The most common element type found in the above papers is the soft or hard linear threshold unit, or **artificial neuron**

$$\text{output(input)} = s(w'\text{input} + b) \tag{1}$$

with parameters w (vector) and b (scalar), and where $s(a)$ could be $1_{a>0}$, $\tanh(a)$, or $\text{sigm}(a) = \frac{1}{1+e^{-a}}$, for example.

Here, we are particularly interested in the Restricted Boltzmann Machine (Smolensky, 1986; Freund & Haussler, 1994; Hinton, 2002; Welling, Rosen-Zvi, & Hinton, 2005; Carreira-Perpiñan & Hinton, 2005), a family of bipartite graphical models with hidden variables (the hidden layer) which are used as components in building Deep Belief Networks (Hinton et al., 2006; Bengio et al., 2007; Salakhutdinov & Hinton, 2007; Larochelle et al., 2007). Deep Belief Networks have yielded impressive performance on several benchmarks, clearly beating the state-of-the-art and other non-parametric learning algorithms in several cases. Currently the most successful learning algorithms for training a Restricted Boltzmann Machine (RBM) is the Contrastive Divergence (CD) algorithm. An RBM represents the joint distribution between a **visible** vector X which is the random variable observed in the data, and a **hidden** random variable H. There is not tractable representation of $P(X, H)$ but conditional distributions $P(H|X)$ and $P(X|H)$ can easily be computed and sampled from. CD-k is based on a Gibbs Monte-Carlo Markov Chain (MCMC) starting at an example $X = x_1$ from the empirical distribution and converging to the RBM's generative distribution $P(X)$. CD-k relies on a biased estimator obtained after a small number k of Gibbs steps (often only 1 step). Each Gibbs step is composed of two alternating sub-steps: sampling $h_t \sim P(H|X = x_t)$ and sampling $x_{t+1} \sim P(X|H = h_t)$, starting at $t = 1$.

The surprising empirical result is that even $k = 1$ (CD-1) often gives good results. An extensive numerical comparison of training with CD-k versus exact log-likelihood gradient has been presented in (Carreira-Perpiñan & Hinton, 2005). In these experiments, taking k larger than 1 gives more precise results, although very good approximations of the solution can be obtained even with $k = 1$.

CD-1 has originally been justified (Hinton, 2002) as an approximation of the gradient of $KL(P(X_2 = \cdot|x_1)||P(X = \cdot)) - KL(\hat{P}(X = \cdot)||P(X = \cdot))$, where KL is Kullback-Leibler divergence and $P(X_2 = \cdot|x_1)$ denotes the distribution of the chain after one step. The term left out in the approximation of the gradient of the KL difference is (Hinton, 2002)

$$\sum_x \frac{\partial KL(P(X_2 = \cdot|x_1)||P(X = \cdot)}{\partial P(X_2 = x|x_1)} \frac{\partial P(X_2 = x|x_1)}{\partial \theta} \tag{2}$$

which was empirically found to be small. On the one hand it is not clear how aligned are the log-likelihood gradient and the gradient with respect to the above KL difference. On the other hand it would be nice to prove that left-out terms are small in some sense. One of the motivations for this paper is to obtain the Contrastive Divergence algorithm from a different route, by which we can prove that the term left-out with respect to the *log-likelihood gradient* is small and converging to zero, as we take k larger.

We show that the log-likelihood and its gradient can be written down as a series where each term is associated with a step of the Gibbs chain. We show that when truncating the gradient series to k steps, the remainder converges to zero at a rate that depends on the mixing rate of the chain. The inspiration for this derivation comes from Hinton et al. (2006): first the idea that the Gibbs chain can be associated with an infinite directed graphical model (which here we associate to an expansion of the log-likelihood and of its gradient), and second that the convergence of the chain justifies Contrastive Divergence (since the k-th sample from the Gibbs chain becomes equivalent to a model sample).

Interestingly, the derivation is independent of the particular parametric formulation of $P(H|X)$ and $P(X|H)$, although the standard CD update is recovered in the case of the standard parametric formulation of RBMs. Finally, we show that when truncating the series to a single sub-step we obtain the gradient of a stochastic reconstruction error. A mean-field approximation of that error is the reconstruction error often used to train autoassociators (Rumelhart, Hinton, & Williams, 1986; Bourlard & Kamp, 1988; Hinton & Zemel, 1994; Schwenk & Milgram, 1995; Japkowicz, Hanson, & Gluck, 2000). Auto-associators can be stacked using the same principle used to stack RBMs into a Deep Belief Network in order to train deeep neural networks (Bengio et al., 2007; Ranzato et al., 2007; Larochelle et al., 2007). Reconstruction error has also been used to monitor progress in training RBMs by CD (Taylor, Hinton, & Roweis, 2006; Bengio et al., 2007), because it can be computed tractably and analytically, without sampling noise.

In the following we drop the $X = x$ notation and use shorthands such as $P(x|h)$ instead of $P(X = x|H = h)$. The t index is used to denote position in the Markov chain, whereas indices i or j denote an element of the hidden or visible vector respectively.

2 Restricted Boltzmann Machines and Contrastive Divergence

2.1 Boltzmann Machines

A Boltzmann Machine (Hinton, Sejnowski, & Ackley, 1984; Hinton & Sejnowski, 1986) is a probabilistic model of the joint distribution between **visible units** x, marginalizing over the values of **hidden units** h,

$$P(x) = \sum_h P(x, h) \tag{3}$$

and where the joint distribution between hidden and visible units is associated with a *quadratic* **energy function**

$$\mathcal{E}(x, h) = -b'x - c'h - h'Wx - x'Ux - h'Vh. \tag{4}$$

with,

$$P(x, h) = \frac{e^{-\mathcal{E}(x,h)}}{Z} \tag{5}$$

where $Z = \sum_{x,h} e^{-\mathcal{E}(x,h)}$ is a normalization constant (called the partition function) and (b, c, W, U, V) are parameters of the model. b_i is called the bias of visible unit x_i, c_i is the bias of visible unit h_i, and the matrices W, U, and V represent **interaction terms** between units. Note that non-zero U and V mean that there are interactions between

units belonging to the same layer (hidden layer or visible layer). Marginalizing over h at the level of the energy yields the so-called **free energy**:

$$\mathcal{F}(x) = -\log \sum_h e^{-\mathcal{E}(x,h)}. \tag{6}$$

We can rewrite the log-likelihood accordingly

$$\log P(x) = \log \sum_h e^{-\mathcal{E}(x,h)} - \log \sum_{\tilde{x},\tilde{h}} e^{-\mathcal{E}(\tilde{x},\tilde{h})} = -\mathcal{F}(x) - \log \sum_{\tilde{x}} e^{-\mathcal{F}(\tilde{x})}. \tag{7}$$

Differentiating the above, the gradient of the log-likelihood can be written as follows:

$$\begin{aligned} \frac{\partial \log P(x)}{\partial \theta} &= -\frac{\sum_h e^{-\mathcal{E}(x,h)} \frac{\partial \mathcal{E}(x,h)}{\partial \theta}}{\sum_h e^{-\mathcal{E}(x,h)}} + \frac{\sum_{\tilde{x},\tilde{h}} e^{-\mathcal{E}(\tilde{x},\tilde{h})} \frac{\partial \mathcal{E}(\tilde{x},\tilde{h})}{\partial \theta}}{\sum_{\tilde{x},\tilde{h}} e^{-\mathcal{E}(\tilde{x},\tilde{h})}} \\ &= -\sum_h P(h|x) \frac{\partial \mathcal{E}(x,h)}{\partial \theta} + \sum_{\tilde{x},\tilde{h}} P(\tilde{x},\tilde{h}) \frac{\partial \mathcal{E}(\tilde{x},\tilde{h})}{\partial \theta}. \end{aligned} \tag{8}$$

Computing $\frac{\partial \mathcal{E}(x,h)}{\partial \theta}$ is straightforward. Therefore, if sampling from the model was possible, one could obtain a stochastic gradient for use in training the model, as follows. Two samples are necessary: h given x for the first term, which is called the **positive phase**, and an $(\tilde{x},\tilde{h})$ pair from $P(\tilde{x},\tilde{h})$ in what is called the **negative phase**. Note how the resulting stochastic gradient estimator

$$-\frac{\partial \mathcal{E}(x,h)}{\partial \theta} + \frac{\partial \mathcal{E}(\tilde{x},\tilde{h})}{\partial \theta} \tag{9}$$

has one term for each of the positive phase and negative phase, with the same form but opposite signs. Let $u = (x,h)$ be a vector with all the unit values. In a general Boltzmann machine, one can compute and sample from $P(u_i|u_{-i})$, where u_{-i} is the vector with all the unit values except the i-th. Gibbs sampling with as many sub-steps as units in the model has been used to train Boltzmann machines in the past, with very long chains, yielding correspondingly long training times.

2.2 Restricted Boltzmann Machines

In a Restricted Boltzmann Machine (RBM), $U = 0$ and $V = 0$ in eq. 4, i.e. the only interaction terms are between a hidden unit and a visible unit, but not between units of the same layer. This form of model was first introduced under the name of **Harmonium** (Smolensky, 1986). Because of this restriction, $P(h|x)$ and $P(x|h)$ factorize and can be computed and sampled from easily. This enables the use of a 2-step Gibbs sampling alternating between $h \sim P(H|X = x)$ and $x \sim P(X|H = h)$. In addition, the positive phase gradient can be obtained exactly because the free energy factorizes:

$$\begin{aligned} e^{-\mathcal{F}(x)} = \sum_h e^{b'x + c'h + h'Wx} &= e^{b'x} \sum_{h_1} \sum_{h_2} \cdots \sum_{h_k} \prod_{i=1}^{k} e^{c_i h_i + (Wx)_i h_i} \\ &= e^{b'x} \sum_{h_1} e^{h_1(c_1 + W_1 x)} \cdots \sum_{h_k} e^{h_k(c_k + W_k x)} \\ &= e^{b'x} \prod_{i=1}^{k} \sum_{h_i} e^{h_i(c_i + W_i x)} \end{aligned}$$

where W_i is the i-th row of W. Using the same type of factorization, one obtains for example in the most common case where h_i is binary

$$-\sum_h P(h|x) \frac{\partial \mathcal{E}(x,h)}{\partial W_{ij}} = E[H_i|x] \cdot x_j, \tag{10}$$

where

$$E[H_i|x] = P(H_i = 1|X = x) = \text{sigm}(c_i + W_i x). \tag{11}$$

The log-likelihood gradient for W_{ij} thus has the form

$$\frac{\partial \log P(x)}{\partial W_{ij}} = P(H_i = 1|X = x) \cdot x_j - E_X[P(H_i = 1|X) \cdot X_j] \tag{12}$$

where E_X is an expectation over $P(X)$. Samples from $P(X)$ can be approximated by running an alternating Gibbs chain $x_1 \Rightarrow h_1 \Rightarrow x_2 \Rightarrow h_2 \Rightarrow \ldots$. Since the model P is trying to imitate the empirical distribution $\hat{P}$, it is a good idea to start the chain from $\hat{P}$, so that we start the chain from a distribution close to the asymptotic one.

In most uses of RBMs (Hinton, 2002; Carreira-Perpiñan & Hinton, 2005; Hinton et al., 2006; Bengio et al., 2007) both h_i and x_i are binary, but many extensions are possible and have been studied, including cases where hidden and/or visible units are continuous-valued (Freund & Haussler, 1994; Welling et al., 2005; Bengio et al., 2007).

2.3 Contrastive Divergence

The k-step Contrastive Divergence (CD-k) (Hinton, 1999, 2002) involves a second approximation besides the use of MCMC to sample from P. This additional approximation introduces some bias in the gradient: we run the MCMC chain for only k steps, starting from the observed example x. Using the same technique as in eq. 8 to express the log-likelihood gradient, but keeping the sums over h inside the free energy, we obtain

$$\begin{aligned}
\frac{\partial \log P(x)}{\partial \theta} &= \frac{\partial(-\mathcal{F}(x) - \log \sum_x e^{-\mathcal{F}(x)})}{\partial \theta} \\
&= -\frac{\partial \mathcal{F}(h)}{\partial \theta} + \frac{\sum_{\tilde{x}} e^{-\mathcal{F}(\tilde{x})} \frac{\partial \mathcal{F}(\tilde{x})}{\partial \theta}}{\sum_{\tilde{x}} e^{-\mathcal{F}(\tilde{x})}} \\
&= -\frac{\partial \mathcal{F}(h)}{\partial \theta} + \sum_{\tilde{x}} P(\tilde{x}) \frac{\partial \mathcal{F}(\tilde{x})}{\partial \theta}.
\end{aligned} \tag{13}$$

The CD-k update after seeing example x is taken proportional to

$$\Delta\theta = -\frac{\partial \mathcal{F}(x)}{\partial \theta} + \frac{\partial \mathcal{F}(\tilde{x})}{\partial \theta} \tag{14}$$

where $\tilde{x}$ is a sample from our Markov chain after k steps. We know that when $k \to \infty$, the samples from the Markov chain converge to samples from P, and the bias goes away. We also know that when the model distribution is very close to the empirical distribution, i.e., $P \approx \hat{P}$, then when we start the chain from x (a sample from $\hat{P}$) the MCMC samples have already converged to P, and we need less sampling steps to obtain an unbiased (albeit correlated) sample from P.

3 Log-Likelihood Expansion via Gibbs Chain

In the following we consider the case where both h and x can only take a finite number of values. We also assume that there is no pair (x, h) such that $P(x|h) = 0$ or $P(h|x) = 0$. This ensures the Markov chain associated with Gibbs sampling is **irreducible** (one can go from any state to any other state), and there exists a unique stationary distribution $P(x, h)$ the chain converges to.

Lemma 3.1. *Consider the irreducible Gibbs chain $x_1 \Rightarrow h_1 \Rightarrow x_2 \Rightarrow h_2 \ldots$ starting at data point x_1. The log-likelihood can be expanded as follows for any path of the chain:*

$$\log P(x_1) = \log P(x_t) + \sum_{s=1}^{t-1} \log \frac{P(x_s|h_s)}{P(h_s|x_s)} + \log \frac{P(h_s|x_{s+1})}{P(x_{s+1}|h_s)} \tag{15}$$

and consequently, since this is true for any path:

$$\log P(x_1) = E[\log P(x_t)|x_1] + \sum_{s=1}^{t-1} E\left[\log \frac{P(x_s|h_s)}{P(h_s|x_s)} + \log \frac{P(h_s|x_{s+1})}{P(x_{s+1}|h_s)} \middle| x_1 \right] \tag{16}$$

where the expectation is over Markov chain sample paths.

Proof. We will expand the expression of the log-likelihood by introducing samples of the Gibbs chain $x_1 \Rightarrow h_1 \Rightarrow x_2 \Rightarrow h_2 \Rightarrow \ldots$:

$$\log P(x_1) \quad = \quad \log \frac{P(x_1)}{P(h_1)} \frac{P(h_1)}{P(x_2)} \frac{P(x_2)}{P(h_2)} \ldots \frac{P(h_{t-1})}{P(x_t)} P(x_t)$$

and substitute

$$\frac{P(x_s)}{P(h_s)} = \frac{P(x_s|h_s)}{P(h_s|x_s)} \tag{17}$$

and

$$\frac{P(h_s)}{P(x_{s+1})} = \frac{P(h_s|x_{s+1})}{P(x_{s+1}|h_s)} \tag{18}$$

to obtain

$$\log P(x_1) = \log P(x_t) + \sum_{s=1}^{t-1} \left(\log \frac{P(x_s|h_s)}{P(h_s|x_s)} + \log \frac{P(h_s|x_{s+1})}{P(x_{s+1}|h_s)} \right). \tag{19}$$

Since this is true for any sequence $h_1, x_2, h_2, \ldots x_t$, it is also true of any weighted average of such sequences. Taking these weights to be $P(h_1, x_2, h_2, \ldots x_t|x_1)$ and summing over all possible paths of the chain, we obtain

$$\log P(x_1) = E[\log P(x_t)|x_1] + \sum_{s=1}^{t-1} E\left[\log \frac{P(x_s|h_s)}{P(h_s|x_s)} + \log \frac{P(h_s|x_{s+1})}{P(x_{s+1}|h_s)} \middle| x_1 \right]. \tag{20}$$

□

Note that $E[\log P(x_t)]$ is the negative entropy of the t-th visible sample of the chain, and it does not become smaller as $t \to \infty$. Therefore it does not seem reasonable to truncate this expansion. However, the gradient of the log-likelihood is more interesting. But first we need a simple lemma.

Lemma 3.2. *For any model $P(Y)$ with parameters θ,*

$$E\left[\frac{\partial \log P(Y)}{\partial \theta} \right] = 0 \tag{21}$$

when the expected value is taken according to $P(Y)$.

Proof. We start from the sum to 1 constraint on $P(Y)$, differentiate and obtain the Lemma. To obtain the last line below we use the fact that for any function $f(\theta)$, we have $\frac{\partial f(\theta)}{\partial \theta} = f(\theta) \frac{\partial \log f(\theta)}{\partial \theta}$.

$$\begin{aligned} E[1] &= \sum_y P(Y = y) = 1 \\ \frac{\partial \sum_y P(Y = y)}{\partial \theta} &= \frac{\partial 1}{\partial \theta} = 0 \\ \sum_y P(Y = y) \frac{\partial \log P(Y = y)}{\partial \theta} &= 0 \end{aligned}$$

□

The lemma is clearly also true for conditional distributions with corresponding conditional expectations.

Theorem 3.3. *Consider the converging Gibbs chain $x_1 \Rightarrow h_1 \Rightarrow x_2 \Rightarrow h_2 \ldots$ starting at data point x_1. The log-likelihood gradient can be expanded in a converging series as follows:*

$$\begin{aligned} \frac{\partial \log P(x_1)}{\partial \theta} &= \sum_{s=1}^{t-1} \left(E\left[\frac{\partial \log P(x_s|h_s)}{\partial \theta} \middle| x_1 \right] + E\left[\frac{\partial \log P(h_s|x_{s+1})}{\partial \theta} \middle| x_1 \right] \right) \\ &+ E\left[\frac{\partial \log P(x_t)}{\partial \theta} \middle| x_1 \right] \end{aligned} \tag{22}$$

and the terms in the sum as well as the final term converge to zero as s and t respectively are taken larger.

Proof. We take derivatives with respect to a parameter θ in the log-likelihood expansion in eq. 15 of Lemma 3.1:

$$\frac{\partial \log P(x_1)}{\partial \theta} = \frac{\partial \log P(x_t)}{\partial \theta} + \sum_{s=1}^{t-1} \frac{\partial \log P(x_s|h_s)}{\partial \theta} - \frac{\partial \log P(h_s|x_s)}{\partial \theta} + \frac{\partial \log P(h_s|x_{s+1})}{\partial \theta} - \frac{\partial \log P(x_{s+1}|h_s)}{\partial \theta}. \tag{23}$$

Then we take expectations with respect to the Markov chain conditional on x_1, getting

$$\begin{aligned} \frac{\partial \log P(x_1)}{\partial \theta} &= E\left[\frac{\partial \log P(x_t)}{\partial \theta} \middle| x_1 \right] \\ &+ E\left[\sum_{s=1}^{t-1} \frac{\partial \log P(x_s|h_s)}{\partial \theta} - \frac{\partial \log P(h_s|x_s)}{\partial \theta} + \frac{\partial \log P(h_s|x_{s+1})}{\partial \theta} - \frac{\partial \log P(x_{s+1}|h_s)}{\partial \theta} \middle| x_1 \right]. \end{aligned} \tag{24}$$

Notice that the terms with a minus correspond to conditionals occuring in the chain, i.e., Lemma 3.2 can be applied to eliminate them, e.g.,

$$E_{x_s}\left[E_{h_s}\left[\frac{\partial \log P(h_s|x_s)}{\partial \theta} \middle| x_s \right] \middle| x_1 \right] = E_{x_s}[0|x_1] = 0, \tag{25}$$

yielding

$$\frac{\partial \log P(x_1)}{\partial \theta} = E\left[\frac{\partial \log P(x_t)}{\partial \theta} \middle| x_1 \right] + \sum_{s=1}^{t-1} \left(E\left[\frac{\partial \log P(x_s|h_s)}{\partial \theta} \middle| x_1 \right] + E\left[\frac{\partial \log P(h_s|x_{s+1})}{\partial \theta} \middle| x_1 \right] \right). \tag{26}$$

In order to prove the convergence of each individual term towards zero, we will use the assumed convergence of the chain, which can be written

$$P(X_t = x|X_1 = x_1) = P(x) + \epsilon_t(x) \tag{27}$$

with $\sum_x \epsilon_t(x) = 1$ and $\lim_{t\to+\infty} \epsilon_t(x) = 0$ for all x. Since x is discrete, $\epsilon_t \stackrel{def}{=} \max_x |\epsilon_t(x)|$ also verifies $\lim_{t\to+\infty} \epsilon_t = 0$. Then we can rewrite the first expectation as follows:

$$\begin{aligned} E\left[\frac{\partial \log P(x_t)}{\partial \theta} \middle| x_1 \right] &= \sum_{x_t} P(x_t|x_1) \frac{\partial \log P(x_t)}{\partial \theta} \\ &= \sum_{x_t} \left(P(x_t) + \epsilon_t(x_t)\right) \frac{\partial \log P(x_t)}{\partial \theta} \\ &= \sum_{x_t} P(x_t) \frac{\partial \log P(x_t)}{\partial \theta} + \sum_{x_t} \epsilon_t(x_t) \frac{\partial \log P(x_t)}{\partial \theta}. \end{aligned}$$

Using Lemma 3.2, the first sum is equal to zero. Thus we can bound this expectation by

$$\begin{aligned} \left| E\left[\frac{\partial \log P(x_t)}{\partial \theta} \middle| x_1 \right] \right| &\leq \sum_{x_t} |\epsilon_t(x_t)| \left| \frac{\partial \log P(x_t)}{\partial \theta} \right| \\ &\leq \left(N_x \max_x \left| \frac{\partial \log P(x)}{\partial \theta} \right| \right) \epsilon_t \end{aligned}$$

where N_x is the number of discrete configurations for the random variable X. This proves the expectation converges to zero as $t \to +\infty$, since $\lim_{t\to+\infty} \epsilon_t = 0$.

To prove that $E\left[\frac{\partial \log P(x_t)}{\partial \theta} \middle| x_1 \right]$ also converges to zero we need to formalize the convergence of the chain for conditional probabilities. For any x, h we have that $\lim_{s\to+\infty} P(X_s = x, H_s = h|X_1 = x_1) = P(x, h)$ and $\lim_{s\to+\infty} P(H_s = h|X_1 = x_1) = P(h)$. It follows that

$$\begin{aligned} \lim_{s\to+\infty} P(X_s = x|H_s = h, X_1 = x_1) &= \lim_{s\to+\infty} \frac{P(X_s = x, H_s = h|X_1 = x_1)}{P(H_s = h|X_1 = x_1)} \\ &= \frac{P(x, h)}{P(h)} = P(x|h). \end{aligned}$$

Thus we can write

$$P(X_s = x|H_s = h, X_1 = x_1) = P(x|h) + \delta_s(x, h) \tag{28}$$

where $\lim_{s\to+\infty} \delta_s(x, h) = 0$. Since x and h are discrete $\delta_s \stackrel{def}{=} \max_{x,h} |\delta_s(x, h)|$ verifies $\lim_{s\to+\infty} \delta_s = 0$. For $s > 1$, the second expectation in eq. 26 becomes

$$\begin{aligned}
E\left[\left.\frac{\partial \log P(x_s|h_s)}{\partial\theta}\right| x_1\right] &= \sum_{x_s,h_s} P(x_s, h_s|x_1)\frac{\partial \log P(x_s|h_s)}{\partial\theta} \\
&= \sum_{h_s} P(h_s|x_1) \sum_{x_s} P(x_s|h_s, x_1)\frac{\partial \log P(x_s|h_s)}{\partial\theta} \\
&= \sum_{h_s} P(h_s|x_1) \sum_{x_s} (P(x_s|h_s) + \delta_s(x_s, h_s))\frac{\partial \log P(x_s|h_s)}{\partial\theta} \\
&= \sum_{h_s} P(h_s|x_1) \sum_{x_s} P(x_s|h_s)\frac{\partial \log P(x_s|h_s)}{\partial\theta} + \sum_{h_s,x_s} P(h_s|x_1)\delta_s(x_s, h_s)\frac{\partial \log P(x_s|h_s)}{\partial\theta}
\end{aligned}$$

Using Lemma 3.2, the first term is equal to zero so that the above expectation can be bounded by

$$\begin{aligned}
\left|E\left[\left.\frac{\partial \log P(x_s|h_s)}{\partial\theta}\right| x_1\right]\right| &\leq \sum_{h_s,x_s} P(h_s|x_1)|\delta_s(x_s, h_s)|\left|\frac{\partial \log P(x_s|h_s)}{\partial\theta}\right| \\
&\leq \left(N_h N_x \max_{x,h}\left|\frac{\partial \log P(x|h)}{\partial\theta}\right|\right)\delta_s
\end{aligned}$$

where N_h and N_x are respectively the number of possible configurations of the hidden and visible layers. As a result, this expectation converges to zero $\lim_{s\to+\infty} \delta_s = 0$. By switching the roles of x and h, a similar argument can be made to prove that the third expectation also verifies

$$\lim_{s\to+\infty} E\left[\left.\frac{\partial \log P(h_s|x_{s+1})}{\partial\theta}\right| x_1\right] = 0. \tag{29}$$

□

One may wonder to what extent the above results still hold in the situation where x and h are not discrete anymore, but instead may take values in infinite (possibly uncountable) sets. We assume $P(x|h)$ and $P(h|x)$ are such that there still exists a unique stationary distribution $P(x, h)$. Lemma 3.1 and its proof remain unchanged. On another hand, Lemma 3.2 is only true for distributions P such that

$$\int_y \frac{\partial P(y)}{\partial\theta} dy = \frac{\partial}{\partial\theta}\int_y P(y)dy. \tag{30}$$

This equation can be guaranteed to be verified under additional "niceness" assumptions on P, and we assume it is the case for distributions $P(x)$, $P(x|h)$ and $P(h|x)$. Consequently, the gradient expansion (eq. 22) in Theorem 3.3 can be obtained in the same way as before. The key point to justify further truncation of this expansion is the convergence towards zero of the residual term

$$E\left[\left.\frac{\partial \log P(x_t)}{\partial\theta}\right| x_1\right]. \tag{31}$$

This convergence is not necessarily guaranteed unless we have convergence of $P(x_t|x_1)$ to $P(x_t)$ in the sense that

$$\lim_{t\to\infty} E\left[\left.\frac{\partial \log P(x_t)}{\partial\theta}\right| x_1\right] = E\left[\frac{\partial \log P(x)}{\partial\theta}\right], \tag{32}$$

where the second expectation is over the stationary distribution P. If the distributions $P(x|h)$ and $P(h|x)$ are such that eq. 32 is verified, then this limit is also zero according to Lemma 3.2, and it makes sense to truncate eq. 22. Note however than eq. 32 does not necessarily hold in the most general case (Hernández-Lerma & Lasserre, 2003).

4 Connection with Contrastive Divergence

The above result justifies truncating the series after k steps, i.e., computing only the first k steps in the chain. Note how the sums in the above expansion can be readily replaced by sampling (which is easy for the first k steps in the Gibbs chain). This gives rise to a stochastic gradient, whose expected value is the exact expression associated with a truncation of the above log-likelihood gradient expansion. Finally, we show that truncating to the first k steps gives a parameter update that is exactly the CD-k update in the case of a binomial RBM, i.e., with binary-valued units.

Corollary 4.1. *When considering only the terms arising from the first k steps in the Gibbs chain $x_1 \Rightarrow h_1 \Rightarrow x_2 \Rightarrow h_2 \Rightarrow \ldots x_k \Rightarrow h_k$, the unbiased stochastic estimator of the gradient of the truncated log-likelihood expansion of Theorem 3.3 (with expectations replaced by samples in the chain) equals the CD-k update in the case of a binomial RBM.*

Proof. Let $W_{\cdot j}$ denote the j-th column of W and $W_{i\cdot}$ its i-th row. In the Restricted Boltzmann Machine (RBM) we have $P(h|x) = \prod_i P(h_i|x)$, with $P(h_i = 1|x) = \text{sigmoid}(b_i + W_{i\cdot}x)$ and similarly $P(x|h) = \prod_j P(x_j|h)$, with $P(x_j = 1|h) = \text{sigmoid}(c_j + h'W_{\cdot j})$. We want to see what the terms in the truncated sum of eq. 22 look like with this parametrization:

$$\frac{\partial \log P(x_t|h_t)}{\partial W_{ij}} = (x_{tj} - P(x_j = 1|h_t))h_{ti} \tag{33}$$

and

$$\frac{\partial \log P(h_t|x_{t+1})}{\partial W_{ij}} = (h_{ti} - P(h_i = 1|x_{t+1}))x_{t+1,j}. \tag{34}$$

Adding the two terms, and taking the expectation over $P(x_{t+1,j}|h_t)$, we notice that in expectation $-P(x_j = 1|h_t))h_{ti}$ cancels $h_{ti}x_{t+1,j}$. That leaves

$$\Delta W_{ij} = x_{tj}h_{ti} - P(h_i = 1|x_{t+1})x_{t+1,j} \tag{35}$$

When $t = k = 1$, this is exactly 1-step contrastive divergence for the RBM. Consider two such consecutive differences, for t and $t + 1$:

$$\begin{aligned} x_{tj}h_{ti} - P(h_i = 1|x_{t+1})x_{t+1,j} \quad &+ \\ x_{t+1,j}h_{t+1,i} - P(h_i = 1|x_{t+2})x_{t+2,j} \quad &. \end{aligned}$$

Again, we notice a cancellation, in expectation, of $-P(h_i = 1|x_{t+1})x_{t+1,j}$ with $x_{t+1,j}h_{t+1,i}$, because $h_{t+1,i}$ is sampled from $P(h_i|x_{t+1})$. Hence for k-step generalized contrastive divergence applied to an RBM, all the intermediate products cancel out, and we are left with only the first and last product:

$$\Delta W_{ij} = x_{1j}h_{1i} - P(h_i = 1|x_{k+1})x_{k+1,j}. \tag{36}$$

This is the standard k-step contrastive divergence update rule for binomial RBMs. Using the same procedure, the same can be shown for the biases. □

Inspection of the proof of the theorem suggests that the mixing rate of the Gibbs chain is an important factor in the quality of the approximation obtained by CD-k for different values of k, since it determines the error in throwing out the final term in the series. When the RBM weights are large it is plausible that the chain will mix more slowly because there is less randomness in each sampling step. Hence it might be advisable to use larger values of k as the weights become larger. However, it is not clear from this analysis how to precisely adjust k.

5 Connection with Autoassociator Reconstruction Error

Theorem 3.3 can be rewritten easily so that the final term is not $E\left[\frac{\partial \log P(x_t)}{\partial \theta}\middle| x_1\right]$ but rather $E\left[\frac{\partial \log P(h_t)}{\partial \theta}\middle| x_1\right]$, and now the expansion has an odd number of terms (in addition to the remainder). If we truncate the gradient series so as to keep only *the first term*, we obtain

$$\sum_{h_1} P(h_1|x_1)\frac{\partial \log P(x_1|h_1)}{\partial \theta} \tag{37}$$

which is an average over $P(h_1|x_1)$, which could be approximated by sampling. Note that this is not quite the negated gradient of the **stochastic reconstruction error**

$$\text{SRE} = -\sum_{h_1} P(h_1|x_1) \log P(x_1|h_1). \tag{38}$$

The gradient of the stochastic reconstruction error would be

$$\frac{\partial \text{SRE}}{\partial \theta} = -\sum_{h_1} P(h_1|x_1) \left(\frac{\partial \log P(x_1|h_1)}{\partial \theta} \; + \; \log P(x_1|h_1) \frac{\partial \log P(h_1|x_1)}{\partial \theta} \right). \tag{39}$$

Here we consider a notion of **mean-field approximation** by which an average $E_X[f(X)]$ over configurations of a random variable X is approximated by $f(E[X])$, i.e., using the mean configuration. Applying such an approximation to either eq. 37 or the above stochastic reconstruction gradient, i.e., instead of summing over all h_1 configurations we replace instances of h_1 by $\hat{h}_1 = E[H_1|x_1]$. Both eq. 37 and eq. 39 become (up to sign)

$$\frac{\partial \log P(x_1|\hat{h}_1)}{\partial \theta} \tag{40}$$

because, e.g., in the case where h_1 is binomial with $E[H_{1i}|x_1] = P(H_{1i} = 1|X_1 = x_1) = \text{sigm}(a_i)$, we have

$$\frac{\partial \log P(h_1|x_1)}{\partial \theta} = \sum_i (h_{1i} - E[H_{1i}|x_1]) \frac{\partial a_i}{\partial \theta} \tag{41}$$

and replacing h_1 by $\hat{h}_1$, we obtain

$$\sum_i (\hat{h}_{1i} - E[H_{1i}|x_1]) \frac{\partial a_i}{\partial \theta} = 0. \tag{42}$$

It is arguable whether the mean-field approximation per se gives us license to include in $\frac{\partial \log P(x_1|\hat{h}_1)}{\partial \theta}$ the effect of θ on $\hat{h}_1$, but if we do not then only the effect of θ on the reconstruction given $\hat{h}_1$ as fixed would be taken into account, and it seems preferable to take also into account the influence of θ on $\hat{h}_1$.

In that case, we find that eq. 40 can be interpreted as the gradient of the **reconstruction error** typically used in training autoassociators (Rumelhart et al., 1986; Bourlard & Kamp, 1988; Hinton & Zemel, 1994; Schwenk & Milgram, 1995; Japkowicz et al., 2000; Bengio et al., 2007; Ranzato et al., 2007; Larochelle et al., 2007),

$$\text{RE} = -\log P(x_1|\hat{h}_1) \tag{43}$$

which is a mean-field approximation of the stochastic reconstruction error.

Whereas CD-1 keeps the first two terms in the series, reconstruction error update thus relies only the first term, and on a mean-field approximation of that term (to convert from stochastic reconstruction error to reconstruction error). The reconstruction error gradient can be seen as a more biased approximation of the log-likelihood gradient than CD-1. Comparative experiments between reconstruction error training and CD-1 training confirm this view (Bengio et al., 2007; Larochelle et al., 2007): CD-1 updating generally has a slight advantage over reconstruction error gradient.

However, reconstruction error can be computed deterministically and has been used as an easy method to monitor the progress of training RBMs with CD, whereas the CD-k itself is generally not the gradient of anything and is stochastic.

6 Conclusion

This paper provides a theoretical analysis of the log-likelihood gradient in graphical models involving a hidden variable h in addition to the observed variable x, and where conditionals $P(h|x)$ and $P(x|h)$ are easy to compute and sample from. That includes the case of Contrastive Divergence for Restricted Boltzmann Machines (RBM). The analysis justifies the use of a short Gibbs chain to obtain a biased but converging estimator of the log-likelihood gradient. Since the left-over term in the gradient expansion depends on the closeness between the t-th sample in the

chain and a sample from the model, it confirms the importance of paying attention to the mixing rate of the Gibbs chain (either by some form of regularization or by selecting k adaptively).

The analysis also shows a connection between reconstruction error and log-likelihood and Contrastive Divergence (CD), which helps understand the better results generally obtained with CD and justify the use of reconstruction error as a monitoring device when training an RBM by CD. The generality of the analysis also opens the door to other learning algorithms in which $P(h|x)$ and $P(x|h)$ do not have the parametric forms of RBMs.

References

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy layer-wise training of deep networks. In Schölkopf, B., Platt, J., & Hoffman, T. (Eds.), *Advances in Neural Information Processing Systems 19*, pp. 153–160. MIT Press.

Bengio, Y., & Le Cun, Y. (2007). Scaling learning algorithms towards AI. In Bottou, L., Chapelle, O., DeCoste, D., & Weston, J. (Eds.), *Large Scale Kernel Machines*. MIT Press.

Bourlard, H., & Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, *59*, 291–294.

Carreira-Perpiñan, M., & Hinton, G. (2005). On contrastive divergence learning. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, Jan 6-8, 2005, Savannah Hotel, Barbados*.

Freund, Y., & Haussler, D. (1994). Unsupervised learning of distributions on binary vectors using two layer networks. Tech. rep. UCSC-CRL-94-25, University of California, Santa Cruz.

Hernández-Lerma, O., & Lasserre, J. B. (2003). *Markov Chains and Invariant Probabilities*. Birkhäuser Verlag.

Hinton, G. E., Osindero, S., & Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*, 1527–1554.

Hinton, G. E., & Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. In Rumelhart, D. E., & McClelland, J. L. (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*. MIT Press, Cambridge, MA.

Hinton, G. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, *14*, 1771–1800.

Hinton, G., Sejnowski, T., & Ackley, D. (1984). Boltzmann machines: Constraint satisfaction networks that learn. Tech. rep. TR-CMU-CS-84-119, Carnegie-Mellon University, Dept. of Computer Science.

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

Hinton, G. E., & Salakhutdinov, R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, *313*, 504–507.

Hinton, G. E., & Zemel, R. S. (1994). Autoencoders, minimum description length, and Helmholtz free energy. *Advances in Neural Information Processing Systems*, *6*, 3–10.

Japkowicz, N., Hanson, S. J., & Gluck, M. A. (2000). Nonlinear autoassociation is not equivalent to pca. *Neural Computation*, *12*(3), 531–545.

Larochelle, H., Erhan, D., Courville, A., Bergstra, J., & Bengio, Y. (2007). An empirical evaluation of deep architectures on problems with many factors of variation. In *Twenty-fourth International Conference on Machine Learning (ICML'2007)*.

Ranzato, M., Poultney, C., Chopra, S., & LeCun, Y. (2007). Efficient learning of sparse representations with an energy-based model. In et al., J. P. (Ed.), *Advances in Neural Information Processing Systems (NIPS 2006)*. MIT Press.

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536.

Salakhutdinov, R., & Hinton, G. (2007). Semantic hashing. In *Proceedings of the 2007 Workshop on Information Retrieval and applications of Graphical Models (SIGIR 2007)*.

Schwenk, H., & Milgram, M. (1995). Transformation invariant autoassociation with application to handwritten character recognition. In Tesauro, G., Touretzky, D., & Leen, T. (Eds.), *Advances in Neural Information Processing Systems 7*, pp. 991–998. MIT Press.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In Rumelhart, D., & McClelland, J. (Eds.), *Parallel Distributed Processing*, Vol. 1, chap. 6, pp. 194–281. MIT Press, Cambridge.

Taylor, G., Hinton, G., & Roweis, S. (2006). Modeling human motion using binary latent variables. In *Advances in Neural Information Processing Systems 20*. MIT Press.

Welling, M., Rosen-Zvi, M., & Hinton, G. (2005). Exponential family harmoniums with an application to information retrieval. In Saul, L., Weiss, Y., & Bottou, L. (Eds.), *Advances in Neural Information Processing Systems 17*. MIT Press.