Department of Information and Computer Science

Derivations of the Enhanced Gradient for the Boltzmann Machine

Tapani Raiko, KyungHyun Cho, and Alexander Ilin

Aalto University

SCIENCE +
TECHNOLOGY

TECHNICAL REPORT

Aalto University publication series
SCIENCE + TECHNOLOGY 20/2011

Derivations of the Enhanced Gradient for the Boltzmann Machine

Tapani Raiko, KyungHyun Cho, Alexander Ilin

Aalto University
School of Science
Department of Information and Computer Science
Bayesian Algorithms for Latent Variable Models

Aalto University publication series
SCIENCE + TECHNOLOGY 20/2011

ISBN 978-952-60-4295-4 (pdf)
ISSN-L 1799-4896
ISSN 1799-490X (pdf)

Aalto Print
Helsinki 2011

Finland

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Tapani Raiko, KyungHyun Cho, and Alexander Ilin

Name of the publication
Derivations of the Enhanced Gradient for the Boltzmann Machine

Publisher School of Science

Unit Department of Information and Computer Science

Series Aalto University publication series SCIENCE + TECHNOLOGY 20/2011

Field of research Machine Learning

Abstract

This technical report extends the conference paper (Cho et al., 2011) and the abstract (Raiko et al., 2011) with detailed derivations and proofs. First we recap notation that we use on the Boltzmann machine and its learning. Then we define transformations for the machine where some of its bits are flipped for all samples, and show the equivalence of the transformed model to the original one. Then we show that traditional update rules are not invariant to the transformations, propose a new update rule called the enhanced gradient, and finally show its invariance to the transformations.

Keywords machine learning, neural networks, deep learning, Boltzmann machine

ISBN (printed) **ISBN (pdf)** 978-952-60-4295-4

ISSN-L 1799-4896 **ISSN (printed)** 1799-4896 **ISSN (pdf)** 1799-490X

Location of publisher Espoo **Location of printing** Helsinki **Year** 2011

Pages 11

Derivations of the Enhanced Gradient for the Boltzmann Machine

Tapani Raiko, KyungHyun Cho, and Alexander Ilin

Aalto University

Abstract

This technical report extends the conference paper [1] and the abstract [2] with detailed derivations and proofs. First we recap notation that we use on the Boltzmann machine and its learning. Then we define transformations for the machine where some of its bits are flipped for all samples, and show the equivalence of the transformed model to the original one. Then we show that traditional update rules are not invariant to the transformations, propose a new update rule called the enhanced gradient, and finally show its invariance to the transformations.

1 Boltzmann Machine

Boltzmann machine has a binary state column vector $\mathbf{x}$, of which part is observed (visible) and part is latent (hidden) $\mathbf{x} = \left[\mathbf{x}_v^T \;\; \mathbf{x}_h^T\right]^T$. The machine has parameters $\boldsymbol{\theta}$ which include a square weight matrix $\mathbf{W}$ and a bias column vector $\mathbf{b}$. The state vector $\mathbf{x}$ varies for each sample[1], whereas parameters are constant. The probability of the machine being in state $\mathbf{x}$ is defined as

$$P(\mathbf{x} \mid \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left[-E(\mathbf{x} \mid \boldsymbol{\theta})\right] \tag{1}$$

$$E(\mathbf{x} \mid \boldsymbol{\theta}) = -\frac{1}{2}\mathbf{x}^T\mathbf{W}\mathbf{x} - \mathbf{b}^T\mathbf{x}, \tag{2}$$

where $Z(\boldsymbol{\theta}) = \sum_{\mathbf{x}} \exp\left[-E(\mathbf{x} \mid \boldsymbol{\theta})\right]$ is a normalizing constant called the partition function, and the $E(\mathbf{x} \mid \boldsymbol{\theta})$ is called the energy function. The weight matrix $\mathbf{W}$ is symmetric ($W_{ij} = W_{ji}$) and the diagonal elements are zero ($W_{ii} = 0$), that is,

[1] We do not use sample indices in the notation of this report.

connections between units are undirected, and a unit is not connected to itself.

2 Traditional Gradient

Parameters are typically learned to maximize likelihood. Data set or distribution d contains samples of the visible part of the state $\mathbf{x}_v$. The traditional gradient is obtained by taking a partial derivative of the log likelihood with respect to the parameters. It is

$$
\begin{aligned}
\frac{\partial \left\langle \log P(\mathbf{x}_v \mid \boldsymbol{\theta}) \right\rangle_d}{\partial W_{ij}} &= \frac{\partial}{\partial W_{ij}} \left\langle \log \sum_{\mathbf{x}_h} P(\mathbf{x}_v, \mathbf{x}_h \mid \boldsymbol{\theta}) \right\rangle_d && (3) \\
&= \frac{\partial}{\partial W_{ij}} \left\langle \log \frac{\sum_{\mathbf{x}_h} \exp[-E(\mathbf{x}_v, \mathbf{x}_h \mid \boldsymbol{\theta})]}{\sum_{\mathbf{x}} \exp[-E(\mathbf{x} \mid \boldsymbol{\theta})]} \right\rangle_d && (4) \\
&= \frac{\partial}{\partial W_{ij}} \left[\left\langle \log \sum_{\mathbf{x}_h} \exp \left(\frac{1}{2} \mathbf{x}^T \mathbf{W} \mathbf{x} + \mathbf{b}^T \mathbf{x} \right) \right\rangle_d \right. \\
&\quad \left. - \log \sum_{\mathbf{x}} \exp \left(\frac{1}{2} \mathbf{x}^T \mathbf{W} \mathbf{x} + \mathbf{b}^T \mathbf{x} \right) \right] && (5) \\
&= \left\langle \frac{\sum_{\mathbf{x}_h} \exp \left(\frac{1}{2} \mathbf{x}^T \mathbf{W} \mathbf{x} + \mathbf{b}^T \mathbf{x} \right) \frac{1}{2} x_i x_j}{\sum_{\mathbf{x}_h} \exp \left(\frac{1}{2} \mathbf{x}^T \mathbf{W} \mathbf{x} + \mathbf{b}^T \mathbf{x} \right)} \right\rangle_d \\
&\quad - \frac{\sum_{\mathbf{x}} \exp \left(\frac{1}{2} \mathbf{x}^T \mathbf{W} \mathbf{x} + \mathbf{b}^T \mathbf{x} \right) \frac{1}{2} x_i x_j}{\sum_{\mathbf{x}} \exp \left(\frac{1}{2} \mathbf{x}^T \mathbf{W} \mathbf{x} + \mathbf{b}^T \mathbf{x} \right)} && (6) \\
&= \frac{1}{2} \left\langle x_i x_j \right\rangle_d - \frac{1}{2} \left\langle x_i x_j \right\rangle_m . && (7)
\end{aligned}
$$

where $\langle \cdot \rangle_d$ is the expectation over the data distribution in which observed units are clamped to the data $\mathbf{x}_v$ and hidden units follow the conditional distribution of the model given the observed units $P(\mathbf{x}_h \mid \mathbf{x}_v, \boldsymbol{\theta})$, and $\langle \cdot \rangle_m$ is the expectation over the model distribution $P(\mathbf{x} \mid \boldsymbol{\theta})$ defined in Equation (1). In order to take into account the restriction $W_{ij} = W_{ji}$ we define as the gradient

$$
\begin{aligned}
\nabla W_{ij} &= \frac{\partial \left\langle \log P(\mathbf{x}_v \mid \boldsymbol{\theta}) \right\rangle_d}{\partial W_{ij}} + \frac{\partial \left\langle \log P(\mathbf{x}_v \mid \boldsymbol{\theta}) \right\rangle_d}{\partial W_{ji}} && (8) \\
&= \left\langle x_i x_j \right\rangle_d - \left\langle x_i x_j \right\rangle_m . && (9)
\end{aligned}
$$

For the bias term, we get the gradient similarly

$$
\begin{aligned}
\nabla b_i &= \frac{\partial \left\langle \log P(\mathbf{x}_v \mid \boldsymbol{\theta}) \right\rangle_d}{\partial b_i} && (10) \\
&= \left\langle \frac{\sum_{\mathbf{x}_h} \exp \left(\frac{1}{2} \mathbf{x}^T \mathbf{W} \mathbf{x} + \mathbf{b}^T \mathbf{x} \right) x_i}{\sum_{\mathbf{x}_h} \exp \left(\frac{1}{2} \mathbf{x}^T \mathbf{W} \mathbf{x} + \mathbf{b}^T \mathbf{x} \right)} \right\rangle_d - \frac{\sum_{\mathbf{x}} \exp \left(\frac{1}{2} \mathbf{x}^T \mathbf{W} \mathbf{x} + \mathbf{b}^T \mathbf{x} \right) x_i}{\sum_{\mathbf{x}} \exp \left(\frac{1}{2} \mathbf{x}^T \mathbf{W} \mathbf{x} + \mathbf{b}^T \mathbf{x} \right)} \\
& && (11) \\
&= \left\langle x_i \right\rangle_d - \left\langle x_i \right\rangle_m . && (12)
\end{aligned}
$$

A gradient ascent update is

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \eta \nabla \boldsymbol{\theta}, \tag{13}$$

where η is a constant step size.

3 Bit-Flip Transformations

Let us define a transformation where some of the bits of a Boltzmann machine are flipped. This can be seen as another representation of the data for the visible units, and another representation of the hidden data for the hidden units. A binary vector $\mathbf{f}$ is a constant to all samples and it indicates which bits are flipped. The transformation is

$$\tilde{x}_i = x_i^{1-f_i}(1-x_i)^{f_i} \tag{14}$$

$$\tilde{W}_{ij} = (-1)^{f_i+f_j} W_{ij} \tag{15}$$

$$\tilde{b}_i = (-1)^{f_i} \left(b_i + \sum_j f_j W_{ij} \right). \tag{16}$$

Theorem 3.1. *The transformed machine is equivalent to the original one, that is,* $P(\tilde{\mathbf{x}} \mid \tilde{\boldsymbol{\theta}}) = P(\mathbf{x} \mid \boldsymbol{\theta})$ *for all states* $\mathbf{x}$.

Proof.

$$\begin{aligned}
E(\tilde{\mathbf{x}} \mid \tilde{\boldsymbol{\theta}}) &= -\frac{1}{2}\tilde{\mathbf{x}}^T \tilde{\mathbf{W}} \tilde{\mathbf{x}} - \tilde{\mathbf{b}}^T \tilde{\mathbf{x}} & (17)\\
&= \sum_{ij} -\frac{1}{2} x_i^{1-f_i}(1-x_i)^{f_i}(-1)^{f_i+f_j} W_{ij} x_j^{1-f_j}(1-x_j)^{f_j} \\
&\quad - \sum_i (-1)^{f_i} \left(b_i + \sum_j f_j W_{ij} \right) x_i^{1-f_i}(1-x_i)^{f_i} & (18)\\
&= \sum_{ij} \frac{1}{2}(x_i - f_i) W_{ij} (x_j - f_j) - \sum_i \left(b_i + \sum_j f_j W_{ij} \right)(x_i - f_i) & (19)\\
&= \sum_{ij} -\frac{1}{2} x_i W_{ij} x_j - \sum_i b_i x_i + \sum_{ij} \frac{1}{2} x_i W_{ij} f_j + \sum_{ij} \frac{1}{2} x_i W_{ij} x_j \\
&\quad - \sum_{ij} \frac{1}{2} f_i W_{ij} f_j - \sum_{ij} f_j W_{ij} x_i + \sum_i \left(b_i + \sum_j f_j W_{ij} \right) f_i & (20)\\
&= \sum_{ij} -\frac{1}{2} x_i W_{ij} x_j - \sum_i b_i x_i \\
&\quad - \sum_{ij} \frac{1}{2} f_i W_{ij} f_j + \sum_i \left(b_i + \sum_j f_j W_{ij} \right) f_i & (21)\\
&= E(\mathbf{x} \mid \boldsymbol{\theta}) + \text{const.} & (22)
\end{aligned}$$

Since the energy functions are the same up to a constant, the probability distributions P are the same. □

Theorem 3.2. *Transforming twice with the same vector* $\mathbf{f}$ *results in the original machine, that is,* $\tilde{\tilde{x}}_i = x_i$, $\tilde{\tilde{W}}_{ij} = W_{ij}$, *and* $\tilde{\tilde{b}}_i = b_i$.

Proof. Flipping a binary state twice results in the original state so $\tilde{\tilde{x}}_i = x_i$ obviously.

$$\begin{aligned}
\tilde{\tilde{W}}_{ij} &= (-1)^{f_i+f_j}\tilde{W}_{ij} && (23)\\
&= (-1)^{f_i+f_j}(-1)^{f_i+f_j}W_{ij} && (24)\\
&= W_{ij}, && (25)\\
\tilde{\tilde{b}}_i &= (-1)^{f_i}\left(\tilde{b}_i + \sum_j f_j \tilde{W}_{ij}\right) && (26)\\
&= (-1)^{f_i}\left[(-1)^{f_i}\left(b_i + \sum_j f_j W_{ij}\right) + \sum_j f_j (-1)^{f_i+f_j} W_{ij}\right] && (27)\\
&= b_i + \sum_j f_j W_{ij} - \sum_j f_j W_{ij} && (28)\\
&= b_i. && (29)
\end{aligned}$$

□

4 Transform, Update, Transform back

Let us study what happens when a Boltzmann machine is first transformed using Equations (14)–(16), then update using the traditional gradient in Equation (13), and finally transformed back using the same transformation vector $\mathbf{f}$. One might expect to get the same updated model regardless of the transformation, but that is not the case.

First we define

$$\begin{aligned}
\mathrm{cov}_P(x_i, x_j) &= \langle x_i x_j \rangle_P - \langle x_i \rangle_P \langle x_j \rangle_P && (30)\\
\langle \cdot \rangle_{dm} &= \frac{\langle \cdot \rangle_d + \langle \cdot \rangle_m}{2} && (31)
\end{aligned}$$

and note that

$$\begin{aligned}
\langle x_i \rangle_{dm} \nabla b_j = \frac{1}{2}\langle x_i \rangle_d \langle x_j \rangle_d + \frac{1}{2}\langle x_i \rangle_m \langle x_j \rangle_d \\
- \frac{1}{2}\langle x_i \rangle_d \langle x_j \rangle_m - \frac{1}{2}\langle x_i \rangle_m \langle x_j \rangle_m && (32)
\end{aligned}$$

to help get the form in Equation (37). The three-step update for the weights is

$$
\begin{aligned}
W_{ij} \leftarrow (-1)^{f_i+f_j}\left[(-1)^{f_i+f_j}W_{ij} + \eta\left(\langle\tilde{x}_i\tilde{x}_j\rangle_d - \langle\tilde{x}_i\tilde{x}_j\rangle_m\right)\right] && (33)\\
= W_{ij} + \eta(-1)^{f_i+f_j}\left[\left\langle x_i^{1-f_i}(1-x_i)^{f_i}x_j^{1-f_j}(1-x_j)^{f_j}\right\rangle_d\right.\\
\left. - \left\langle x_i^{1-f_i}(1-x_i)^{f_i}x_j^{1-f_j}(1-x_j)^{f_j}\right\rangle_m\right] && (34)\\
= W_{ij} + \eta\left[\langle(x_i-f_i)(x_j-f_j)\rangle_d - \langle(x_i-f_i)(x_j-f_j)\rangle_m\right] && (35)\\
= W_{ij} + \eta\left[\langle x_ix_j\rangle_d - \langle x_ix_j\rangle_m - f_i\nabla b_j - f_j\nabla b_i\right] && (36)\\
= W_{ij} + \eta\left[\mathrm{cov}_d(x_i,x_j) - \mathrm{cov}_m(x_i,x_j) + \left(\langle x_i\rangle_{dm} - f_i\right)\nabla b_j + \left(\langle x_j\rangle_{dm} - f_j\right)\nabla b_i\right]. && (37)
\end{aligned}
$$

Let us use a shorthand $\nabla_{\mathbf{f}}W_{ij} = \langle x_ix_j\rangle_d - \langle x_ix_j\rangle_m - f_i\nabla b_j - f_j\nabla b_i$ for the resulting gradient when using the transformation $\mathbf{f}$.

The three-step update for the bias is

$$
b_i \leftarrow (-1)^{f_i}\left\{\tilde{b}_i + \eta\left(\langle\tilde{x}_i\rangle_d - \langle\tilde{x}_i\rangle_m\right) + \sum_j f_j\left[\tilde{W}_{ij} + \eta\left(\langle\tilde{x}_i\tilde{x}_j\rangle_d - \langle\tilde{x}_i\tilde{x}_j\rangle_m\right)\right]\right\} \quad (38)
$$

$$
\begin{aligned}
= (-1)^{f_i}\left\{(-1)^{f_i}\left(b_i + \sum_j f_jW_{ij}\right) + \eta\left[\left\langle x_i^{1-f_i}(1-x_i)^{f_i}\right\rangle_d\right.\right.\\
\left.\left. - \left\langle x_i^{1-f_i}(1-x_i)^{f_i}\right\rangle_m\right] + \sum_j f_j(-1)^{f_i+f_j}\left[W_{ij} + \eta\nabla_{\mathbf{f}}W_{ij}\right]\right\} && (39)\\
= b_i + \sum_j f_jW_{ij} + \eta\left(\langle x_i\rangle_d - \langle x_i\rangle_m\right) + \sum_j f_j\left(-W_{ij} - \eta\nabla_{\mathbf{f}}W_{ij}\right) && (40)\\
= b_i + \eta\left(\nabla b_i - \sum_j f_j\nabla_{\mathbf{f}}W_{ij}\right). && (41)
\end{aligned}
$$

From Equations (37) and (41) we can note that the transformation vectors $\mathbf{f}$ do not cancel out and thus we end up with different parameters after the update depending on the transformation.

5 Enhanced Gradient

We propose a so called enhanced gradient as a better alternative to the traditional gradient for learning Boltzmann machines. The enhanced gradient is

$$\begin{aligned}
\nabla_e W_{ij} &= \mathrm{cov}_d(x_i, x_j) - \mathrm{cov}_m(x_i, x_j) && (42)\\
&= \langle x_i x_j \rangle_d - \langle x_i x_j \rangle_m - \frac{\langle x_i \rangle_d + \langle x_i \rangle_m}{2} \nabla b_j - \frac{\langle x_j \rangle_d + \langle x_j \rangle_m}{2} \nabla b_i && (43)\\
\nabla_e b_i &= \nabla b_i - \sum_j \langle x_j \rangle_{dm} \nabla_e W_{ij} && (44)\\
&= \nabla b_i - \sum_j \frac{\langle x_j \rangle_d + \langle x_j \rangle_m}{2} \nabla_e W_{ij} && (45)
\end{aligned}$$

and we will show in Section 6 that it is invariant to the bit-flipping transformations.

The enhanced gradient is a weighted average of all possible updates with different transformation vectors $\mathbf{f}$ as described in the previous section. We chose the weights for the different updates inspired by Equation (37) to be

$$\prod_i \langle x_i \rangle_{dm}^{f_i} \left(1 - \langle x_i \rangle_{dm}\right)^{1-f_i} \qquad (46)$$

that sum up to one when considering all the exponentially many alternative transformation vectors $\mathbf{f}$.

The enhanced gradient for the weights is derived as follows

$$\begin{aligned}
\nabla_e W_{ij} = &\sum_{\mathbf{f} \in \{0,1\}^n} \left[\prod_k \langle x_k \rangle_{dm}^{f_k} \left(1 - \langle x_k \rangle_{dm}\right)^{1-f_k} \right]\\
&\left[\mathrm{cov}_d(x_i, x_j) - \mathrm{cov}_m(x_i, x_j) + (\langle x_i \rangle_{dm} - f_i) \nabla b_j + (\langle x_j \rangle_{dm} - f_j) \nabla b_i \right] && (47)\\
= &\sum_{f_i, f_j \in \{0,1\}} \prod_{k, i \neq k \neq j} \left[\langle x_k \rangle_{dm} + (1 - \langle x_k \rangle_{dm}) \right]\\
&\langle x_i \rangle_{dm}^{f_i} \left(1 - \langle x_i \rangle_{dm}\right)^{1-f_i} \langle x_j \rangle_{dm}^{f_j} \left(1 - \langle x_j \rangle_{dm}\right)^{1-f_j}\\
&\left[\mathrm{cov}_d(x_i, x_j) - \mathrm{cov}_m(x_i, x_j) + (\langle x_i \rangle_{dm} - f_i) \nabla b_j + (\langle x_j \rangle_{dm} - f_j) \nabla b_i \right] && (48)\\
= &\sum_{f_i, f_j \in \{0,1\}} \langle x_i \rangle_{dm}^{f_i} \left(1 - \langle x_i \rangle_{dm}\right)^{1-f_i} \langle x_j \rangle_{dm}^{f_j} \left(1 - \langle x_j \rangle_{dm}\right)^{1-f_j}\\
&\left[\mathrm{cov}_d(x_i, x_j) - \mathrm{cov}_m(x_i, x_j) + (\langle x_i \rangle_{dm} - f_i) \nabla b_j + (\langle x_j \rangle_{dm} - f_j) \nabla b_i \right] && (49)
\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{cov}_d(x_i, x_j) - \mathrm{cov}_m(x_i, x_j) \\
&\quad + (1 - \langle x_i \rangle_{dm})(1 - \langle x_j \rangle_{dm}) \left[\langle x_i \rangle_{dm} \nabla b_j + \langle x_j \rangle_{dm} \nabla b_i \right] \\
&\quad + \langle x_i \rangle_{dm} (1 - \langle x_j \rangle_{dm}) \left[(\langle x_i \rangle_{dm} - 1) \nabla b_j + \langle x_j \rangle_{dm} \nabla b_i \right] \\
&\quad + (1 - \langle x_i \rangle_{dm}) \langle x_j \rangle_{dm} \left[\langle x_i \rangle_{dm} \nabla b_j + (\langle x_j \rangle_{dm} - 1) \nabla b_i \right] \\
&\quad + \langle x_i \rangle_{dm} \langle x_j \rangle_{dm} \left[(\langle x_i \rangle_{dm} - 1) \nabla b_j + (\langle x_j \rangle_{dm} - 1) \nabla b_i \right] && (50) \\
&= \mathrm{cov}_d(x_i, x_j) - \mathrm{cov}_m(x_i, x_j) && (51)
\end{aligned}
$$

For the bias term, it would be possible to derive similarly

$$
\begin{aligned}
\nabla'_e b_i = &\sum_{\mathbf{f} \in \{0,1\}^n} \left[\prod_k \langle x_k \rangle_{dm}^{f_k} (1 - \langle x_k \rangle_{dm})^{1-f_k} \right] \\
&\left[\nabla b_i - \sum_j f_j \left(\nabla W_{ij} - f_i \nabla b_j - f_j \nabla b_i \right) \right] && (52) \\
= &\nabla b_i - \sum_j \sum_{f_i, f_j \in \{0,1\}} \langle x_i \rangle_{dm}^{f_i} (1 - \langle x_i \rangle_{dm})^{1-f_i} \langle x_j \rangle_{dm}^{f_j} (1 - \langle x_j \rangle_{dm})^{1-f_j} \\
&f_j \left(\nabla W_{ij} - f_i \nabla b_j - f_j \nabla b_i \right) && (53)
\end{aligned}
$$

$$
= \nabla b_i - \sum_j \langle x_j \rangle_{dm} \left[(1 - \langle x_i \rangle_{dm})(\nabla W_{ij} - \nabla b_i) + \langle x_i \rangle_{dm} \left(\nabla W_{ij} - \nabla b_j - \nabla b_i \right) \right] \tag{54}
$$

$$
= \nabla b_i - \sum_j \langle x_j \rangle_{dm} \left(\nabla W_{ij} - \nabla b_i - \langle x_i \rangle_{dm} \nabla b_j \right). \tag{55}
$$

However, there is an alternative formulation that turns out to be more elegant in form and to work slightly better in practice. The update of the bias depends on the update of the weights as seen in Equation (41). Rather than using a different weight update $\nabla_{\mathbf{f}} \mathbf{W}$ for each bias update, we use the enhanced gradient for the weights when doing each bias update, that is, replacing Equation (41) by $b_i \leftarrow b_i + \eta \left(\nabla b_i - \sum_j f_j \nabla_e W_{ij} \right)$. Now we get the final formulation of the enhanced gradient for the bias:

$$
\begin{aligned}
\nabla_e b_i = &\sum_{\mathbf{f} \in \{0,1\}^n} \left[\prod_k \langle x_k \rangle_{dm}^{f_k} (1 - \langle x_k \rangle_{dm})^{1-f_k} \right] \left(\nabla b_i - \sum_j f_j \nabla_e W_{ij} \right) && (56) \\
&= \nabla b_i - \sum_j \sum_{f_i, f_j \in \{0,1\}} \langle x_i \rangle_{dm}^{f_i} (1 - \langle x_i \rangle_{dm})^{1-f_i} \langle x_j \rangle_{dm}^{f_j} (1 - \langle x_j \rangle_{dm})^{1-f_j} f_j \nabla_e W_{ij} && (57) \\
&= \nabla b_i - \sum_j \langle x_j \rangle_{dm} \left[(1 - \langle x_i \rangle_{dm}) \nabla_e W_{ij} + \langle x_i \rangle_{dm} \nabla_e W_{ij} \right] && (58) \\
&= \nabla b_i - \sum_j \langle x_j \rangle_{dm} \nabla_e W_{ij}. && (59)
\end{aligned}
$$

6 Invariance of the Enhanced Gradient

Theorem 6.1. *The enhanced gradient*

$$\begin{aligned}\nabla_e W_{ij} &= \mathrm{cov}_d(x_i, x_j) - \mathrm{cov}_m(x_i, x_j) && (60)\\ \nabla_e b_i &= \nabla b_i - \sum_j \langle x_j \rangle_{dm} \nabla_e W_{ij} && (61)\end{aligned}$$

is invariant to the bit-flipping transformations as described in Section 3.

Proof. We again compose a three-step update consisting of a transformation, update by enhanced gradient, and transformation back. If the resulting model is the same regardless of the transformation vector $\mathbf{f}$, we have proven the claim. The combined update for the weights is

$$\begin{aligned} W_{ij} &\leftarrow (-1)^{f_i+f_j}\left[\tilde{W}_{ij} + \eta\left(\mathrm{cov}_d(\tilde{x}_i, \tilde{x}_j) - \mathrm{cov}_m(\tilde{x}_i, \tilde{x}_j)\right)\right] && (62)\\ &= W_{ij} + (-1)^{f_i+f_j}\eta\Big[\,\mathrm{cov}_d(x_i^{1-f_i}(1-x_i)^{f_i}, x_j^{1-f_j}(1-x_j)^{f_j})\\ &\quad - \mathrm{cov}_m(x_i^{1-f_i}(1-x_i)^{f_i}, x_j^{1-f_j}(1-x_j)^{f_j})\Big] && (63)\\ &= W_{ij} + \eta\left[\mathrm{cov}_d(x_i, x_j) - \mathrm{cov}_m(x_i, x_j)\right]. && (64)\end{aligned}$$

The combined update for the bias is

$$\begin{aligned} b_i &\leftarrow (-1)^{f_i}\left[\tilde{b}_i + \eta\nabla_e\tilde{b}_i + \sum_j f_j(\tilde{W}_{ij} + \eta\nabla_e\tilde{W}_{ij})\right] && (65)\\ &= (-1)^{f_i}\Bigg\{(-1)^{f_i}\left(b_i + \sum_j f_j W_{ij}\right) + \eta\Big[\,\langle\tilde{x}_i\rangle_d - \langle\tilde{x}_i\rangle_m\\ &\quad - \sum_j \langle\tilde{x}_j\rangle_{dm}\left(\langle\tilde{x}_i\tilde{x}_j\rangle_d - \langle\tilde{x}_i\tilde{x}_j\rangle_m - \langle\tilde{x}_j\rangle_{dm}\langle\tilde{x}_i\rangle_d + \langle\tilde{x}_j\rangle_{dm}\langle\tilde{x}_i\rangle_m - \langle\tilde{x}_i\rangle_{dm}\langle\tilde{x}_j\rangle_d + \langle\tilde{x}_i\rangle_{dm}\langle\tilde{x}_j\rangle_m\right)\Big]\\ &\quad + \sum_j f_j\Big[(-1)^{f_i+f_j}W_{ij} + \eta\left(\langle\tilde{x}_i\tilde{x}_j\rangle_d - \langle\tilde{x}_i\tilde{x}_j\rangle_m - \langle\tilde{x}_i\rangle_d\langle\tilde{x}_j\rangle_d + \langle\tilde{x}_i\rangle_m\langle\tilde{x}_j\rangle_m\right)\Big]\Bigg\}\\ & && (66)\end{aligned}$$

$$
\begin{aligned}
&= b_i + \eta \Bigg\{ \langle x_i - f_i \rangle_d - \langle x_i - f_i \rangle_m - \sum_j \Bigg[\langle x_j - f_j \rangle_{dm} \Big(\langle (x_i - f_i)(x_j - f_j) \rangle_d \\
&\quad - \langle (x_i - f_i)(x_j - f_j) \rangle_m - \langle x_j - f_j \rangle_{dm} \langle x_i - f_i \rangle_d + \langle x_j - f_j \rangle_{dm} \langle x_i - f_i \rangle_m \\
&\quad - \langle x_i - f_i \rangle_{dm} \langle x_j - f_j \rangle_d + \langle x_i - f_i \rangle_{dm} \langle x_j - f_j \rangle_m \Big) \\
&\quad + f_j \left(\langle x_i x_j \rangle_d - \langle x_i x_j \rangle_m - \langle x_i \rangle_d \langle x_j \rangle_d + \langle x_i \rangle_m \langle x_j \rangle_m \right) \Bigg] \Bigg\}
\end{aligned} \tag{67}
$$

$$
\begin{aligned}
&= b_i + \eta \Bigg\{ \nabla b_i - \sum_j \Bigg[\left(\langle x_j \rangle_{dm} - f_j \right) \\
&\quad \left(\nabla W_{ij} - f_j \nabla b_i - f_i \nabla b_j - \langle x_i \rangle_{dm} \nabla b_j + f_i \nabla b_j - \langle x_j \rangle_{dm} \nabla b_i + f_j \nabla b_i \right) \\
&\quad + f_j \left(\nabla W_{ij} - \langle x_i \rangle_{dm} \nabla b_j - \langle x_j \rangle_{dm} \nabla b_i \right) \Bigg] \Bigg\}
\end{aligned} \tag{68}
$$

$$
\begin{aligned}
&= b_i + \eta \Bigg[\nabla b_i - \sum_j \Big(\langle x_j \rangle_{dm} \nabla W_{ij} - \langle x_j \rangle_{dm} \langle x_i \rangle_{dm} \nabla b_j - \langle x_j \rangle_{dm}^2 \nabla b_i \\
&\quad - f_j \nabla W_{ij} + f_j \langle x_i \rangle_{dm} \nabla b_j + f_j \langle x_j \rangle_{dm} \nabla b_i \\
&\quad + f_j \nabla W_{ij} - f_j \langle x_i \rangle_{dm} \nabla b_j - f_j \langle x_j \rangle_{dm} \nabla b_i \Big) \Bigg]
\end{aligned} \tag{69}
$$

$$
= b_i + \eta \left[\nabla b_i - \sum_j \langle x_j \rangle_{dm} \left(\nabla W_{ij} - \langle x_i \rangle_{dm} \nabla b_j - \langle x_j \rangle_{dm} \nabla b_i \right) \right] \tag{70}
$$

$$
= b_i + \eta \left(\nabla b_i - \sum_j \langle x_j \rangle_{dm} \nabla_e W_{ij} \right). \tag{71}
$$

□

Bibliography

[1] KyungHyun Cho, Tapani Raiko, and Alexander Ilin. Enhanced gradient and adaptive learning rate for training restricted Boltzmann machines. In *Proceedings of the International Conference on Machine Learning (ICML 2011)*, Bellevue, Washington, USA, June 2011.

[2] Tapani Raiko, KyungHyun Cho, and Alexander Ilin. Enhanced gradient for learning Boltzmann machines (abstract). In *The Learning Workshop*, Fort Lauderdale, Florida, April 2011.

This technical report extends the conference paper (Cho et al., 2011) and the abstract (Raiko et al., 2011) with detailed derivations and proofs. First we recap notation that we use on the Boltzmann machine and its learning. Then we define transformations for the machine where some of its bits are flipped for all samples, and show the equivalence of the transformed model to the original one. Then we show that traditional update rules are not invariant to the transformations, propose a new update rule called the enhanced gradient, and finally show its invariance to the transformations.

Aalto-ST 20/2011

ISBN 978-952-60-4295-4 (pdf)
ISBN 978-952-60-4294-7
ISSN-L 1799-4896
ISSN 1799-490X (pdf)
ISSN 1799-4896

Aalto University
School of Science
Department of Information and Computer Science
www.aalto.fi

BUSINESS +
ECONOMY

ART +
DESIGN +
ARCHITECTURE

SCIENCE +
TECHNOLOGY

CROSSOVER

DOCTORAL
DISSERTATIONS