On the Convergence Properties of Contrastive Divergence

Ilya Sutskever
University of Toronto

Tijmen Tieleman
University of Toronto

Abstract

Contrastive Divergence (CD) is a popular method for estimating the parameters of Markov Random Fields (MRFs) by rapidly approximating an intractable term in the gradient of the log probability. Despite CD's empirical success, little is known about its theoretical convergence properties.

In this paper, we analyze the CD_1 update rule for Restricted Boltzmann Machines (RBMs) with binary variables. We show that this update is not the gradient of any function, and construct a counterintuitive "regularization function" that causes CD learning to cycle indefinitely. Nonetheless, we show that the regularized CD update has a fixed point for a large class of regularization functions using Brower's fixed point theorem.

1 INTRODUCTION

Markov Random Fields (MRFs) are an important class of probabilistic models that are useful for denoising, prediction, and density estimation (Cross and Jain, 1981; Malfait and Roose, 1997; Portilla et al., 2003; Roth and Black, 2005; Li, 1994; Wainwright, 2008). In particular, MRFs subsume the Restricted Boltzmann Machines (RBMs) (Hinton, 2002; Smolensky, 1986), which are essential for learning Deep Belief Networks (Hinton et al., 2006; Bengio et al., 2007; Hinton and Salakhutdinov, 2006).

Nearly every application of MRFs requires estimating their parameters from data. The natural maximum-likelihood parameter estimation is challenging, because the log probability's gradient is the difference of two expectations, of which one cannot be easily computed. As a result, a number of approximate parameter estimation methods have been developed. Pseudolikelihood (Besag, 1977) and Score Matching (Hyvarinen, 2006) are tractable alternatives to the log probability objective which are easier to optimize, and Loopy Belief Propagation and its variants (Wainwright, 2008) directly approximate the intractable expectation in the gradient. This paper focuses on Contrastive Divergence (CD) (Hinton, 2002), which directly approximates the intractable expectation with an easy Monte Carlo estimate. Being trivial to implement, CD is widely used (Hinton et al., 2006), but its convergence properties are not entirely understood.

In this paper we gain a better understanding of the noiseless CD_1 update rule for binary RBMs, and report the following results:

- We provide two proofs showing that the CD update is not the gradient of any objective function. This result was first proved by Tieleman (2007) and stated by Bengio and Delalleau (2009).
- We construct an example of a nonconvex regularization function that causes the CD update to cycle indefinitely.
- The CD update is shown to have at least one fixed point when used with L_2 regularization.

2 RELATED WORK

There has been much work attempting to elucidate the convergence properties of CD. Some of this work shows that CD minimizes a known cost function when used with specific Markov chains. For example, if the Markov chain used to estimate the intractable expectation (Hinton, 2002) is the Langevin Monte Carlo method, then CD computes the gradient of the score-matching objective function; similarly, when the Markov chain samples a random component of the data vector from its conditional distribution, then CD becomes the gradient of the log pseudo-likelihood of the model (Hyvarinen, 2007). Other work has provided general conditions under which CD converges to the maximum likelihood solution (Yuille, 2004), which

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

mainly depend on the rate of convergence of the said Markov chain. A continuous-time version of CD is also known to converge to the maximum-likelihood solution for Gaussian models under broad conditions (Movellan, 2008). An analysis of CD in terms of an expansion of the log probability is found by Bengio and Delalleau (2009). However, we are unaware of convergence-related theoretical results that are applicable to the commonly used CD_1 for training binary Restricted Boltzmann Machines.

3 PRELIMINARIES

In this section, we describe and define the Restricted Boltzmann Machine and the Contrastive Divergence update.

3.1 RESTRICTED BOLTZMANN MACHINES

A binary RBM defines a probability distribution over binary vectors $V \in \{0,1\}^n$ and $H \in \{0,1\}^m$ by the expression

$$P(V,H) = \frac{\exp(-E(V,H))}{Z} \tag{1}$$

where the energy $E(V,H)$ is defined as

$$E(V,H) = -V^\top W H - V^\top b_V - H^\top b_H \tag{2}$$

and the partition function Z is

$$Z = \sum_{V\in\{0,1\}^n} \sum_{H\in\{0,1\}^m} \exp(-E(V,H)) \tag{3}$$

The marginal probability $P(V)$ is $\sum_{H\in\{0,1\}^m} P(V,H)$, and its logarithm $\log P(V)$ is equal to

$$V^\top b_V + \sum_{j=1}^{m} \log\left(1+\exp\left(V^\top W_{(:,j)} + b_{H\,j}\right)\right) - \log Z$$

A standard way of estimating the RBM's parameters from a training set $\{V_1,\ldots,V_N\}$ is by finding the parameters that maximize the average log probability

$$L = E_{D(V)}[\log P(V)] \tag{4}$$

where $D(V)$ is the empirical data distribution (which is a uniform mixture of delta distributions, one for each training point). The parameters maximizing the average log probability L are typically found with a gradient-based optimization method. The gradient of L with respect to the weights of the RBM is given by

$$\begin{aligned}
\frac{\partial L}{\partial b_V} &= E_{D(V,H)}[V] - E_{P(V,H)}[V] \qquad (5)\\
\frac{\partial L}{\partial b_H} &= E_{D(V,H)}[H] - E_{P(V,H)}[H]\\
\frac{\partial L}{\partial W} &= E_{D(V,H)}\left[VH^\top\right] - E_{P(V,H)}\left[VH^\top\right]
\end{aligned}$$

where $D(V,H) = P(H|V)D(V)$. In each equation, expectations with respect to $D(V,H)$ are easy to estimate because $D(V)$ is trivial to sample and the distribution $P(H|V)$ is factorial (eq. 1 implies $P(H|V) = \prod_{j=1}^{m} P(H_j|V)$). In contrast, there is no easy way to estimate expectation with respect to $P(V,H)$.

3.2 CONTRASTIVE DIVERGENCE

Contrastive Divergence (CD) (Hinton, 2002) is an algorithmically efficient procedure for RBM parameter estimation. The CD update is obtained by replacing the distribution $P(V,H)$ with a distribution $R(V,H)$ in eq. 5

$$\begin{aligned}
\Delta^{b_V} CD(W) &= E_{D(V,H)}[V] - E_{R(V,H)}[V] \qquad (6)\\
\Delta^{b_H} CD(W) &= E_{D(V,H)}[H] - E_{R(V,H)}[H]\\
\Delta^{W} CD(W) &= E_{D(V,H)}\left[VH^\top\right] - E_{R(V,H)}\left[VH^\top\right]
\end{aligned}$$

Drawing samples from $R(V,H)$ is cheaply done as follows:

1. Let $V' \sim D(V)$ be a random training point.
2. Sample H' from $P(H'|V')$.
3. Sample V from $P(V|H')$.
4. Sample H from $P(H|V)$.
5. return (V,H).

Although the distribution R is obtained by starting at the data distribution and running the Gibbs sampling Markov chain for one step, the term $E_{R(V,H)}\left[VH^\top\right]$ still reflects the kind of data the model "likes", causing the difference to often point in a direction of improvement. See the discussion in (Hinton, 2002).

In addition, CD has an "objective function",

$$CD(W) = E_{P(H|V)D(V)}[-E(V,H)] - E_{R(V,H)}[-E(V,H)], \tag{7}$$

whose gradients $\nabla CD(W)$ are close—but not equal—to the CD update $\Delta CD(W)$ (Hinton, 2002).

Standard CD learning proceeds by repeatedly changing the RBM parameters W according to the CD update $\Delta CD(W)$ with some learning rate[1]. The regularized CD update, $\Delta CD_F(W)$, with the regularization function F, is defined by the equation

$$\Delta CD_F(W) = \Delta CD(W) + \nabla F(W) \tag{8}$$

so, for example, if $F(W) = -\lambda/2 \cdot \|W\|$, then $\Delta CD_F(W)$ is the CD update with L_2 weight decay.

[1]We abuse the notation ΔCD and "the CD update": an actual update must involve a learning rate which we omit.

4 THE CD UPDATE DIRECTION IS NOT THE GRADIENT OF ANY FUNCTION

Finding a function $H(W)$ whose gradients are precisely equal to ΔCD would shed light on the type of solutions that CD tends to find, and would also let us conclude that CD always converges. However, in this section we prove that ΔCD is not a gradient of a function. Even more surprisingly, we construct a regularization function that causes the CD update to cycle indefinitely around a circle.

4.1 FIRST PROOF

We present two proofs showing that the CD update is not the gradient of any function.

Assume, in order to obtain a contradiction, that there is a function H such that $\nabla H = \Delta CD$. Consider an RBM with one visible unit and one hidden unit, and a single training point where the visible unit is in the zero state. The bias to the hidden unit is fixed to zero, so that there are only two parameters: w, the connection between the two units, and b_V, the bias to the visible unit. The parameters are jointly written as $W = (w, b_V)$.

Now, because ΔCD is sufficiently smooth, basic calculus states that

$$\begin{aligned} \frac{\partial^2 H(W)}{\partial w \partial b_V} &= \frac{\partial^2 H(W)}{\partial b_V \partial w} \\ \frac{\partial}{\partial w}\frac{\partial}{\partial b_V} H(W) &= \frac{\partial}{\partial b_V}\frac{\partial}{\partial w} H(W) \\ \frac{\partial}{\partial w}\Delta^{b_V} CD(W) &= \frac{\partial}{\partial b_V}\Delta^{w} CD(W) \end{aligned} \tag{9}$$

is valid for all W.

We investigated whether those are equal, and found, with $b_V = 0$ and $w = 1$, that they are not. A straightforward but tedious derivation reveals that

$$\frac{\partial}{\partial w}\Delta^{b_V} CD - \frac{\partial}{\partial b}\Delta^{w} CD = \frac{e \cdot (e-1) \cdot (e+3)}{8(e+1)^3} \tag{10}$$

where $e = \exp(1)$. None of the terms in the numerator are zero, so these two partial derivatives are not equal, implying that no such function H exists.

4.2 SECOND PROOF

Our second proof involves a simulation and is important for constructing a regularization function that causes CD to cycle; it is essentially the proof in (Tieleman, 2007). As before, suppose that ΔCD were the gradient of a function H. We can use the CD update to compute the change in the value of H as we travel along a path in the parameter space. If the path's initial and final points are equal, then the total change in the value of H, as measured by ΔCD, its gradient, must be zero. Therefore, if we use the CD update to numerically evaluate the change in H along a closed loop, and find that the cumulative change in the value of this presumed H is nonzero, then there cannot be any function H of which ΔCD is the gradient.

Our example is the same 2-parameter RBM as before. The path we follow, $\gamma(t)$, traverses a circle whose radius is $1/(2\pi)$ and whose center is $(2, 0)$ as t traverses the interval $[0, 1]$; this radius ensures that the length of the path is 1. The path γ begins at $(2, 1/(2\pi))$ and proceeds in counter-clockwise direction. In particular, $\gamma(0) = \gamma(1)$.

We computed the total change in the supposed function H along the path γ using the CD update, which turned out to be 0.008242 (see fig. 1); if the CD update were the gradient of a function, this total change would be zero. The total change is computed by traversing the path with 10^5 equally-spaced steps, where at each step, the change in the function's value is proportional to the length of the projection of the CD update onto the path's direction.[2]

4.3 STRONGER NEGATIVE STATEMENTS

While we have shown that ΔCD is not the gradient of any objective function, CD can still be minimizing a function by having a positive inner-product with this function's gradient. This would imply that CD learning is guaranteed to find a local optimum of H, and that CD learning never cycles.

We did not manage to prove that there cannot be such an objective function. However, the following is true: if there is no function whose gradient has a positive inner-product with CD, then CD learning must sometimes fail to converge.

Indeed, if CD always converges then the total length of the path followed by CD learning, from θ to the convergence point that it reaches, is clearly being minimized by CD. Therefore, this function's gradient must be positively correlated with CD.

[2] We are reasonably certain that the reported value of the function H is computed accurately, because we tried traversing the path with 10^6 steps, and obtained an answer whose 6 significant digits match 0.008242. Since the rectangle method is used to compute this integral (whose error is $O(n^{-2})$), we can be fairly confident in this value.

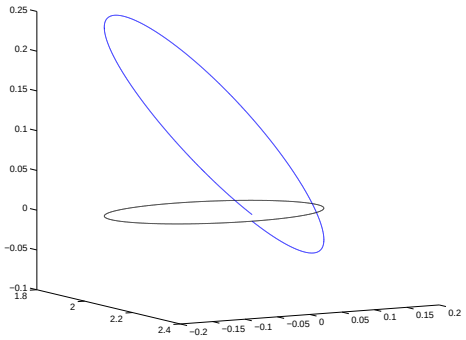

Figure 1: The surface of the supposed function H implied by the CD update along the path γ described in the text (blue curve). The figure shows that the altitude of the final point is slightly greater than the altitude of the initial point. Although the total elevation along this closed path is slight, its existence is sufficient to deduce that there is no function whose gradient is equal to the CD update. The black curve is the projection of the blue circle to the $z = 0$ plane.

4.4 CONVERGENCE-PREVENTING REGULARIZATION

Using the above result, we construct a "regularization" function F for which the update[3]

$$\Delta CD_F(W) = \Delta CD(W) + \nabla F(W) \tag{11}$$

does not converge for the previously-described RBM, and causes the optimization to cycle indefinitely. It is not a function likely to be used in practice for regularization, but it demonstrates how the fact that the CD update is not a gradient can cause the optimization to behave in unexpected ways.

We achieve this effect by choosing a function F satisfying the following criteria:

1. The function F causes $\Delta CD_F(W)$ to seem like it is always "ascending" along the path γ.
2. The function F severely penalizes deviations from the path γ.

Condition 2 ensures that parameters following $\Delta CD_F(W)$ tend to stay near the path γ, while condition 1 causes the parameters cycle around it. As we will see, a function satisfying condition 1 can be constructed precisely because the CD update "thinks" that traversing γ causes a total increase of 0.008 in H.

[3]If ΔCD were the gradient of H and we wished to optimize $F + H$, then its gradient would be eq. 11.

We now define the function F more formally. Let $G(t)$ be the change, as computed by ΔCD, between the values of the presumed function H at $\gamma(0)$ and $\gamma(t)$. Specifically, $G(t)$ is defined by the integral

$$G(t) = \int_0^t \Delta CD(\gamma(\tau))^\top \gamma'(\tau) d\tau \tag{12}$$

The main fact about G is that $G(1) \approx 0.008 > 0$. Using G, we can define the values of the function F along the path γ with the equation

$$F(\gamma(t)) = -G(t) + G(1) \cdot t \tag{13}$$

For this definition to be consistent, the above equation must have $F(\gamma(0))$ equal to $F(\gamma(1))$ (simply because $\gamma(0) = \gamma(1)$). And indeed,

$$\begin{aligned} F(\gamma(0)) &= -G(0) + G(1) \cdot 0 \\ &= 0 \\ &= -G(1) + G(1) \cdot 1 = F(\gamma(1)) \end{aligned} \tag{14}$$

In fact, F is differentiable on each point of the circle γ including $\gamma(0) = \gamma(1)$ because

$$\begin{aligned} \frac{\partial F(\gamma(0))}{\partial t} &= -G'(0) + G(1) \\ &= -G'(1) + G(1) \\ &= \frac{\partial F(\gamma(1))}{\partial t} \end{aligned} \tag{15}$$

because $G'(0) = G'(1)$ by eq. 12. Finally, we extend F to $\mathbb{R}^2$: $F(x)$ is the value of F at the closest point to x on γ (and 0 at the circle's center), so F is differentiable everywhere except in the circle's center (which we set to zero).

As a result, the update $\Delta CD_F(W)$ cancels out the irregular effect of CD and preserves only the elevation $t \cdot G(1)$, which is can be seen by computing the magnitude of ΔCD_F's projection on the direction of γ:

$$\begin{aligned} \Delta CD_F(\gamma(t))^\top \gamma'(t) &= \Delta CD(\gamma(t))^\top \gamma'(t) + F'(\gamma(t))^\top \gamma'(t) \\ &= \Delta CD(\gamma(t))^\top \gamma'(t) - G'(t) + G(1) \\ &= G'(t) - G'(t) + G(1) \\ &= G(1) \\ &> 0 \end{aligned} \tag{16}$$

where we used the identity $\Delta CD(W)^\top \gamma'(t) = G'(t)$ from the definition of $G(t)$. This causes the parameters to move in a constant speed along the circle γ provided that the point stays near γ—which is easily arranged using regularization.

For the regularization function, we added $R(x) = -10 \cdot \|x - \gamma\|^{1.3}$ to F, where $\|x - \gamma\|$ is the distance from

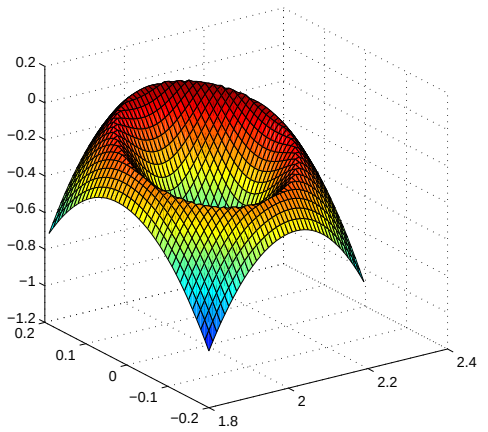

Figure 2: The surface of the "regularization" function F that causes CD learning to cycle. The plot clearly shows the contribution from G which is reflected in slope of the circle, as well as the contribution from the regularization.

x to the nearest point on γ. This regularization was chosen to make the simulation work successfully. For the simulation, we implemented F as described above by computing approximate values for $F(\gamma(t))$ at 20,000 equally-spaced points (using the rectangle method, as $F(\gamma(t))$ is an integral; eq. 13), and using cubic splines to extend $F(\gamma(t))$'s definition to all $0 \leq t \leq 1$. The gradient ∇F was computed with numerical differentiation (with a 10^{-7} stepsize); if the stepsize of CD is set to 10^{-5}, the update $\Delta CD_F(W)$ makes the parameters cycle around γ indefinitely.

The resulting F is not differentiable at 0, which we fix by multiplication with a smooth function whose value is mostly 1 but is zero in a small neighborhood of the circle's center. The resulting product will be unchanged in the region of interest, but the function will be globally differentiable. This is done for the sake of formality, because $\Delta CD_F(W)$ keeps the parameters near γ which is far from the circle's center, so F's values at its neighborhood are irrelevant.

5 THE FIXED POINTS OF CD

We previously demonstrated negative statements about the CD update, but now we show a more positive result. While we wish to show that CD converges, we show instead that the L_2-regularized CD has fixed points, where a fixed point is a setting of the parameters that is unchanged by the CD update. This is interesting because if fixed points did not exist, we could not even hope for CD to converge. The main result of this section states that the regularized CD update of fully visible Boltzmann Machines (BM) (Ackley et al., 1985) has fixed points. A fully visible BM with binary variables $V \in \{0,1\}^n$ defines a probability distribution by the expression

$$P(V) = \exp\left(V^\top W V/2\right)/Z \tag{17}$$

where W are its parameters, and its CD update is

$$\Delta CD(W) = E_{D(V)}\left[VV^\top\right] - E_{R(V)}\left[VV^\top\right] \tag{18}$$

where a sample from $R(V)$ is obtained by running a number of Gibbs sampling sweeps initialized at the distribution $D(V)$. Although our theorem is also valid for standard RBMs, there is a subtlety that makes the theorem less interesting in this case which we explain later.

Our simple technical result is fairly general, stating that any L_2-regularized continuous weight update $U : \mathbb{R}^n \to \mathbb{R}^n$ of bounded magnitude has a fixed point. It implies the above, because the CD update is continuous and bounded for binary (R)BMs. This results in the following theorem.

Theorem 1. *If $U : \mathbb{R}^n \to \mathbb{R}^n$ is a continous bounded function and $0 < \lambda < 1$, then the regularized update $U_\lambda(W) = U(W) - \lambda W$ has a fixed point: there is a setting of the parameters W^* so that $0 = U_\lambda(W^*) = U(W^*) - \lambda W^*$.*

When the distribution to be estimated, $D(V)$,[4] is exactly representable by a distribution of a BM with parameters W^* (say $P(V)$; so $D(V) = P(V)$), then CD is consistent in the sense that $\Delta CD(W^*) = 0$. This is because the Gibbs sampling Markov chain leaves $P(V)$ invariant, causing the distributions $D(V)$, $P(V)$, and $R(V)$ to be equal. However, no such result is known in the general case, when $D(V)$ cannot be precisely expressed as the distribution of an BM—which is the typical situation when working with finite training sets. Nonetheless, CD was suspected to have fixed points on the basis of careful empirical evidence (Carreira-Perpinan and Hinton, 2005).

The statement of the theorem is about the regularized CD update, which does not apply to the pure CD update ΔCD, which is a drawback. Nonetheless, there are two reasons for which such regularization is desirable. First, parameter estimation methods are rarely used without regularization, so the result is relevant to the way CD is used in practice. Second, unregularized, even the gradient of the log probability does not necessarily have a fixed point, and, in particular, does not

[4] If we have an RBM, then $D(V,H) = D(V)P(H|V)$. However, if we have a fully visible BM, then we only work with $D(V)$.

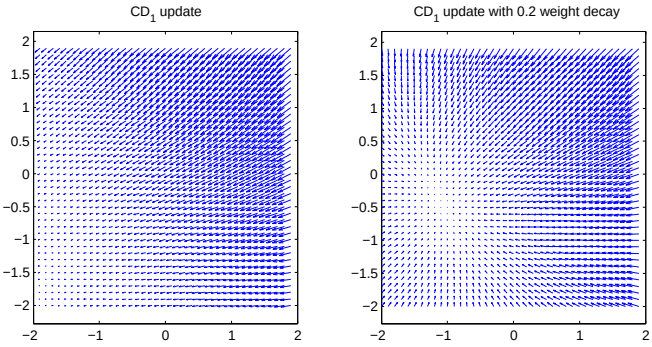

Figure 3: An illustration of Theorem 2 and of the effect of the L_2 regularization on the CD update. **Left**: the update defined by CD without regularization. The update has no fixed point in the plotted region (although it may have a fixed point outside of it). **Right**: the update defined by CD with 0.2 L_2 regularization. Observe that all the arrows on the boundary of the square point to its inside, and hence, according to Brower's fixed point theorem, it must have a fixed point (which is found near $(-1, -1)$). In both figures, the axes correspond to the parameters of the RBM described in section 4.

converge. To see why, consider the fully-visible binary Boltzmann machine with one unit:

$$\begin{aligned} P(V=1) &= \exp(\theta)/(\exp(\theta)+\exp(0)) \\ P(V=0) &= \exp(0)/(\exp(\theta)+\exp(0)) \end{aligned}$$

If the training data consists of the single point $\{1\}$, then the gradient of dataset's log probability,

$$\nabla_\theta \log P(V=1) = 1 - P(V=1),$$

is never zero, implying that regularization is necessary if we wish to obtain a statement of convergence or fixed points.

5.1 THE FORMAL PROOF

In this section we prove Theorem 1. The proof uses Brower's fixed point theorem (e.g., Henle, 1994):

Theorem 2. *Let B be any closed ball in $\mathbb{R}^k$, and let $f : B \to B$ be any continuous function whose outputs are contained in B. Then f has a fixed point—namely, there is an $x^* \in B$ such that $f(x^*) = x^*$.*

To prove Theorem 1, consider the function

$$f(W) = W - U_\lambda(W) = W - U(W) - \lambda W \qquad (19)$$

Let u be the upper bound on U, so $\|U(W)\| \le u$ for all W.

Now, for all W such that $\|W\| \le \frac{u}{\lambda}$,

$$\begin{aligned} \|W - \lambda W + U(W)\| &\le (1-\lambda)\|W\| + \|U(W)\| \\ &\le (1-\lambda)\frac{u}{\lambda} + u \\ &= \frac{u}{\lambda} \end{aligned} \qquad (20)$$

where we used the triangle inequality and the fact that $1 - \lambda > 0$. Therefore, when $R = u/\lambda$ and B is the closed ball $\{W : \|W\| \le R\}$, the function f satisfies $f(B) \subseteq B$; furthermore, f is continuous since both U update and the L_2 regularization are continuous. This lets us apply Brower's fixed point theorem to f and conclude that there is a $W^* \in B$ such that $f(W^*) = W^*$, or equivalently, that $U_\lambda(W^*) = 0$.

5.2 APPLICATION TO RESTRICTED BOLTZMANN MACHINES

We stated earlier that the above theorem is applicable to the fully visible BM as well as to the standard RBM, although we deemphasized the application to the standard RBM. This was done because the theorem is uninteresting for RBMs, as we can show that fixed points trivially exist. Namely, by setting the weights and the hidden biases to zero ($W = 0$ and $b_H = 0$ as in eq. 2), and fitting the visible biases b_V so that the marginal distributions $P(V_i)$ match the data marginal distributions $D(V_i)$, the unregularized CD can be seen to leave this parameter setting unchanged. The result can be seen to remain valid even if we introduce L_2 regularization and slightly modify the visible biases. This effect does not occur when CD is used with fully visible BMs, because they do not have trivial fixed points. Nonetheless, our theorem is valid when we regularize the parameters to a non-zero point, or, more ambitiously, blend the regularization gradient with the gradient of an autoencoder objective, both of which will prevent the zero weights from being stable.

6 DISCUSSIONS AND CONCLUSIONS

In this paper we gave proofs showing that CD is not the gradient of any function and that it is possible to construct regularization functions that cause it to fail to converge. We also showed that regularized CD has fixed points, which must be the case if CD really is convergent. However, the main task of proving that CD converges remains open.

7 ACKNOWLEDGEMENTS

We thank the anonymous referees for helpful comments.

References

Ackley, D., G. Hinton, and T. Sejnowski (1985). A learning algorithm for Boltzmann machines. *Cognitive science 9*(1), 147–169.

Bengio, Y. and O. Delalleau (2009). Justifying and generalizing contrastive divergence. *Neural Computation 21*(6), 1601–1621.

Bengio, Y., P. Lamblin, D. Popovici, H. Larochelle, and U. Montreal (2007). Greedy layer-wise training of deep networks. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, pp. 153. The MIT Press.

Besag, J. (1977). Efficiency of pseudolikelihood estimation for simple Gaussian fields.

Carreira-Perpinan, M. and G. Hinton (2005). On contrastive divergence learning. In *Artificial Intelligence and Statistics*, Volume 2005.

Cross, G. and A. Jain (1981). Markov random field texture models. In *Conference on Pattern Recognition and Image Processing, Dallas, TX*, pp. 597–602.

Henle, M. (1994). *A combinatorial introduction to topology*. Dover publications.

Hinton, G. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation 14*(8), 1771–1800.

Hinton, G., S. Osindero, and Y. Teh (2006). A fast learning algorithm for deep belief nets. *Neural Computation 18*(7), 1527–1554.

Hinton, G. and R. Salakhutdinov (2006). Reducing the dimensionality of data with neural networks.

Hyvarinen, A. (2006). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research 6*(1), 695.

Hyvarinen, A. (2007). Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables. *IEEE Transactions on Neural Networks 18*(5), 1529.

Li, S. (1994). Markov random field models in computer vision. *Lecture Notes in Computer Science 801*, 361–370.

Malfait, M. and D. Roose (1997). Wavelet-based image denoising using a Markov random field a priori model. *IEEE transactions on image processing 6*(4), 549–565.

Movellan, J. (2008). Contrastive divergence in gaussian diffusions. *Neural Computation 20*(9), 2238–2252.

Portilla, J., V. Strela, M. Wainwright, and E. Simoncelli (2003). Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Transactions on Image Processing 12*(11), 1338–1351.

Roth, S. and M. Black (2005). Fields of experts: A framework for learning image priors. In *IEEE Computer Society Conference on Computer Vision andn Pattern Recognition, 2005. CVPR 2005*, Volume 2.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. *Parallel distributed processing: Explorations in the microstructure of cognition 1*, 194–281.

Tieleman, T. (2007). Some investigations into energy-based models. Master's thesis, University of Toronto.

Wainwright, M. (2008). *Graphical Models, Exponential Families, and Variational Inference.* Now Publishers.

Yuille, A. (2004). The convergence of contrastive divergences. *Advances in neural information processing systems 17*.