ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

Maximum likelihood learning of RBMs with Gaussian visible units on the Stiefel manifold

Ryo Karakida[1], Masato Okada[1,2] and Shun-ichi Amari[2]

[1]The University of Tokyo - Department of Complexity Science and Engineering, Chiba - Japan

[2]RIKEN Brain Science Institute, Saitama - Japan

Abstract. The restricted Boltzmann machine (RBM) is a generative model widely used as an essential component of deep networks. However, it is hard to train RBMs by using maximum likelihood (ML) learning because many iterations of Gibbs sampling take too much computational time. In this study, we reveal that, if we consider RBMs with Gaussian visible units and constrain the weight matrix to the Stiefel manifold, we can easily compute analytical values of the likelihood and its gradients. The proposed algorithm on the Stiefel manifold achieves comparable performance to the standard learning algorithm.

1 Introduction

The restricted Boltzmann machine (RBM) is a bipartite graphical model that is widely used as a building block of deep neural networks [1], but it is hard to train by using maximum likelihood (ML) learning. Computation of the likelihood is analytically intractable and requires many iterations of Gibbs sampling, which entails a lengthy computation. Here, Contrastive divergence (CD) learning has been developed as an approximate way of computing the gradient of ML learning and is commonly used in practice [1, 2]. CD learning computes the ML gradient with samples obtained by a limited number of Gibbs samplings and empirically converges close enough to the ML solutions in a short time [3]. However, in general, there are almost no theoretical guarantees of convergence or maximization of the likelihood in CD learning [4].

In this study, we reveal that the likelihood and its gradients in RBMs with Gaussian visible units are analytically tractable when the weight matrix is constrained to the Stiefel manifold. We propose a novel algorithm based on geodesic flow on the Stiefel manifold for Gaussian-Bernoulli RBM and demonstrate its effectiveness in experiments on natural image patches. Because our algorithm has a tractable likelihood, there are advantages to using it to monitor the convergence and maximization of the likelihood. Moreover, we prove theoretically that the proposed method arrives at essentially the same solution as standard ML learning in Gaussian-Gaussian RBM.

2 Geodesic flow of ML learning on Stiefel Manifold

Let us consider the problem of minimizing a cost function L with regards to a parameter matrix $A \in \mathbb{R}^{M \times N}$. We assume $M \leq N$ and that M row vectors of

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

A are mutually orthogonal N-dimensional unit vectors satisfying $AA^T = I_M$. The set of all such matrices is known as the Stiefel manifold. In particular, the Stiefel manifold with $N = M$ reduces to the orthogonal group. When one minimizes L by using the steepest descent algorithm, the update rule is given by $A_{t+1} = A_t - \varepsilon \Delta A$, where ε is a small learning constant and $\Delta A = dL/dA$. It should be noted that, because A_{t+1} is not in the manifold, it is necessary to project A_{t+1} to the manifold in each iteration [5].

In contrast, by considering an extension of the natural gradient method, one can minimize L along geodesic flows on the Stiefel Manifold as follows [6]:

$$A_{t+1} \leftarrow A_t \exp\left(\epsilon(A_t^T \Delta A - \Delta A^T A_t)/2\right). \tag{1}$$

Because A_t being on the Stiefel manifold ensures that A_{t+1} will also be on it, we can omit the projection to the manifold in each iteration. In practical applications such as ICA, this geodesic algorithm has outperformed the standard steepest decent algorithms [5, 6]. In the following, we apply it to two types of RBM with Gaussian visible units.

2.1 Gaussian-Bernoulli RBM

The model distribution of a Gaussian-Bernoulli RBM is defined as follows [1]:

$$p(\mathbf{h}, \mathbf{v}) = \exp\left(-\frac{1}{2\sigma^2}|\mathbf{v} - \mathbf{b}|^2 + \frac{1}{\sigma}\mathbf{h}^T W \mathbf{v} + \mathbf{c}^T \mathbf{h}\right)/Z. \tag{2}$$

Let us denote binary hidden variables as $h_i = \{0, 1\}$ $(i = 1, ..., M)$ and continuous visible variables as v_i $(i = 1, ..., N)$. Moreover, we denote the variance of the visible units by σ^2 and the normalization constant by Z. We estimate the weight matrix $W \in \mathbb{R}^{M \times N}$ and bias vectors $\mathbf{b}$ and $\mathbf{c}$.

The maximum likelihood (ML) estimate is obtained by minimizing the negative log-likelihood $L = -\int q(\mathbf{v}) \ln p(\mathbf{v}) d\mathbf{v}$, where $q(\mathbf{v})$ denotes the input distribution and $p(\mathbf{v})$ denotes the marginal model distribution. The steepest gradient of the negative log-likelihood is given by $W_{t+1} = W_t + \epsilon \Delta W$ with $\Delta W =< \mathbf{h}\mathbf{v}^T >_q - < \mathbf{h}\mathbf{v}^T >_p$ [1]. Let us denote the average over training examples generated from $q(\mathbf{v})$ as $< \cdot >_q$ and the average over the model distribution as $< \mathbf{h}\mathbf{v}^T >_p$. After marginalizing $\mathbf{h}$ in the first term and $\mathbf{v}$ in the second term, the update ΔW can be transformed into

$$\Delta W =< g(W\mathbf{v}/\sigma + \mathbf{c})\mathbf{v}^T >_q - (\sigma < \mathbf{h}\mathbf{h}^T >_{p(\mathbf{h})} W + < \mathbf{h} >_{p(\mathbf{h})} \mathbf{b}^T), \tag{3}$$

where $g(\mathbf{x})$, a function with a vector argument $\mathbf{x}$, denotes a vector whose i-th element is a sigmoid function $g(x_i)$. The second term is analytically intractable because one needs to take a summation over an exponential number of hidden states obeying the following model distribution:

$$p(\mathbf{h}) = \exp\left(|W^T \mathbf{h}|^2/2 + (W\mathbf{b}/\sigma + \mathbf{c})^T \mathbf{h}\right)/Z. \tag{4}$$

Surprisingly, we can avoid this analytically intractable summation by constraining the parameter space of W. Let us assume that the weight matrix W

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

is constrained to $W = DA$, where A is included in the Stiefel Manifold and D is a diagonal matrix $D = \mathrm{diag}(d_1, d_2, ..., d_M)$. Substituting $W = DA$ into (4), the model distribution $p(\mathbf{h})$ becomes independent among the hidden variables,

$$p(\mathbf{h}) = \prod_{i=1}^{M} g(y_i)^{h_i}(1 - g(y_i))^{1-h_i}, \tag{5}$$

where we define $y_i = d_i^2/2 + d_i \mathbf{a}_i \mathbf{b}/\sigma + c_i$ and denote the i-th row vector of A by $\mathbf{a}_i$. The constraint of $W = DA$ corresponds not only to reducing the dimension of the search space W but also to giving a prior that hidden units act independently.

Using the independent distribution (5), we can compute the update rule (3) analytically as follows:

$$\Delta A = D[< g(W\mathbf{v}/\sigma + \mathbf{c})\mathbf{v}^T >_q - (\sigma K W + g(\mathbf{y})\mathbf{b}^T)], \tag{6}$$

where $K_{ii} = g(y_i) \quad (i = 1, ..., M)$ and $K_{ij} = g(y_i)g(y_j) \quad (i \neq j)$. The update rule of A is given by (1) with (6). In addition, the update rules of the other parameters are given by the ordinary steepest directions

$$d_i \leftarrow d_i + \epsilon[< g(d_i \mathbf{a}_i \mathbf{v}/\sigma + c_i)\mathbf{a}_i \mathbf{v} >_q - g(y_i)(\mathbf{a}_i \mathbf{b} + \sigma d_i)], \tag{7}$$

$$\mathbf{b} \leftarrow \mathbf{b} + \epsilon[< \mathbf{v} >_q - (\mathbf{b} + \sigma W^T g(\mathbf{y}))], \tag{8}$$

$$\mathbf{c} \leftarrow \mathbf{c} + \epsilon[< g(W\mathbf{v}/\sigma + \mathbf{c}) >_q - g(\mathbf{y})]. \tag{9}$$

Moreover, the constraint $W = DA$ makes the negative log-likelihood analytically tractable as follows:

$$L = < \frac{1}{2\sigma^2}|\mathbf{v} - \mathbf{b}|^2 - \Sigma_i \ln(1 + e^{d_i \mathbf{a}_i \mathbf{v}/\sigma + c_i}) >_q + \Sigma_i \ln(1 + e^{y_i}) + N \ln(\sqrt{2\pi}\sigma). \tag{10}$$

In general, L is analytically intractable and difficult to use to monitor the progress of learning in the Gaussian-Bernoulli RBM. In contrast, our method can monitor how well and determine where the learning trajectory converges.

2.2 Gaussian-Gaussian RBM

We can also obtain the geodesic ML learning rule in Gaussian-Gaussian RBM, whose model distribution is defined as follows [7]:

$$p(\mathbf{h}, \mathbf{v}) = \exp\left(-\frac{1}{2s^2}|\mathbf{h} - \mathbf{c}|^2 - \frac{1}{2\sigma^2}|\mathbf{v} - \mathbf{b}|^2 + \frac{1}{s\sigma}\mathbf{h}^T W \mathbf{v}\right)/Z, \tag{11}$$

where both visible and hidden units take continuous real values. Note that the likelihood and its gradient in the standard ML learning without any constraint are analytically tractable in Gaussian-Gaussian RBM [7]. In this study, we consider ML learning with $W = DA$ in order to obtain an insight into how this constraint changes the solution compared with that of standard ML learning. The geodesic ML learning on the Stiefel manifold is given by

$$A \leftarrow A \exp\left(\epsilon(A^T D^2 A C - C A^T D^2 A)/2\right), \tag{12}$$

$$d_i \leftarrow d_i + \epsilon d_i[< (\mathbf{a}_i \mathbf{v})^2 >_q - \sigma^2/(1 - d_i^2)], \tag{13}$$

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

where C is the data covariance matrix of the input distribution $q(\mathbf{v})$. We set the mean values of the input data to $\int \mathbf{v}q(\mathbf{v})d\mathbf{v} = 0$ and set the bias parameters to $\mathbf{b} = \mathbf{c} = 0$ for simplicity, but we can also formulate and analyze the general case in the same way.

3 Analysis of Gaussian-Gaussian RBM

Here, we provide a theoretical guarantee that ML learning on the Stiefel manifold arrives at essentially the same ML solution as standard ML learning in Gaussian-Gaussian RBM. Due to space limitations, we will consider only the case of $M = N$ in the following analysis.

Let us assume that the data covariance matrix C has non-degenerate eigenvalues λ_i $(i = 1, ..., N)$ satisfying $\lambda_1 > \cdots > \lambda_k > \sigma^2 > \lambda_{k+1} > \cdots > \lambda_N$ and is diagonalized such that $C = V^T \mathrm{diag}(\lambda_1, ..., \lambda_N)V$, where V is an $N \times N$ orthogonal matrix. Under these assumptions, the previous study found that the stable solution of standard ML learning is limited to the following $\bar{W}$ [7]:

$$\bar{W} = U \mathrm{diag}\left(\sqrt{1 - \sigma^2/\lambda_1}, ..., \sqrt{1 - \sigma^2/\lambda_k}, 0, ..., 0\right) V, \tag{14}$$

where U is an arbitrary $N \times N$ orthogonal matrix. At the stable solution $\bar{W}$, the model distribution becomes a Gaussian distribution: $p(\mathbf{v}) = \mathcal{N}(\mathbf{v}; \mathbf{0}, V^T \mathrm{diag}(\lambda_1, ..., \lambda_k, \sigma^2, ..., \sigma^2)V)$. This means that the trained Gaussian-Gaussian RBM extracts only the largest k principal components, whose eigenvalues are larger than the model variance σ^2.

Under the assumptions of the previous study, we find that the ML learning on the Stiefel manifold (12, 13) has the following analytical solutions:

Proposition. *The stable equilibrium solution of ML learning on the Stiefel manifold (12, 13) is* $\bar{W} = \mathrm{diag}\left(\sqrt{1 - \sigma^2/\lambda_1}, ..., \sqrt{1 - \sigma^2/\lambda_k}, 0, ..., 0\right) V$.

Proof. The update rule (12) stops at $A^T D^2 AC = CA^T D^2 A$. Because $A^T D^2 A$ and C are commutative, these matrices are simultaneously diagonalizable, and we obtain an equilibrium $\bar{A} = V$. In addition, substituting $\bar{A} = V$ into (13), we get $\bar{d}_i = 0$ or $\bar{d}_i = \sqrt{1 - \sigma^2/\lambda_i}$ with $\lambda_i > \sigma^2$.

Next, we check the stability of the equilibrium solution in a similar process as shown in the standard ML learning [7]. We can represent the perturbation of A along the Stiefel manifold as $\Delta A \equiv A \exp(A^T \Delta X A) - A$, where ΔX is an $N \times N$ alternative matrix whose entries satisfy $\Delta X_{ij} = -\Delta X_{ji}$. Because the perturbation ΔX_{ij} takes an infinitesimal value $|\Delta X_{ij}| \ll 1$, we get $\Delta A \sim \Delta X A$. The stable solution requires the following inner product to become negative,

$$\begin{aligned} &\mathrm{Tr}\left(\Delta A^T \Delta F_A\right) + \Sigma_i \Delta d_i \Delta F_{d_i} \\ &\sim \sum_{a<b}^{N} \Delta X_{ab}^2 (d_a^2 - d_b^2)(\lambda_b - \lambda_a) + \sum_{i=1}^{N} \Delta d_i^2 \left\{\lambda_i - (1 + d_i^2)/(1 - d_i^2)^2 \sigma^2\right\}, \end{aligned} \tag{15}$$

where Δd_i denotes the perturbation of d_i. When one transforms the update rule (12) into the form $A_{t+1} - A_t = F_A(A_t, d_{i,t})$ and the update rule (13) into

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

$d_{i,t+1} - d_{i,t} = F_{d_i}(A_t, d_{i,t})$, the perturbations of the gradients, ΔF_A and ΔF_{d_i}, are given by $\Delta F_A = F_A(\bar{A} + \Delta A, \bar{d}_i + \Delta d_i)$ and $\Delta F_{d_i} = F_{d_i}(\bar{A} + \Delta A, \bar{d}_i + \Delta d_i)$. We can easily confirm that the inner product (15) becomes negative if and only if $\bar{D} = \mathrm{diag}\left(\sqrt{1-\sigma^2/\lambda_1}, ..., \sqrt{1-\sigma^2/\lambda_k}, 0, ..., 0\right)$. Therefore, the ML learning on the Stiefel manifold has the global minimum $\bar{W} = \bar{D}\bar{A}$. □

We have thus proven the remarkable fact that constraining the weight space to $W = DA$ corresponds to eliminating the rotational degrees of freedom caused by U in the standard ML stable solution. In addition, the model distribution $p(\mathbf{v})$ coincides with that of standard ML learning. Therefore, ML learning on the Stiefel Manifold gives essentially the same solution as standard ML learning.

In the case of $M < N$, we can also analytically obtain the ML solutions. Although some local minima appear in the ML learning with $W = DA$, the global minima coincide with those of the standard ML solutions, as is shown with $M = N$.

4 Experiments on Gaussian-Bernoulli RBM

To confirm the effectiveness of our method, we trained a Gaussian-Bernoulli RBM with the algorithm (6-9) on natural image patches sampled from the van Hateren natural image database [8]. In the preprocessing, we applied global contrast normalization and ZCA whitening to 50,000 image patches of 14x14 pixels. The data set consisted of 40,000 training cases and 10,000 test cases. We set σ^2 to be on the same scale as the variance of the data.

As is shown in Fig. 1A, our method achieved a test error comparable to that of the persistent contrastive divergence (CD) algorithm [2] in the Gaussian-Bernoulli RBM with $M = 16$ and $N = 196$. We computed the test error by the negative log-likelihood on the test cases. The thick line in Fig. 1A represent the mean over 10 training runs with different random initializations and the dotted lines represent the standard deviation. The CD algorithm with no constraint on W has an intractable likelihood, but we set the number of hidden units to

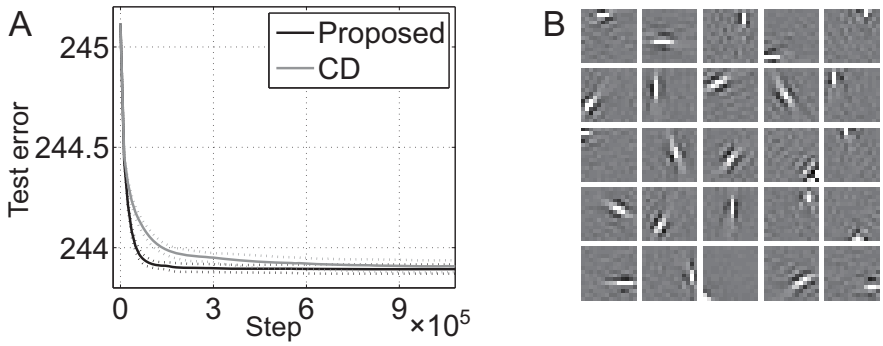

Fig. 1: Negative log-likelihood of test data on natural image dataset: (A) Comparison of proposed method and CD algorithm ($M = 16$, $N = 196$); (B) Gabor-like filters obtained by proposed method ($M = N = 196$)

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

a small value and computed the exact value of the likelihood. Our method was constrained to $W = DA$ and had fewer free parameters than the CD algorithm. Nevertheless, it achieved the same test error as the CD algorithm.

We also trained a Gaussian-Bernoulli RBM with $N = M = 196$ with our method. Because we can compute the likelihood function by using (10) even in the case of large M, we can easily monitor the convergence and test error. We randomly selected the learned filters and the reshaped rows of W with $d_i \neq 0$, and show them in Fig. 1B. As Gabor-like filters are extracted, our method would seem to be useful in feature extraction.

5 Conclusion and Future work

We proposed a novel algorithm to train RBMs with continuous visible units, where the constraint on the Stiefel manifold enables us to compute analytical values of the likelihood and its gradients. In the experiments on Gaussian-Bernoulli RBM, our method achieved comparable performance with the CD algorithm. For Gaussian-Gaussian RBM, we provided a theoretical guarantee that the proposed method obtains the essentially same solution as that of standard ML learning.

A further direction of study is to apply a similar constraint on the parameter space to more general forms of RBMs, such as exponential family harmoniums and stacked RBMs. In deep networks, it has been suggested that orthogonal weight matrices obtained by layerwise pre-training accelerate the convergence of supervised learning after pre-training [9]. It remains to be explored how to apply our method to pre-training of deep networks.

References

[1] G. E. Hinton and R. Salakhutdinov, Reducing the dimensionality of data with neural networks, *Science*, 313(5786):504-507, 2006.

[2] T. Tieleman and G. E. Hinton, Using fast weights to improve persistent contrastive divergence. In *Proceedings of the* 26^{th} *International Conference on Machine Learning* (ICML), pp. 1033-1040, 2009.

[3] M. A. Carreira-Perpinan, and G. E. Hinton, On contrastive divergence learning, In *Proceedings of the* 10^{th} *International Conference on Artificial Intelligence and Statistics* (AISTATS), pp. 33-40, 2005.

[4] I. Sutskever and T. Tieleman, On the convergence properties of contrastive divergence, In *Proceedings of the* 13^{th} *International Conference on Artificial Intelligence and Statistics* (AISTATS), pp. 789-795, 2010.

[5] Y. Nishimori and S. Akaho, Learning algorithms utilizing quasi-geodesic flows on the Stiefel manifold, *Neurocomputing*, 67:106-135, 2005.

[6] S. Fiori, Quasi-Geodesic Neural Learning Algorithms Over the Orthogonal Group: A Tutorial, *JMLR*, 6:743-781, 2005.

[7] R. Karakida, M. Okada and S. Amari, Analyzing Feature Extraction by Contrastive Divergence Learning in RBMs, *Deep Learning and Representation Learning Workshop: NIPS 2014*, Montreal (Canada), December 2014.

[8] H, Lee, C. Ekanadham, and A. Ng, Sparse deep belief net model for visual area V2, In *Proceedings of Advances in Neural Information Processing Systems 20* (NIPS), 2007.

[9] A. Saxe, J. L. McClelland and S. Ganguli, Exact solutions to the nonlinear dynamics of learning in deep linear neural networks, *arXiv preprint*, 1312.6120, 2013.