Information Geometry of Contrastive Divergence

Shotaro Akaho
AIST
Tsukuba, 305-8568 Japan

Kazuya Takabatake
AIST
Tsukuba, 305-8568 Japan

Abstract ***The contrastive divergence(CD) method proposed by Hinton finds an approximate solution of the maximum likelihood of complex probability models. It is known empirically that the CD method gives a high-quality estimation in a small computation time. In this paper, we give an intuitive explanation about the reason why the CD method can approximate well by using the information geometry. We further propose an improved method that is consistent with the maximum likelihood (or MAP) estimation, while the CD method is biased in general.***

Keywords: Information geometry, Markov chain Monte Carlo, Gradient descent, graphical model

1 Introduction

In recent years, probabilistic models called graphical models have been taken a large role in various applications from wireless communication to datamining[6, 10]. Although graphical models are very flexible, we often encounter computational difficulties and many efficient approximation methods have been proposed.

In this paper, we consider a parameter estimation problem of complex probabilistic models like graphical models from sample data. The biggest bottleneck in the computation is to calculate a normalization parameter. Recently, Hinton[8] has proposed an efficient approximation method called "Contrastive Divergence (CD) method".

Empirically it is known that the CD method gives a good solution, and some theoretical results on the behavior of the CD method have been reported[15, 16, 4].

This paper provides two main contributions. One is to give an intuitive explanation why the CD method can approximate well, based on a discussion from an information geometrical point of view. The other is to propose an improved method that gives a solution that is consistent with the maximum likelihood (or MAP) estimation, while the CD method gives a biased solution in general.

We consider the Boltzmann distribution of a random vector $X \in \mathcal{X}$ with a parameter w,

$$p(X;w) = \frac{1}{Z(w)} \exp(-l(X;w)) \tag{1}$$

where $l(X;w)$ is called an energy function, and $Z(w)$ is a normalization constant

$$Z(w) = \sum_{X \in \mathcal{X}} \exp(-l(X;w)). \tag{2}$$

For just a notational convention, we assume that $\mathcal{X}$ is finite.

Here we consider the problem of finding the maximum likelihood (ML) estimation from finite number of samples of X (extension from ML to MAP is easy and straightforward just by adding the term corresponding to the prior distribution). The calculation of $Z(w)$ is often computationally intensive.

First, let us formulate the ML estimation problem. The average of log-likelihood over given samples $\mathcal{D} = \{X(1), \ldots, X(n)\}$ is written as

$$\begin{aligned} L(w) &= \frac{1}{n} \sum_{i=1}^{n} \log p(X(i);w) \\ &= \langle \log p(X(i);w) \rangle_{p_0} \\ &= -\langle l(X;w) \rangle_{p_0} - \log Z(w), \end{aligned} \tag{3}$$

where $\langle f(X) \rangle_p$ denotes the average of $f(X)$ over the distribution $p(X)$, and p_0 is the empirical distribution

$$p_0(X) = \frac{1}{n} \sum_{i=1}^{n} \delta(X - X(i)). \tag{4}$$

A general approach to maximize $L(w)$ is a gradient algorithm, in which the parameter is updated by the gradient,

$$\begin{aligned} w_{t+1} - w_t &= \gamma_t \partial_w L(w) \\ &= -\gamma_t \left\{ \langle \partial_w l(X;w) \rangle_{p_0} - \langle \partial_w l(X;w) \rangle_{p_w} \right\}, \end{aligned} \tag{5}$$

where $\gamma_t > 0$ is a learning coefficient that may depend on t in general, ∂_w denotes the derivative by w, and

$$p_w(X) = p(X; w) \tag{6}$$

is the model distribution.

We encounter a difficulty in computing the second term of (5) that depends on $p(X; w)$, because it is still necessary to calculate $Z(w)$.

2 Contrastive divergence

One approach to avoid the direct calculation of average with respect to $p(X; w)$ is Markov chain Monte Carlo (MCMC)[7], in which the random samples of X is generated by a Markov chain $q(X^{(k+1)} \mid X^{(k)})$, so that the distribution of $X^{(k)}$ converges to $p(X; w)$ as $k \to \infty$. By taking an appropriate Markov chain $q(X^{(k+1)} \mid X^{(k)})$, the stationary distribution is guaranteed to converge to $p(X; w)$.

However, introducing MCMC is still computationally intensive, because MCMC is an inner loop of each gradient descent step. One idea to resolve this problem is to use a distribution $r_w(X)$ which approximates $p(X; w)$. The update rule is thus given by

$$w_{t+1} = w_t - \gamma_t \left\{ \langle \partial_w l(X; w) \rangle_{p_0} - \langle \partial_w l(X; w) \rangle_{r_w} \right\}. \tag{7}$$

Hinton[8] has proposed the contrastive divergence (CD) method based on such an idea. CD algorithm applies just a few steps of the MCMC iteration in each gradient step, where the empirical distribution $p_0(X)$ is taken as an initial state of the MCMC. The algorithm is summarized as follows:

[CD algorithm]

1. Initialize the parameter w_0.
2. Gradient descent loop: repeat the following MCMC loop for $t = 0, 1, 2, 3, \ldots$ until convergence
   (a) MCMC loop by population: for each i, let the initial state $X^{(0)}(i)$ be each sample $X(i)$ $(i = 1, 2, \ldots, n)$, then perform MCMC for K steps to get the MCMC sample $X^{(K)}(i)$.
   (b) Update w_t by (7) in which the $r_w(X)$ be the empirical distribution after K-step MCMC

$$r_w(X) = p^{(K)}(X) \equiv \frac{1}{n} \sum_{i=1}^{n} \delta(X - X^{(K)}(i)) \tag{8}$$

It is clear that the CD method is equivalent to the Monte Carlo version of the ML (maximum likelihood) gradient descent as the number of MCMC steps K goes to infinity, because the distribution of $X^{(K)}(i)$ converges to $p(X; w)$. It is known that the CD method gives a good solution, even when K is relatively small. We will give an intuitive interpretation about the reason why CD method can approximate well by means of information geometry.

3 Gibbs sampler

3.1 Single component MCMC

Although there are many kinds of Markov chain Monte Carlo algorithms, we focus on a specific class of MCMC called single component MCMC which updates only one component at each step.

Suppose X is an m-dimensional vector $X = (X_1, X_2, \ldots, X_m) \in \mathcal{X}^m$, single component MCMC updates one component X_j at each step,

$$q(X^{(k+1)}|X^{(k)}) = q(X_j^{(k+1)}|X^{(k)})\delta(X_{-j}^{(k+1)} - X_{-j}^{(k)}) \tag{9}$$

where X_{-j} denotes the $m-1$ dimensional vector excluding the component X_j. The transition probability $q(X^{(k+1)} \mid X^{(k)})$ has to satisfy the condition that the distribution of $X^{(k)}$ approaches the target distribution $\pi(X) = p(X; w)$ as k increases. One well-known sufficient condition for the convergence is so-called "detailed balance condition",

$$q(X^{(k)}|X^{(k+1)})\pi(X^{(k+1)}) = q(X^{(k+1)}|X^{(k)})\pi(X^{(k)}). \tag{10}$$

Summing up by $X_j^{(k)}$, we obtain the stationarity condition,

$$\pi(X^{(k+1)}) = \sum_{X^{(k)} \in \mathcal{X}} q(X^{(k+1)} \mid X^{(k)})\pi(X^{(k)}). \tag{11}$$

In this paper, we consider only the transition matrix that satisfies the detailed balance condition. Here, we introduce a simple notational convention. First, let $p_t^{(k)}$ be $|\mathcal{X}|^m$ dimensional row vector representing the state probability at the k-th MCMC step of the t-th gradient descent step. The component corresponding to the state of X is denoted by $p_t^{(k)}(X)$. Next, let $Q_t^{(k)}$ be $|\mathcal{X}|^m \times |\mathcal{X}|^m$ matrix representing the state transition matrix whose stationary distribution is $p(X; w_t)$. The transition probability from the state X and Y is given by the (X, Y) component of $Q_t^{(k)}$.

By using this vector-matrix notation, the state transition from $p_t^{(k)}$ by $Q_t^{(k)}$ is given by

$$p_t^{(k+1)} = p_t^{(k)} Q_t^{(k)} = \cdots = p_t^{(0)} \prod_{j=1}^{k} Q_t^{(j)} = p_t^{(0)} Q_t^{[k]}, \tag{12}$$

where

$$Q_t^{[k]} = \prod_{j=1}^{k} Q_t^{(j)}. \tag{13}$$

The condition that single component MCMC satisfies is given by

$$p_t^{(k+1)}(X_{-i}) = p_t^{(k)}(X_{-i}) \tag{14}$$

for any X and i, where $p_t^{(k)}(X_{-i})$ denotes the $|\mathcal{X}|^{m-1}$ vector corresponding to the marginal probability of X_{-i}. Note that the above equation is a linear constraint. Further, the stationary condition (11) can be written in a simple form, $\pi = \pi Q_t^{(k)}$.

3.2 Gibbs sampler

An important class of the single component MCMC is Gibbs sampler, in which the state transition probability is given by the conditional probability of the target distribution,

$$q(X_j^{(k+1)} \mid X^{(k)}) = \pi(X_j^{(k+1)} \mid X_{-j}^{(k)}). \tag{15}$$

The Gibbs sampler has several particular properties. The first one is that the transition probability for X_j does not depend on the current X_j but only depends on the other elements. Another property is that this transition probability satisfies the detailed balance condition. In MCMC, one often require "rejection step" in order to satisfy the detailed balance condition. However, the Gibbs sampler does not require the rejection step.

4 Information Geometry of Contrastive Divergence

4.1 Information geometry of single component MCMC

Information geometry is a natural differential geometry for a space of probability distributions[2].A lot of information processing algorithms have been explained by means of information geometry[1, 3, 5, 9, 11], which gives intuitive understanding of the algorithms and also gives a mathematical framework for analysis. We just outline some necessary concepts of the information geometry.

Once the probability model is designed, each probability distribution can be specified by its parameter. Therefore, the probability distribution can be regarded as a point in a space spanned by the parameter. We can introduce a natural Riemannian structure in that space, where Riemannian metric is given by Fisher information matrix, and the natural affine connection is given by α-connection that is specified by one real valued parameter α.

When we consider the exponential family model,

$$p(X;\theta) = \exp(\theta \cdot F(X) - C(X) - \psi(\theta)) \tag{16}$$

α-connection is significant in particular for $\alpha = \pm 1$, because "flat" structure appears in the space.

For $\alpha = 1$, the natural coordinate θ is the affine coordinate (e-coordinate) in the sense that any geodesic is represented by a linear line, $t\theta_1 + (1-t)\theta_2$, which is called e-geodesic ('e' stands for exponential) and S is called e-flat with respect to this connection. On the other hand, the case of $\alpha = -1$ is dually related to the case of $\alpha = 1$, where the dual coordinate η (m-coordinate) can be obtained by Legendre transform from the e-coordinate, which is given by $\eta = E_X[F(X)]$ for the exponential family, and geodesics are given by linear lines of η called m-geodesics ('m' stands for mixture), and hence S is also called m-flat. As a consequence, the exponential family is dually flat for $\alpha = \pm 1$ connections.

One important concept in the statistical inference is a subspace (submanifold) and a projection onto the subspace. Because there are two kinds of affine coordinate systems, we also define two kinds of flat (autoparallel) subspaces, each of which is a linear subspace with respect to the affine coordinate. The projection can be also defined in two ways: e-projection is the e-geodesic that is orthogonal with respect to the Riemannian metric at the projection point, and m-projection is defined in a similar way.

An important fact about the (flat) subspace and projection is given by the following theorem illustrated in fig.1.

Theorem 1 (projection theorem[2]) *Suppose M be a submanifold of the space of exponential family S, then e-projection from $p(X) \in S$ to M is given by a stationary point of m-divergence that is defined by Kullback-Leibler (KL) divergence*

$$K(q(X), p(X)) = \sum_{X \in \mathcal{X}} q(X) \log \frac{q(X)}{p(X)}. \tag{17}$$

In particular, if M is m-flat, the e-projection is unique and it minimizes the m-divergence. The same

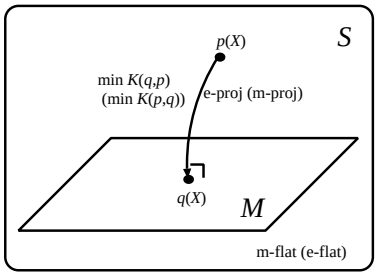

Figure 1: The projection theorem: for a flat submanifold, dual projection is unique that minimizes the KL divergence.

statement holds if we exchange m- and e- completely, where e-divergence is defined by $K(p(X), q(X))$.

4.2 Optimality of the Gibbs sampler

Let us go back to our problem. A set of all probability distribution over $\mathcal{X}^m$ can be regarded as an exponential family S. The parameter $p_t^{(k)}$ forms an m-coordinate system for S, where only $|\mathcal{X}|^m - 1$ elements are independent, because the sum of the components of $p_t^{(k)}$ should be 1. Single component MCMC satisfies (14), which is a linear constraint for $p_t^{(k+1)}$, which yields the following lemma.

Lemma 1 (Flatness[13]) *If we fix the distribution* $p_t^{(k)}$ *obtained by k steps of single component MCMC, the set of possible distributions* $\{p_t^{(k+1)} = p_t^{(k)} Q_t^{(k)}\}$ *forms an m-flat submanifold* M_k*, where* $Q_t^{(k)}$ *is any possible transition matrix of single component MCMC for the target distribution* $p(X; w_t)$.

From the projection theorem, the e-projection onto an m-flat submanifold is unique, and Takabatake[13] has shown the following theorem that guarantees the Gibbs sampler to be the e-projection from the target distribution onto M_k (fig.2).

Theorem 2 (Optimality of Gibbs sampler[13]) *Let* $p_t^{(k)}$ *be the current state vector, and* $Q_t^{(k)}$ *be the transition matrix of the Gibbs sampler, then the e-projection* $\tilde{p}_t^{(k+1)}$ *of the target distribution* $\pi(X)$ *onto* M_k *is given by* $\tilde{p}_t^{(k+1)} = p_t^{(k)} Q_t^{(k)}$.

This theorem means that the Gibbs sampler can be regarded as an alternating optimization procedure in which the optimization is performed within a subspace defined by (14).

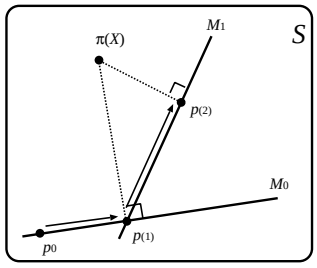

Figure 2: Optimality of the Gibbs sampler

proof (outline) The e-projection of $\pi(X)$ is given by

$$\arg\min_{p_t^{(k+1)}} K(p_t^{(k+1)}, \pi) \tag{18}$$

which is equivalent to

$$\min_{q(X_i|X_{-i})} K\left(p_t^{(k)}(X_{-i}), \pi(X_{-i})\right) + \left\langle K\left(q(X_i \mid X_{-i}), \pi(X_i \mid X_{-i})\right)\right\rangle_{p_t^{(k)}(X_{-i})} \tag{19}$$

for the single component MCMC, where $q(X_i \mid X_{-i})$ is the transition probability of single component MCMC from $p_t^{(k)}$ to $p_t^{(k+1)}$. The first term does not depend on the distribution of X_i, and the second term takes the minimum value 0 when $q(X_i \mid X_{-i}) = \pi(X_i \mid X_{-i})$ which is equivalent of the update rule of the Gibbs sampler.

4.3 Information geometry of CD method

Now we are ready to give a geometrical interpretation of the CD method. For simplicity, we assume the distribution $p(X)$ generating samples belongs to the exponential family (16), and the model space M is e-flat submanifold of S. The latter assumption is not essential and it can be easily generalized to more general case.

Here we need one more convenient notion of information geometry, that is, a "mixed coordinate" (fig. 3). In the exponential family, the way of taking natural parameter θ has an ambiguity because any regular linear transformation $\theta' = A\theta$ can give the same distribution by taking $F'(X) = A^{-1}F(X)$ and changing the normalization term $\psi(\theta)$ appropriately. Therefore, there exists a natural coordinate θ such that M is represented by $(\theta_{\rm I}, 0)_{\rm e} = (w, 0)_{\rm e}$, where $()_{\rm e}$ denotes the e-coordinate representation, and θ is divided into two parts like $(\theta_{\rm I}, \theta_{\rm II})_{\rm e}$. Mixed coordinate is defined by $(\eta_{\rm I}; \theta_{\rm II})_{\rm mix}$, when e-coordinate and

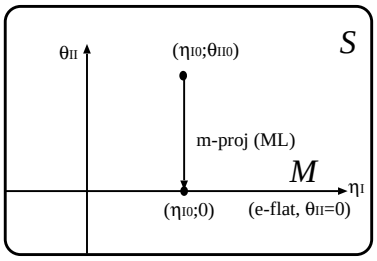

Figure 3: Mixed coordinate

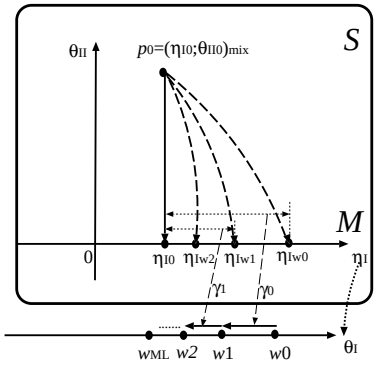

Figure 4: Geometry of Maximum likelihood

m-coordinate are given by $(\theta_{\rm I}, \theta_{\rm II})_{\rm e}$ and $(\eta_{\rm I}, \eta_{\rm II})_{\rm m}$ respectively. Mixed coordinate can specify any points of S uniquely.

Let us consider the geometrical interpretation of the ML and CD method. First, the empirical distribution $p_0(X)$ can be represented as a mixed coordinate $(\eta_{\rm I0}; \theta_{\rm II0})_{\rm mix}$.

It is known that the ML estimation is equivalent to the m-projection from the empirical distribution to the model submanifold, which is given in a closed form of the mixed coordinate $(\eta_{\rm I0}; 0)_{\rm mix}$ (fig. 4). Note that this does not mean the ML solution is given in a closed form, because it is not easy to obtain e-coordinate representation from the mixed coordinate in general.

The average equation of the general gradient descent optimization (7) is given by

$$\langle w_{t+1} - w_t \rangle = -\gamma_t(\eta_{\rm I0} - \eta_{\rm I}(r_w)), \qquad (20)$$

where $\eta_{\rm I}(r)$ is $\eta_{\rm I}$ coordinate specified by $r_w(X)$. In the stationary point $w = w^*$ of this equation, the solution satisfies the equation $\eta_{\rm I0} = \eta_{\rm I}(r_{w^*})$. Therefore, for the ML solution, it holds

$$\eta_{\rm I0} = \eta_{\rm I}(p_{w^*}), \quad p_{w^*} = p(X; w^*), \qquad (21)$$

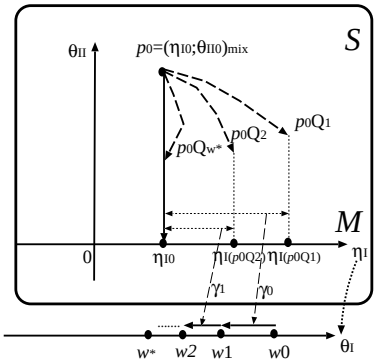

Figure 5: Geometry of Contrastive Divergence

and on the other hand, the CD method converges to the point

$$\eta_{\rm I0} = \eta_{\rm I}(p_0 Q_{w^*}^{[K]}), \qquad (22)$$

where $Q_{w^*}^{[K]}$ denotes the transition matrix of K steps of MCMC for the target distribution $p(X; w^*)$.

From the above consideration, we get a simple expression of the condition of CD and ML being equivalent.

Theorem 3 *If the model manifold is e-flat, both the stationary points of CD and ML are equal to w^* when it holds*

$$\eta_{\rm I}(p_{w^*}) = \eta_{\rm I}(p_0 Q_{w^*}^{[K]}). \qquad (23)$$

This is equivalent to the condition given by Yuille[16] but it is much simpler, even though we assumed the model is e-flat.

Let us consider qualitatively about the relation between ML and CD. In order that CD may approximates ML well, it is not necessary that $p_0 Q_t^{[K]}$ is close to the model manifold M, but it is sufficient if the projection of $p_0 Q_t^{[K]}$ to M approximates $\eta_{\rm I}(p_w)$ well. In particular, if we use the Gibbs sampler, the direction of $p_0 Q_t^{[K]}$ may not be so different from the direction of $\eta_{\rm I}(p_w)$ because of the optimality of the Gibbs sampler. It is one reason why the solution of CD well approximates well.

There is another reason related to the Monte Carlo nature of the algorithm. Although the analysis of this section has been based on the average equation, the algorithm includes fluctuations due to the finiteness of the number of Monte Carlo sampling in practice. Therefore, even for the ML (MCMC with infinite steps), the solution is suffered by this fluctuation. If it is larger than the approximation precision by the CD method, the CD method will be a good approximator of the ML.

4.4 Stochastic approximation

In this section, we consider the convergence property when the algorithm includes random fluctuation. It can be analyzed by using the stochastic approximation theory, and firstly we will show the result essentially equivalent to Yuille[16]:

Theorem 4 *By the CD method, w converges to w^* (that is different from ML solution in general) in probability when the following conditions are satisfied.*

$$\gamma_t = 1/t, \quad |\nabla_w E(w)| < \infty \tag{24}$$

$$(\eta_{\rm I0} - \eta_{\rm I}(p_0 Q^{[K]}_{w^*}))(w - w^*) \geq {}^{\exists} C_1 |w - w^*|^2, \tag{25}$$

We just review the stochastic approximation theory briefly. The purpose of the theory is to analyze the stochastic difference equation

$$w_{k+1} = w_k + \gamma_k s(w_k, y_k) \tag{26}$$

where w is a parameter and y is any random variable. Let us define the average difference by

$$\bar{s}(w) = \langle s(w, y) \rangle_{p(y|w)} \tag{27}$$

and Lyapunov function $V(w)$ by

$$V(w) \geq 0, \quad \nabla_w V(w^*) = 0, \tag{28}$$

$$(\nabla_w V(w))^T \bar{s}(w) < 0, {}^{\forall} w \neq w^*, \tag{29}$$

then the following theorem holds.

Theorem 5 (Stochastic approximation) *w_k converges to w^* in probability when*

$$-(\nabla_w V(w))^T \bar{s}(w) \geq C_1 V(w), \tag{30}$$

$$\sum_{k=1}^{\infty} \gamma_k = \infty, \quad \sum_{k=1}^{\infty} \gamma_k^2 < \infty \tag{31}$$

$$E_y[\|s(w, y)\|^2] < C_2(1 + V(w)) \tag{32}$$

Typical choice of γ_k that satisfies the above condition is $\gamma_k = 1/k$. Taking account the average difference of CD method is given by $\bar{s}(w) = -(\eta_{\rm I0} - \eta_{\rm I}(p_0 Q^{[K]}_w))$ and letting $V(w) = \|w - w^*\|^2$, we obtain Theorem 4.

5 Progressive CD method

5.1 Proposed method

One reason why the CD method does not always converge to the ML solution is that the initial distribution of MCMC is reset to the empirical distribution $p_0(X)$.

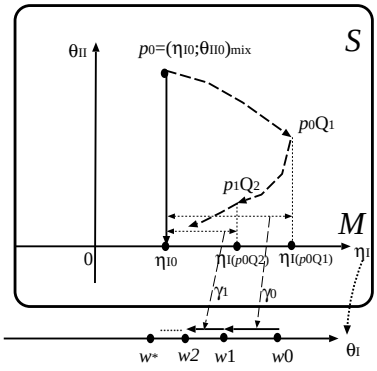

Figure 6: Geometry of Progressive Contrastive Divergence

Our basic idea to improve the CD method is to use the MCMC samples of the current step as the initial population of the next step of the MCMC. We call this method as "Progressive contrastive divergence" (PCD), because the initial population for each MCMC is updated progressively as time goes by.

[PCD algorithm]

1. Initialize the parameter w_0.
2. Initialize the working sample set by the given samples: $\{\hat{X}(1), \hat{X}(2), \ldots, \hat{X}(n)\} \Leftarrow \{X(1), X(2), \ldots, X(n)\}$
3. Gradient descent loop: repeat the following MCMC loop for $t = 0, 1, 2, 3, \ldots$ until convergence
   (a) MCMC loop by population: for each i, let the initial state $X_i^{(0)}$ be each working sample $\hat{X}(i)$ $(i = 1, 2, \ldots, n)$, then perform MCMC for K steps to get the MCMC sample $X^{(K)}(i)$.
   (b) Update w_t by (7) in which the $r_w(X)$ be the empirical distribution after K-step MCMC
   $$r_w(X) = \frac{1}{n} \sum_{i=1}^{n} \delta(X - X^{(K)}(i)) \tag{33}$$
   (c) Update the working sample set by
   $$\hat{X}(i) \Leftarrow X^{(k)}(i) \tag{34}$$

The distribution of working sample set is given by

$$p_0 \prod_{t=1}^{T} Q_t^{[K]} \tag{35}$$

at the T-th step of the gradient descent.

5.2 Convergence property

We can prove the weak consistency (unbiasedness) of the PCD method as a direct extension of the case of CD method.

Theorem 6 *If the PCD method converges in probability, the convergent point is consistent with the stationary point of the original gradient descent equation (5) of the ML solution, which is a local optimum in general.*

However, the convergence of the PCD method is not always guaranteed. We will give the condition for convergence based on the stochastic approximation.

Theorem 7 *By PCD method, w converges to w^* in probability when the following conditions are satisfied.*

$$\gamma_t = 1/t, \quad |\nabla_w E(w)| < \infty \tag{36}$$

$$(\eta_{\mathrm{I}0} - \eta_{\mathrm{I}}(pQ_{w^*}^{[K]}))(w - w^*) \geq {}^{\exists}C_1|w - w^*|^2, \tag{37}$$

for any state vector p.

This theorem gives a very loose sufficient condition for convergence, because it requires the condition for any state vector p instead of the given sample distribution p_0. PCD may have worse convergence property (it might have larger variance) than CD, but it is unbiased.

Therefore, in practice, we need some kinds of technique to avoid unstability of the algorithm, for example, if we increase the population size, the algorithm approaches the average behavior that is much milder and more stable. In the early stage of the algorithm, CD method will show a good performance. PCD can be used to converge to the unbiased solution in the later stage.

6 Conclusion

We have given an information geometrical interpretation of the contrastive divergence (CD) method, and have clarified why CD method can approximate the maximum likelihood solution well. Further, we proposed the progressive contrastive divergence method (PCD) as an unbiased version of the CD method, while PCD might have larger variance (we need further theoretical analysis). The condition of convergence is still very loose and obtain tighter condition is left as a future work.

References

[1] S. Akaho, The e-PCA and m-PCA: dimension reduction by information geometry, *Proc. of Int. Joint Conf. on Neural Networks (IJCNN)*, 2004.

[2] S. Amari, *Differential Geometrical Methods in Statistics, Lecture Notes in Statistics*, Vol. 28, Springer-Verlag, 1985.

[3] S. Amari, Information Geometry of the EM and em Algorithms for Neural Networks, *Neural Networks*, Vol. 8, No. 9, pp. 1379–1408, 1995.

[4] M.Á. Carreira-Perpiñán and G.E. Hinton, On Contrastive Divergence Learning, *Artificial Intelligence and Statistics*, 2005

[5] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, 1981.

[6] S. Geman and D. Geman, Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, *IEEE Trans. on PAMI*, Vol. 6, No. 6, pp. 721–741, 1984.

[7] W.R. Gilks, S. Richardson and D.J. Spiegelhalter (eds.), *Markov Chain Monte Carlo in Practice*, Chapman & Hall, 1996.

[8] G.E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, Vol. 14, No. 8, pp. 1771–1800, 2002.

[9] S. Ikeda, T. Tanaka, S. Amari, Information geometry of turbo and low-density parity-check codes, *IEEE Trans. on Information Theory*, Vol. 50, No. 6, pp. 1097–1114, 2004.

[10] M.I. Jordan (eds.), *Learning in Graphical Models*, MIT Press, 1998.

[11] N. Murata, S. Eguchi, T. Takenouchi, T. Kanamori, Information Geometry of U-Boost and Bregman Divergence, *Neural Computation*, Vol. 16, No. 7, pp. 1437–1481, 2004.

[12] E. Seneta, *Non-negative Matrices and Markov Chains, Second Edition*, Springer-Verlag, 1981.

[13] K. Takabatake, Information Geometry of Gibbs Sampler, *Proc. of WSEAS Int. Conf. on Neural Networks and Applications (NNA)*, 2004.

[14] P.J.M. van Laarhoven and E.H.L. Aarts, *Simulated Annealing: Theory and Applications*, Kluwer Academic Publishers, 1987.

[15] C.K.I. Williams and F.V.Agakov, An Analysis of Contrastive Divergence Learning in Gaussian Boltzmann Machines, *Technical Report*, EDI-INF-RR-0120, Division of Informatics, University of Edinburgh, 2002.

[16] A. Yuille, The convergence of contrastive divergences, *NIPS04*, 2004.