Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables

Revised submission to IEEE TNN

Aapo Hyvärinen

Dept of Computer Science and HIIT

University of Helsinki *

21st February 2007

Abstract

Score matching and contrastive divergence are two recently proposed methods for estimation of non-normalized statistical methods without computation of the normalization constant (partition function). Although they are based on very different approaches, we show in this paper that they are equivalent in a special case: in the limit of infinitesimal noise in a specific Monte Carlo method. Further, we show how these methods can be interpreted as approximations of pseudolikelihood.

1 Introduction

Denote by $\mathbf{x}_t, t = 1, \ldots, T$ as sample of data in an n-dimensional real space. Denote by $q(\mathbf{x}; \boldsymbol{\theta})$ an unnormalized pdf parameterized by the vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$. The problem is to estimate the model parameters when the computation of the normalization constant or partition function $Z(\boldsymbol{\theta}) = \int q(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x}$ is computationally very difficult.

*Dept of Computer Science, P.O.Box 68, FIN-00014 University of Helsinki, Finland. Fax: +358-9-191 51120, email: aapo.hyvarinen@helsinki.fi

A recently proposed approach is contrastive divergence [3]. While there are different versions of contrastive divergence (CD), we concentrate here on the version which is used in practice: the gradient-based version. Denote by s the iteration index and by α a step size parameter. The iteration consists of

$$\boldsymbol{\theta}_{s+1} = \boldsymbol{\theta}_s + \alpha[\frac{\partial \log q(\mathbf{x}_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \frac{\partial \log q(\mathbf{x}_t^*(\boldsymbol{\theta}_s); \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}] \tag{1}$$

where $\mathbf{x}_t^*$ denotes a point obtained by making a single step of a Markov Chain Monte Carlo (MCMC) method starting from the point $\mathbf{x}_t$. CD is fundamentally based on these two ideas: taking a single step of the Markov chain instead of many, and starting the iterations from the observed data points instead of random points. We write $\mathbf{x}_t^*$ as a function of $\boldsymbol{\theta}$ because the distribution of $\mathbf{x}_t^*$ depends on the value of $\boldsymbol{\theta}$ used in the MCMC method. The data point $\mathbf{x}_t$ is randomly chosen from the sample at each iteration of (1).

An important point is that here the value of $\boldsymbol{\theta}$ in the MCMC step is fixed to the current value $\boldsymbol{\theta}_s$, i.e. the differentiation on the right-hand side considers $\boldsymbol{\theta}_s$ as a constant. Thus, the differentiation is simplified because the dependence of the MCMC method on the value of $\boldsymbol{\theta}$ is not taken into account.

Another approach to estimation of non-normalized models is score matching [5]. The starting point is the gradient of the log-density with respect to the data vector. For simplicity, we call this the score function, although according the conventional definition, it is actually the score function with respect to a hypothetical location parameter [8]. For the normalized model density p, we denote the score function by $\boldsymbol{\psi}(\boldsymbol{\xi}; \boldsymbol{\theta})$:

$$\boldsymbol{\psi}(\boldsymbol{\xi}; \boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial \log p(\boldsymbol{\xi}; \boldsymbol{\theta})}{\partial \xi_1} \\ \vdots \\ \frac{\partial \log p(\boldsymbol{\xi}; \boldsymbol{\theta})}{\partial \xi_n} \end{pmatrix} = \begin{pmatrix} \psi_1(\boldsymbol{\xi}; \boldsymbol{\theta}) \\ \vdots \\ \psi_n(\boldsymbol{\xi}; \boldsymbol{\theta}) \end{pmatrix} = \nabla_{\boldsymbol{\xi}} \log p(\boldsymbol{\xi}; \boldsymbol{\theta})$$

The point in using the score function is that it does not depend on $Z(\boldsymbol{\theta})$. In fact we obviously have

$$\boldsymbol{\psi}(\boldsymbol{\xi}; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\xi}} \log q(\boldsymbol{\xi}; \boldsymbol{\theta}) \tag{2}$$

Likewise, we denote by $\boldsymbol{\psi}_{\mathbf{x}}(.) = \nabla_{\boldsymbol{\xi}} \log p_{\mathbf{x}}(.)$ the score function of the distribution of observed data.

The basic principle is that the score function should be equal for the observed distribution and the model distribution. Thus, we minimize a function of the form

$$J(\boldsymbol{\theta}) = \frac{1}{2} \int_{\boldsymbol{\xi} \in \mathbb{R}^n} p_{\mathbf{x}}(\boldsymbol{\xi}) \|\boldsymbol{\psi}(\boldsymbol{\xi}; \boldsymbol{\theta}) - \boldsymbol{\psi}_{\mathbf{x}}(\boldsymbol{\xi})\|^2 d\boldsymbol{\xi} \tag{3}$$

This could in principle be accomplished by computing the gradient of the logarithm of a non-parametric estimate of the pdf, but it was shown in [5] that no such computations are necessary. The function in (3) is equal to an expression which only uses simple derivatives of q. A sample version of the resulting objective function is given by

$$J_{SM}(\boldsymbol{\theta}) = \sum_t \sum_{i=1}^{n} \frac{1}{2}\psi_i(\mathbf{x}_t;\boldsymbol{\theta})^2 + \psi_{ii}(\mathbf{x}_t;\boldsymbol{\theta}) \tag{4}$$

where $\psi_{ii} = \partial\psi_i/\partial\xi_i$. The resulting estimator can be shown to be (locally) consistent [5].

Thus, these two methods use completely different approaches, and do not seem to have any connection with each other. However, in this paper, we show a deep connection between these two methods. Furthermore, we show an approximative connection to a third method, pseudolikelihood.

2 Score matching as deterministic contrastive divergence

Here we consider a Langevin Monte Carlo method [7], in which a new point $\mathbf{x}_t^*$ is obtained from the current point (here: observed data point) $\mathbf{x}_t$ as

$$\mathbf{x}_t^*(\boldsymbol{\theta}_s) = \mathbf{x}_t + \frac{\mu^2}{2}\nabla_{\mathbf{x}} \log q(\mathbf{x}_t;\boldsymbol{\theta}_s) + \mu\,\mathbf{n} \tag{5}$$

where $\mathbf{n}$ is standardized white gaussian noise. This is the uncorrected version of the Langevin method. Below, we will consider the limit of infinitesimal μ, in which case the uncorrected version is the same as a corrected one [7]. Such a method can be considered a first-order approximation (in the limit of infinitesimal μ) of more sophisticated Monte Carlo methods [7].

Denote the second partial derivatives of $\log q$ by

$$\psi_{ij}(\mathbf{x};\boldsymbol{\theta}) = \frac{\partial^2 \log q(\mathbf{x};\boldsymbol{\theta})}{\partial x_i x_j} \tag{6}$$

A simple Taylor expansion gives

$$\log q(\mathbf{x}_t^*(\boldsymbol{\theta}_s);\boldsymbol{\theta}) = \log q(\mathbf{x}_t;\boldsymbol{\theta}) + \sum_i \psi_i(\mathbf{x}_t;\boldsymbol{\theta})[\frac{\mu^2}{2}\psi_i(\mathbf{x}_t;\boldsymbol{\theta}_s) + \mu\, n_i] + \frac{1}{2}\sum_{i,j} n_i n_j \psi_{ij}(\mathbf{x}_t;\boldsymbol{\theta})\mu^2 + o(\mu^2)$$

Here, we take a second-order expansion, because it turns out that in the averaged version, the first-order terms disappear, and the second-order terms are the most significant.

Consider the difference of the unnormalized log-likelihoods as in (1). Let us compute the expectation of the difference with respect to the Monte Carlo noise $\mathbf{n}$; denote the result by J_{CD}. We

obtain

$$J_{CD}(\boldsymbol{\theta}, \boldsymbol{\theta}_s, \mathbf{x}_t) = E_{\mathbf{n}}\{\log q(\mathbf{x}_t; \boldsymbol{\theta}) - \log q(\mathbf{x}_t^*(\boldsymbol{\theta}_s); \boldsymbol{\theta})\}$$
$$= -\frac{\mu^2}{2}[\sum_i \psi_i(\mathbf{x}_t; \boldsymbol{\theta})\psi_i(\mathbf{x}_t; \boldsymbol{\theta}_s) + \sum_i \psi_{ii}(\mathbf{x}_t; \boldsymbol{\theta})] + o(\mu^2) \quad (7)$$

because the expectation of n_i is zero, and it is white and independent from the $\mathbf{x}_t$.

Now, we can analyze the averaged behaviour of contrastive divergence by looking at the gradient of the difference in (7) with respect to $\boldsymbol{\theta}$, averaged over all $\mathbf{x}_t$. Note that J_{CD} does not provide a proper objective function for the algorithm because it depends on the current value $\boldsymbol{\theta}_s$ as well. However, the gradient of J_{CD} with respect to $\boldsymbol{\theta}$ (for fixed $\boldsymbol{\theta}_s$) does give us the original CD iteration in (1) averaged over the Monte Carlo noise $\mathbf{n}$. For notational simplicity, we consider the partial derivatives with respect to the θ_k. We obtain

$$\frac{\partial J_{CD}(\boldsymbol{\theta}, \boldsymbol{\theta}_s, \mathbf{x}_t)}{\partial \theta_k} = -\frac{\mu^2}{2}[\sum_i \frac{\partial \psi_i(\mathbf{x}_t; \boldsymbol{\theta})}{\partial \theta_k}\psi_i(\mathbf{x}_t; \boldsymbol{\theta}_s) + \sum_i \frac{\partial \psi_{ii}(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_k}] + o(\mu^2) \quad (8)$$

In the algorithm, the running value of the estimate, $\boldsymbol{\theta}$, will be equal to $\boldsymbol{\theta}_s$. The step size μ will be infinitesimal, so the term $o(\mu^2)$ can be ignored. Thus, let us consider functions of the form

$$j_k(\boldsymbol{\theta}) = -\sum_t \sum_i \frac{\partial \psi_i(\mathbf{x}_t; \boldsymbol{\theta})}{\partial \theta_k}\psi_i(\mathbf{x}_t; \boldsymbol{\theta}) + \sum_i \frac{\partial \psi_{ii}(\mathbf{x}_t; \boldsymbol{\theta})}{\partial \theta_k} \quad (9)$$

which take the average over all the $\mathbf{x}_t$.

The average behaviour of the CD algorithm, averaged over both the Monte Carlo noise $\mathbf{n}$ and the sample index t, can now be described as the addition of the $j_k(\boldsymbol{\theta})$, multiplied by a very small step size, to the current estimates of θ_k. This approximation is valid if 1) the step size α is annealed to zero as required in a classic stochastic approximation scheme, and 2) we consider the limit of the step size μ being infinitesimal.

Our main result is that the CD algorithm with Langevin Monte Carlo can then be interpreted as a gradient descent on the score matching objective function [5]. In fact, it is easy to see that

$$j_k(\boldsymbol{\theta}) = -\frac{\partial J_{SM}(\boldsymbol{\theta})}{\partial \theta_k} \quad (10)$$

where J_{SM} is the score matching objective function in (4).

Thus, we have proven that score matching is an infinitesimal deterministic variant of contrastive divergence using the Langevin Monte Carlo method. In particular, computation of the second derivative of the log-pdf in (4) is performed by a numerical Monte Carlo method in CD, adding

Gaussian noise. In contrast, score matching is based on explicit algebraic formulae, in which all derivatives and integrals can typically be computed without resorting to numerical differentiation or Monte Carlo integration.

3 Simulations

To investigate the practical validity of the result given above, we performed simulations where contrastive divergence, score matching, and maximum likelihood were used to estimate a basic ICA model in two dimensions [5]. We took 100 observations from the model, and computed the gradients of score matching objective function and likelihood at the point of the true separating matrix. For CD, we computed the expected "gradient" over 5,000,000 Monte Carlo samples, with μ^2 given values $0.1, 0.01, 0.001, 0.0001$, and 0.00001. These three matrix gradients were normalized to unit Frobenius norm, and the Frobenius distances between them compared over 10 runs. Specifically, we computed the ratio of the distance between the gradients for contrastive divergence and score matching, to the distance between the gradients for score matching and maximum likelihood. If this ratio is very small our approximation is valid.

The results are shown in Figure 1. For moderate values of μ^2, the mean ratio is only a few percent, obtaining a minimum of 3.4% for $\mu^2 = 0.001$. However, when μ is larger or smaller, the average ratios are large. For a large μ, this is expected because our approximation assumes infinitesimal μ. For smaller μ, the reason seems to be that for then the noise in the Langevin equation dominates over the deterministic part, and thus the gradient is very noisy. So, the plots shows a large error because the expected CD gradient is not computed exactly enough using this sample size. This was confirmed by additional simulations (not shown) in which the number of Monte Carlo samples was increased by a factor of 10: The approximation was then quite good for $\mu^2 = 0.0001$ as well, with an average ratio of 7.8% (max 31.3%, min 1.3% out of six trials). Larger MC sample sizes were not investigated because they required excessive computation time.

Thus, overall the SM and CD gradients were much more similar to each other than they were to the likelihood gradient, which is in line with our theoretical result.

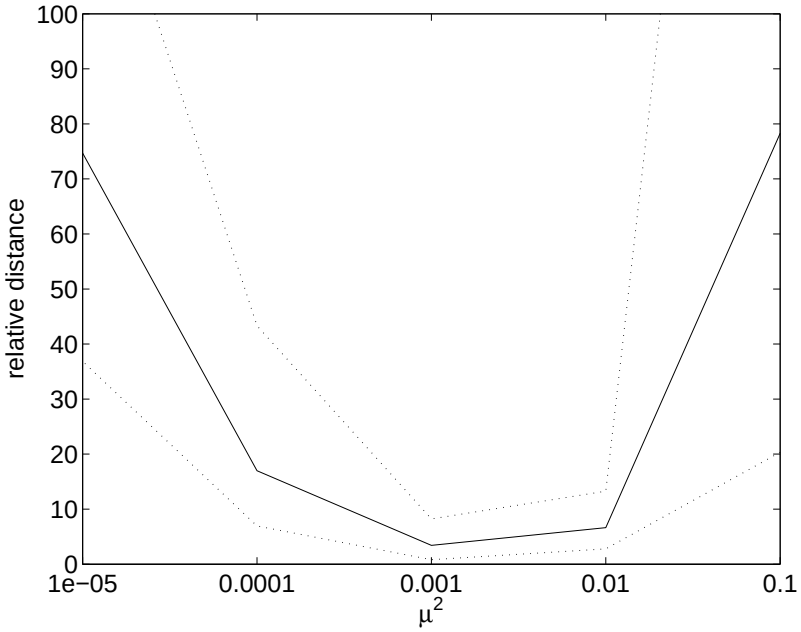

Figure 1: The relative distance between CD and score matching gradients as a function of μ^2. This is defined as the ratio, in percentages, of the distance between the expected CD gradient and the score matching gradient to the distance between the score matching and likelihood gradients. Solid line shows the mean value over 10 trials, dotted lines show the minimum and maximum values over the trials.

4 Connection with pseudolikelihood

A related method for estimating non-normalized models is pseudolikelihood [2].The basic idea is to consider the conditional pdf's $p(x_i|x_1,\dots,x_{i-1},x_{i+1},\dots,x_n;\boldsymbol{\theta})$, i.e. conditional densities of the random variable given all other variables, where $\boldsymbol{\theta}$ denotes the parameter vector. Let us denote by $\mathbf{x}^{\notin i}$ the vector with x_i removed:

$$\mathbf{x}^{\notin i} = (x_1,\dots,x_{i-1},x_{i+1},\dots,x_n) \tag{11}$$

and the logarithms of the conditional pdf's by

$$c_i(x_i;\mathbf{x}^{\notin i},\boldsymbol{\theta}) = \log p(x_i|\mathbf{x}^{\notin i},\boldsymbol{\theta}) = \log p(\mathbf{x};\boldsymbol{\theta}) - \log \int p(\mathbf{x};\boldsymbol{\theta})dx_i \tag{12}$$

We then estimate the model by maximizing these conditional pdf's in the same way as one would maximize ordinary likelihood. Given a sample $\mathbf{x}(1),\dots,\mathbf{x}(T)$, the pseudolikelihood (normalized as a function of sample size by dividing by T) is thus of the form

$$J_{PL}(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n} c_i(x_i(t);\mathbf{x}^{\notin i}(t),\boldsymbol{\theta}) \tag{13}$$

Pseudolikelihood is, in some special cases, equivalent to contrastive divergence. This was shown in [6] for a fully-visible Boltzmann machine, with CD based on Gibbs sampling. However, such a result is unlikely to hold in general.

The gradient of pseudolikelihood can be computed as

$$\nabla_{\boldsymbol{\theta}} J_{PL}(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x};\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} \log \int p(\mathbf{x};\boldsymbol{\theta})dx_i \tag{14}$$

As shown in [1], such a gradient can be equivalently expressed as

$$\nabla_{\boldsymbol{\theta}} J_{PL}(\boldsymbol{\theta}) = \sum_{i=1}^{n} E_{\mathbf{x}} \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x};\boldsymbol{\theta}) - E_{x_i|x^{\notin i}} \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x};\boldsymbol{\theta}) \tag{15}$$

where $E_{\mathbf{x}}$ means expectation with respect to the observed (sample) distribution of $\mathbf{x}$, and $E_{x_i|\mathbf{x}^{\notin i}}$ means expectation with respect to the conditional distribution of x_i given all other $x_j, i \neq j$.

Here we see that with continuous-valued data, pseudolikelihood is computationally much more complex than SM, because computation of the conditional pdf in (12) needs numerical integration with respect to x_i, or, equivalently, sampling from the conditional distribution $x_i|\mathbf{x}^{\notin i}$ in (15).

Now consider what happens if we approximate this numerical integration with an MCMC method, and only take a single step in the MCMC method, *à la* contrastive divergence. Specifically, consider

a Langevin iteration which is only run with respect to x_i, keeping all other variables fixed:

$$x_{i,t}^{+}(\boldsymbol{\theta}_s) = x_{i,t} + \frac{\mu^2}{2}\frac{\partial \log q(\mathbf{x}_t;\boldsymbol{\theta}_s)}{\partial x_i} + \mu\,\nu \tag{16}$$

where ν is a standardized gaussian variable. If such an iteration were run infinitely, we would obtain a sample of the conditional distribution $x_i|\mathbf{x}^{\not\in i}$. However, we *approximate* such a Langevin method by using a single iteration.

Consider a variable $\mathbf{x}'$ which is obtained by applying the iteration in (16) on a randomly chosen index k. Then the distribution of $\mathbf{x}'$ can be simply expressed using the original Langevin iteration in (5): Define a new noise variable $\mathbf{n}'$ which is zero in all other directions expect for the variable x_k where k is chosen randomly. Then, equation (5) holds, with these new definitions of $\mathbf{x}'$ instead of $\mathbf{x}^*$ and $\mathbf{n}'$ instead of $\mathbf{n}$. The point is that since (15) takes the sum over all i, it is in the expectation equal to using the newly defined $\mathbf{x}'$ as the sampling distribution, i.e.

$$\nabla_{\boldsymbol{\theta}} J_{PL}(\boldsymbol{\theta}) = \sum_{i=1}^{n} E_{\mathbf{x}} \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x};\boldsymbol{\theta}) - E_{\mathbf{x}'} \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}';\boldsymbol{\theta}) \tag{17}$$

which brings us back to the definition of CD in (1), in an averaged form. All our analysis of the Langevin iteration is equally valid for $\mathbf{n}'$, because our analysis only used the mean and the covariance of $\mathbf{n}$, and they are the same for $\mathbf{n}$ and $\mathbf{n}'$. (Actually, the variance of n_i' is scaled by a constant that depends on the dimension of the data space, but this scaling effect can be cancelled by rescaling μ accordingly.)

Thus, we have shown that score matching and contrastive divergence can be considered as approximations of pseudolikelihood, obtained by running a single step of an MCMC method to compute the conditional pdf needed in pseudolikelihood.

5 Conclusion

We have shown that score matching and contrastive divergence are equivalent in the limit of infinitesimal step size, and with a particular MCMC method. Further, we showed how these methods can be considered approximations of pseudolikelihood.

Our results imply that the statistical results on score matching can be applied on this variant of contrastive divergence, including consistency [5] and optimality in signal restoration [4]. However, as the equivalency only holds in the limit of infinitesimal μ, and other Monte Carlo methods are

often preferred in practice, the actual performance of contrastive divergence may differ from that of score matching. The performance of pseudolikelihood can be even more different because the approximation is unlikely to be at all exact in practice and has mainly theoretical interest.

Acknowledgements

The basic idea in this paper was developed in discussions with Geoffrey Hinton, who, however, preferred not to be a coauthor because the paper contained too many equations. Funding from the Canadian Institute for Advanced Research and the Academy of Finland are gratefully acknowledged.

References

[1] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9:147–169, 1985.

[2] J. Besag. Statistical analysis of non-lattice data. *The Statistician*, 24:179–195, 1975.

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

[4] A. Hyvärinen. Optimal approximation of signal priors. Submitted manuscript. Available at `www.cs.helsinki.fi/aapo.hyvarinen/papers/et.shtml`.

[5] A. Hyvärinen. Estimation of non-normalized statistical models using score matching. *J. of Machine Learning Research*, 6:695–709, 2005.

[6] A. Hyvärinen. Consistency of pseudolikelihood estimation of fully visible Boltzmann machines. *Neural Computation*, 18(10):2283–2292, 2006.

[7] R. M. Neal. Probabilistic inference using Markov Chain Monte Carlo methods. Technical report, Dept of Computer Science, University of Toronto, 1993.

[8] M. Schervish. *Theory of Statistics*. Springer, 1995.