LETTER Communicated by Aapo Hyvärinen

Contrastive Divergence in Gaussian Diffusions

Javier R. Movellan
movellan@mplab.ucsd.edu
Institute for Neural Computation, University of California San Diego, La Jolla, CA 92093-0515, U.S.A.

This letter presents an analysis of the contrastive divergence (CD) learning algorithm when applied to continuous-time linear stochastic neural networks. For this case, powerful techniques exist that allow a detailed analysis of the behavior of CD. The analysis shows that CD converges to ML solutions only when the network structure is such that it can match the first moments of the desired distribution. Otherwise, CD can converge to solutions arbitrarily different from the log-likelihood solutions, or they can even diverge. This result suggests the need to improve our theoretical understanding of the conditions under which CD is expected to be well behaved and the conditions under which it may fail. In, addition the results point to practical ideas on how to improve the performance of CD.

1 Introduction

CD is a recent learning rule found to work well in practice despite still unclear theoretical underpinnings (Hinton, 2002; Hinton & Salakhutdinov, 2006; Hyvarinen, 2006; MacKay, 2001; ¨ Carreira-Perpinan & Hinton, 2005; Roth & Black, 2005; Williams & Agakov, 2002; Yuille, 2004). This letter presents an analysis of CD in Gaussian diffusions—a linear, continuous-time, continuous-state version of RNNs. These networks are of interest for two reasons: (1) powerful analytical tools exist that allow comparing the behavior of CD to other algorithms, like MLE, and (2) many nonlinear systems of interest for which CD has proven useful have multiple attractors about which the systems behave locally like gaussian diffusions. Thus, the analysis of the gaussian diffusion case may provide clues for a better understanding of CD in more general conditions. The analysis presented here shows that convergence of CD is guaranteed if the first moment of the Gaussian diffusion is at equilibrium. In this case, CD and MLE converge to the same solution; otherwise, CD may converge to arbitrarily different solutions from MLE or diverge altogether.

Neural Computation **20, 2238–2252** (2008)

In this letter, we pursue a continuous-time formulation of CD that makes possible the use of stochastic calculus tools. The continuous-time case can be seen as the limit of the dynamics induced by the uncorrected discrete time Langevin MCMC method (Neal, 1996). In addition, it should be noted that CD is typically interpreted as a method for learning equilibrium distributions while here we also examine it as a method for learning finite time distributions.

Ornstein-Uhlenbeck process

Consider a stochastic process $X = \{X_t : t \in \mathcal{R}_+\}$ defined by the SDE,

$$dX_t = \theta(\gamma - X_t)dt + \sqrt{2\tau}\, dB_t, \tag{1.1}$$

$$X_0 \sim \mathcal{N}(\mu_0, \sigma_0), \tag{1.2}$$

Here we interpret the process as a neural network, where θ is a sym. p. d. matrix of synaptic connections, γ is a fixed vector of synaptic biases that determine the mean of the equilibrium distribution, $\tau > 0$ is a fixed parameter that controls the degree of noise in the network, and dB_t is a Brownian motion differential. The solution to this equation is (Movellan, 2006b; Oksendal, 1992):

Ito formula

$$X_t = e^{-t\theta}\left(X_0 + (e^{t\theta} - I)\gamma + \sqrt{2\tau}\int_0^t e^{s\theta}\, dB_s\right). \tag{1.3}$$

Thus, $X_t \sim \mathrm{N}(\mu_t, \sigma_t)$ (Movellan, 2006b):

$$\mu_t \stackrel{\text{def}}{=} \mathbb{E}[X_t] = e^{-t\theta}\mu_0 + (I - e^{-t\theta})\gamma, \tag{1.4}$$

$$\sigma_t \stackrel{\text{def}}{=} \mathbb{C}\mathrm{ov}[X_t] = \tau\theta^{-1} + (\sigma_0 - \tau\theta^{-1})e^{-2t\theta}. \tag{1.5}$$

At equilibrium,

$$\mu_\infty \stackrel{\text{def}}{=} \lim_{t\to\infty} \mu_t = \gamma, \tag{1.6}$$

$$\sigma_\infty \stackrel{\text{def}}{=} \lim_{t\to\infty} \sigma_t = \tau\theta^{-1}, \tag{1.7}$$

and therefore

$$\mu_t = \mu_\infty + e^{-t\theta}(\mu_0 - \mu_\infty) = \mu_0 + (I - e^{-t\theta})(\mu_0 - \mu_\infty), \tag{1.8}$$

$$\sigma_t = \sigma_\infty + e^{-2t\theta}(\sigma_\infty - \sigma_0) = \sigma_0 + (I - e^{-2t\theta})(\sigma_\infty - \sigma_0). \tag{1.9}$$

Express the distribution of X_t in the following Boltzmann form,

$$p(x_t) \propto e^{\phi_t(x_t)}, \tag{1.10}$$

$$\phi_t(x) \stackrel{\text{def}}{=} x'\sigma_t^{-1}\mu_t - \frac{1}{2}x'\sigma_t^{-1}x, \tag{1.11}$$

where $-\phi_t$ is the potential at time t.

2 ML and CD

The process X induces a family of distributions parameterized by t, θ, and γ. For now, we will treat the equilibrium mean γ as a fixed value and the connectivity matrix θ as an adaptive parameter. We will define learning as the process of finding values of θ under which the distribution of X_t approximates the distribution of a target rv ξ.

The method of ML calls for values of θ that maximize the likelihood function . Local maxima can be found by progressively changing θ in the direction of the log-likelihood gradient . For Boltzmann ~~distributions , the~~ log-likelihood gradient takes the following form (the appendix, lemma 1),

$$\nabla_\theta \mathbb{E}[\log p_{X_t}(\xi)] = \mathbb{E}[\Psi_t(\xi)] - \mathbb{E}[\Psi_t(X_t)], \tag{2.1}$$

where $\Psi_t(x)$ is the unnormalized Fisher score function: $\Psi_t(x) \stackrel{\text{def}}{=} \nabla_\theta \phi_t(x)$.

CD was designed for situations in which the equilibrium potential $-\phi_\infty \stackrel{\text{def}}{=} \lim_{t\to\infty} -\phi_t$ is known but the finite time potentials are unknown. Rather than waiting for equilibrium conditions, CD operates with a finite $t > 0$ and progressively changes θ in the direction of Hinton's CD statistic:

$$H_t \stackrel{\text{def}}{=} \mathbb{E}[\Psi_\infty(\xi)] - \mathbb{E}[\Psi_\infty(X_t)]. \tag{2.3}$$

In Gaussian diffusions, there are analytical expressions for the potentials at all times, thus allowing a direct comparison between ML and CD. It can be shown that the Fisher score function (the appendix, theorem 1),

$$\boxed{\Psi_t(x) = x\mu_t' c_t + t x \sigma_t^{-1} e^{-t\theta}(\gamma - \mu_0)' - \frac{1}{2} x x' c_t,} \tag{2.4}$$

where c_t is a p.d. matrix:

$$c_t \stackrel{\text{def}}{=} \tau \sigma_t^{-2} \left(\theta^{-2}(I - e^{-2t\theta}) - 2t\left(\theta^{-1} - \frac{1}{\tau}\sigma_0\right) e^{-2t\theta} \right). \tag{2.5}$$

Thus, considering that $\boxed{\lim_{t\to\infty} c_t = 1/\tau}$, it follows that

$$\Psi_\infty(x) = \frac{1}{\tau} x \left(\gamma - \frac{1}{2} x \right)'. \tag{2.6}$$

eq (2.1) <- (2.4) ==>

$$\nabla_\theta \mathbb{E}[\log p_{X_t}(\xi)] = \frac{1}{2}(\mathbb{E}[X_t X_t'] - \mathbb{E}[\xi\xi'])c_t + (\mathbb{E}[\xi] - \mu_t)\mu_t' c_t + t\sigma_t^{-1} e^{-t\theta}[\mathbb{E}(\xi) - \mu_t](\gamma - \mu_0)'. \tag{2.7}$$

The gradient for the equilibrium distribution can be obtained by taking the limit as $t \to \infty$:

$$\nabla_\theta \mathbb{E}[\log p_{X_\infty}(\xi)] = \frac{1}{2\tau}[\mathbb{E}(X_\infty X_\infty') - \mathbb{E}(\xi\xi')] + \frac{1}{\tau}[\mathbb{E}(\xi) - \gamma]\gamma'. \tag{2.8}$$

eq (2.3) <- (2.6) ==> Hinton's CD statistic:

$$H_t = \frac{1}{2\tau}[\mathbb{E}(X_t X_t') - \mathbb{E}(\xi\xi')] + \frac{1}{\tau}[\mathbb{E}(\xi) - \mu_t]\gamma'. \tag{2.9}$$

Note (2.9) -->

$$\nabla_\theta \mathbb{E}[\log p_{X_t}(\xi)] = \tau H_t c_t + R_t, \tag{2.10}$$

where the residual term R_t is:

$$R_t \overset{\text{def}}{=} t\sigma_t^{-1} e^{-t\theta}[\mathbb{E}(\xi) - \mu_t](\gamma - \mu_0)' + (\mathbb{E}[\xi] - \mu_t)(\mu_t - \gamma)' c_t. \tag{2.11}$$

Note

$$\boxed{\lim_{t\to\infty} H_t = \nabla_\theta \mathbb{E}[\log p_{X_t}(\xi)],} \tag{2.12}$$

Thus, in the limit Hinton's statistic becomes the gradient of the log likelihood. Hinton (2002) derived the H_t statistic as an approximation to the gradient of the difference between two K-L divergences : $\mathbb{D}(\xi, X_\infty) - \mathbb{D}(X_t, X_\infty)$. It can be shown (see the appendix, theorem 2) that[1]

$$\nabla_\theta \left(\mathbb{D}(\xi, X_\infty)\right) - \mathbb{D}(X_t, X_\infty)) = -(H_t + \tilde{R}_t), \tag{2.13}$$

where the residual $\tilde{R}_t$ is a covariance statistic:

$$\tilde{R}_t \stackrel{\text{def}}{=} \mathbb{C}\text{ov}[\phi_t(X_t) - \phi_\infty(X_t), \Psi_t(X_t)]. \tag{2.14}$$

Hinton (2002) proposed that this residual may be ignored in practice, resulting in the CD learning rule: $\Delta\theta \propto H_t$.

We are now ready to examine four learning rules:

- ML_t: MLE for the finite time process,
$$\Delta\theta \propto \nabla_\theta \mathbb{E}[\log p_{X_t}(\xi)] = \tau H_t\, c_t + R_t. \tag{2.15}$$
- ML_∞: MLE for the process at stochastic equilibrium,
$$\Delta\theta \propto \nabla_\theta \mathbb{E}[\log p_{X_\infty}(\xi)] = H_\infty. \tag{2.16}$$
- ECD: Exact CD,
$$\Delta\theta \propto \nabla_\theta \left(\mathbb{D}(\xi, X_\infty)\right) - \mathbb{D}(X_t, X_\infty)) = H_t + \tilde{R}_t. \tag{2.17}$$
- CD: $\Delta\theta \propto H_t$. (2.18)

First, note that as $t \to \infty$ (i.e., if we let the network settle to equilibrium), R_t and $\tilde{R}_t$ vanish, and the four rules converge to the same solution (see the appendix, remark 1). A more interesting question is what happens when t is finite and obviously not enough time has been given for the network to achieve stochastic equilibrium. In this case, the learning rules may converge to different solutions. In fact, when large values of $\mu_0 - \gamma$ are chosen, the residual term R_t in equation 2.11 can be made arbitrarily large to the point that CD may not converge at all or may converge to solutions arbitrarily different from ML_t and ML_∞. However, there are cases of interest in which the learning rules converge to the same results:

- **Case 1**: $\mu_\infty = \mu_0$. It follows that $\mu_t = \mu_\infty = \gamma$. Thus, the residual term $R_t = 0$, and the gradient of the log-likelihood equals Hinton's CD statistic H_t times the p.d. matrix c_t. Thus, in this case, CD and ML_t converge to the same solution.

[1]This result holds for more general processes, not just Gaussian diffusions.

- **Case 2**: $\mu_t = \mathbb{E}(\xi)$. In this case the first moment of the desired distribution has already been learned. Note the residual term R_t in equation 2.11 also vanishes, and thus CD and ML_t converge to the same estimate.
- **Case 3**: This case combines case 1 and case 2: $\mu_\infty = \mu_0 = \mu_t = \mathbb{E}(\xi)$. Under these conditions (see the appendix, remark 1),

$$H_t = H_\infty(I - e^{-2t\theta}). \tag{2.19}$$

Since $I - e^{-2t\theta}$ is a p.d. matrix and H_∞ is proportional to the gradient of ML_∞ ==> CD, ML_t, and ML_∞ have positive inner products with each other and converge to the same solution.

2.1 Summary of Results. The analysis reveals the importance of initializing the network so that the first moment of the states is at equilibrium. If the first moment is not at equilibrium, then CD may converge to solutions arbitrarily different from ML solutions or diverge altogether. If at equilibrium, then CD and ML_t converge to the same solution. If, in addition, $\mu_0 = \mathbb{E}[\xi]$, then CD, ML_t, and ML_∞ converge to the same solution. There currently is nothing in the theory of CD to explain why it converges when the first moment is at equilibrium but may diverge otherwise.

3 Learning the Equilibrium Means

So far we have treated the equilibrium mean, γ, as a fixed vector. This was purposely done to establish that there are conditions under which CD may not converge. In this section, we study what happens if we treat the connectivity matrix θ as a fixed parameter and the bias parameter γ as adaptive. In this case, it can be shown that

$$H_t = \frac{1}{\tau}\theta\left(\mathbb{E}[\xi] - \mu_t]\right), \tag{3.1}$$

$$\nabla_\gamma \mathbb{E}[\log p_{X_t}(\xi)] = \frac{1}{\tau}\theta(I - e^{-2t\theta})(\mathbb{E}[\xi] - \mu_t) = (I - e^{-2t\theta})H_t. \tag{3.2}$$

Thus, since $I - e^{-2t\theta}$, is a p.d. matrix, when applied to the bias parameter γ, both CD and ML_t have positive inner products with each other. In addition, they converge when the first moments of the desired and obtained distributions are matched.

4 The Partially Observable Case

In many cases of interest, the state vector X_t can be divided into a vector of observable units Y_t and a vector of hidden units Z_t: $X_t' = (Y_t', Z_t')$. The goal in this case is for the observable units to approximate the distribution of the target vector ξ. The expectation maximization algorithm (EM) reduces the partially observable case to the fully observable case (Dempster, Laird, & Rubin, 1977). EM operates in an iterative manner. At iteration k, we are given a fixed parameter $\theta^{(k)}$ and a target value ξ for the observable units. The goal then becomes to learn the fully observable joint distribution of $X_t = (\xi, Z_t^{(k)})$, where Z_t^k is the distribution of samples of hidden states given observable state ξ and parameter $\theta^{(k)}$. The parameter $\theta^{(k+1)}$ that optimizes this joint likelihood becomes the starting point for the next iteration. Thus, the results obtained for the fully observable case generalize to the partially observable case.

5 Conclusion

We analyzed the behavior of CD in gaussian diffusion processes. We showed that in this case, CD converges to maximum likelihood solutions if the first moment of the state distribution is at equilibrium; otherwise, CD may diverge. There is nothing in the current theory of CD that would explain the difference in behavior between these two cases. In gaussian diffusion processes, once the first-order moments of the desired distribution have been matched, the CD learning rule achieves positive inner products with the log-likelihood gradients. The nonlinear systems for which CD has proven useful have potential functions with multiple attractors, around which the systems may behave like gaussian diffusions. This may help explain why CD works well in such systems. This view of CD suggests techniques to improve its performance. For example, since the residual term R_t vanishes when the first moment of the state distribution is at equilibrium, a two-stage process could be used: on each learning trial, the system can be run using zero temperature deterministic dynamics, thus allowing it to quickly find the equilibrium mean, followed by the stochastic dynamics to estimate the H_t statistic. In addition, H_t could be estimated more efficiently using deterministic sampling methods, like the unscented transform (Julier, Uhlmann, & Durrant-Whyte, 1995).

Appendix: Derivations

A.1 Notational Conventions. The appendix assumes the processes defined in the main body of the letter. Unless otherwise stated, capital letters are used for rvs, lowercase letters for specific values taken by random variables, and Greek letters for fixed parameters. The operators $\mathbb{E}$ and $\mathbb{D}$ stand, respectively, for expected value and K-L divergence. $\mathcal{R}$ is the set of real numbers. We leave implicit the properties of the probability space in which the rvs are defined. To simplify the notation, we identify probability functions by their arguments, and when it does not lead to confusion, we leave implicit dependencies on network parameters—for example,

$$p_t(x) \equiv p_{X_t(\theta)}(x), \tag{A.1}$$

$$X_t \equiv X_t(\theta), \tag{A.2}$$

$$\phi(x) \equiv \phi(x, \theta). \tag{A.3}$$

Lemma 1. *Let ξ be a target rv, θ be a random parameter, and X be a rv with a Boltzmann distribution:*

$$p_X(u \mid \theta) \stackrel{def}{=} p(X = u \mid \theta) = \frac{1}{Z} e^{\phi(u)}, \tag{A.4}$$

(A.5)

Then

$$\nabla_\theta [\log p_X(\xi) \mid \theta] = [\Psi(\xi) \mid \theta] - \mathbb{E}[\Psi(X) \mid \theta], \tag{A.6}$$

where $\Psi(x, \theta) \stackrel{def}{=} \nabla_\theta \phi(x, \theta)$.

Lemma 2. *Let* $\Sigma_t : Rn \times Rn \to \mathcal{R}^n \times \mathcal{R}^n$,

$$\Sigma_t(\theta) \stackrel{def}{=} \tau \left(\theta^{-1} + \left(\frac{1}{\tau} \sigma_0 - \theta^{-1} \right) e^{-2t\theta} \right), \quad \textit{for } \theta \in \mathcal{R}^n \times \mathcal{R}^n. \tag{A.9}$$

Let $\theta \in \mathcal{R}^n \times \mathcal{R}^n$ be a fixed symmetric invertible matrix and let $a \in \mathcal{R}^n \times \mathcal{R}^n$ be a fixed matrix. Let $\sigma_t \stackrel{def}{=} \Sigma_t(\theta)$, and let $\epsilon \in \mathcal{R}$. Then

$$\begin{aligned} &\frac{d}{d\epsilon} \Sigma_t^{-1}(\theta + \epsilon a) \,|\, 0 \\ &\quad = \frac{1}{\tau} \sigma_t^{-1} \left(\theta^{-1} a \theta^{-1} (I - e^{-2t\theta}) - 2t \left(\theta^{-1} - \frac{1}{\tau} \sigma_0 \right) e^{-2t\theta} a \right) \sigma_t^{-1}. \end{aligned} \tag{A.10}$$

Proof. To first order,

$$
\begin{aligned}
(\theta+\epsilon a)^{-1} &\approx \theta^{-1} - \epsilon\theta^{-1} a \theta^{-1}, && \text{(A.11)}\\
e^{-\epsilon 2ta} &\approx I - \epsilon\, 2ta\,. && \text{(A.12)}
\end{aligned}
$$

Thus, to first order,

$$
\begin{aligned}
\Sigma_t^{-1}(\theta+\epsilon a) &\stackrel{\text{def}}{=} \frac{1}{\tau}\left((\theta+\epsilon a)^{-1} + \left(\frac{1}{\tau}\sigma_0 - (\theta+\epsilon a)^{-1}\right) e^{-2t(\theta+\epsilon a)}\right)^{-1}\\
&\approx \frac{1}{\tau}\Bigg(\theta^{-1} - \epsilon\theta^{-1} a\theta^{-1}\\
&\quad + \left(\frac{1}{\tau}\sigma_0 - \theta^{-1} + \epsilon\theta^{-1} a \theta^{-1}\right) e^{-2t\theta}\,(I - \epsilon 2ta)\Bigg)^{-1}. \qquad \text{(A.13)}
\end{aligned}
$$

Separating out the constant, linear, and quadratic terms wrt ϵ,

$$
\begin{aligned}
\Sigma_t^{-1}(\theta+\epsilon a) &\approx \frac{1}{\tau}\left(\theta^{-1} + \left(\frac{1}{\tau}\sigma_0 - \theta^{-1}\right) e^{-2t\theta}\right)\\
&\quad + \frac{\epsilon}{\tau}\left(-\theta^{-1} a\theta^{-1}(I - e^{-2t\theta}) - 2t\left(\frac{1}{\tau}\sigma_0 + \theta^{-1}\right) e^{-2t\theta} a\right)\\
&\quad + \frac{\epsilon^2}{\tau} 2t\theta^{-1} a\theta^{-1} e^{-2ta}. \qquad \text{(A.14)}
\end{aligned}
$$

Using ~~equation 1.9~~,

$$
\frac{1}{\tau}\sigma_t = \theta^{-1} + \left(\frac{1}{\tau}\sigma_0 - \theta^{-1}\right) e^{-2t\theta}, \qquad \text{(A.15)}
$$

and eliminating residual terms quadratic on ϵ, it follows that to first order,

$$
\begin{aligned}
\Sigma_t^{-1}(\theta+\epsilon a) &\approx \frac{1}{\tau}\Bigg(\frac{1}{\tau}\sigma_t + \epsilon\Bigg(-\theta^{-1} a\theta^{-1}(I - e^{-2t\theta})\\
&\quad - 2t\left(\frac{1}{\tau}\sigma_0 - \theta^{-1}\right) e^{-2t\theta} a\Bigg)\Bigg)^{-1}. \qquad \text{(A.16)}
\end{aligned}
$$

Using equation A.11,

$$
\begin{aligned}
\Sigma_t^{-1}(\theta+\epsilon a) &\approx \sigma_t^{-1} + \epsilon\tau\sigma_t^{-1}\Bigg(\theta^{-1} a\theta^{-1}(I - e^{-2t\theta})\\
&\quad - 2t\left(\theta^{-1} - \frac{1}{\tau}\sigma_0\right) e^{-2t\theta} a\Bigg)\sigma_t^{-1}. \qquad \text{(A.17)}
\end{aligned}
$$

Thus, $\dfrac{d}{d\epsilon}\Sigma_t^{-1}(\theta+\epsilon a)\ =\ldots$

Lemma 3.

$$\nabla_\theta \; x'\Sigma_t^{-1}(\theta)x = xx'c_t, \tag{A.19}$$

where c_t *is a p.d. matrix:*

$$c_t \stackrel{def}{=} \tau\sigma_t^{-2}\left(\theta^{-2}(I - e^{-2t\theta}) - 2t\left(\theta^{-1} - \frac{1}{\tau}\sigma_0\right)e^{-2t\theta}\right). \tag{A.20}$$

Proof. Using lemma 2 and considering the symmetry of the matrices at hand,

$$\begin{aligned}\frac{\partial}{\partial\theta_{ij}}\Sigma_t^{-1}(\theta) &= \frac{d}{d\epsilon}\Sigma_t^{-1}\left(\theta + \epsilon\frac{1_i1_j' + 1_j1_i'}{2}\right)\\ &= \frac{1_i1_j' + 1_j1_i'}{2}\tau\sigma_t^{-2}\left(\theta^{-2}(I - e^{-2t\theta}) - 2t\left(\theta^{-1} - \frac{1}{\tau}\sigma_0\right)e^{-2t\theta}\right)\\ &= \frac{1_i1_j' + 1_j1_i'}{2}c_t, \end{aligned}\tag{A.21}$$

where 1_i is a vector of Krönecker delta terms $1_i \stackrel{\text{def}}{=} (\delta_{1,i}, \ldots, \delta_{n,i})'$. Thus,

$$\frac{\partial}{\partial\theta_{ij}}x'\Sigma_t^{-1}(\theta)x = x'\frac{1_i1_j' + 1_j1_i'}{2}c_tx, \tag{A.22}$$

$$\nabla_\theta \; x'\Sigma_t^{-1}(\theta)x = xx'c_t. \tag{A.23}$$

We will now show that c_t is a p.d. matrix. First, note

$$c_t = 2t\sigma_t^{-2}\sigma_0e^{-2t\theta} + \tau\sigma_t^{-2}\theta^{-2}(I - e^{-2t\theta} - 2t\theta e^{-2t\theta}). \tag{A.24}$$

The first term is a p.d. matrix for $t > 0$. The second term has two factors: one is a p.d. matrix $\tau\sigma_t^{-2}\theta^{-2}$ and the other p.d.
for $t = 0$ and with a p.d. derivative wrt time:

$$\frac{d}{dt}((I - e^{-2t\theta}) - 2t\theta e^{-2t\theta}) = 4t\theta^2e^{-2t\theta}. \tag{A.25}$$

Lemma 4. *Let* $M_t : \mathcal{R}^n \times \mathcal{R}^n \to \mathcal{R}^n$

$$M_t(\theta) \stackrel{def}{=} e^{-t\theta}\mu_0 + (I - e^{-t\theta})\gamma. \tag{A.26}$$

Let $\theta \in \mathcal{R}^n \times \mathcal{R}^n$ *be a fixed symmetric invertible matrix, and let* $\mu_t \stackrel{def}{=} M_t(\theta)$*. Then*

$$\nabla_\theta\, x'\Sigma_t^{-1}(\theta)M_t(\theta) = x\mu_t'c_t + tx\sigma_t^{-1}e^{-t\theta}\,(\gamma - \mu_0)'\,. \tag{A.27}$$

with Σ_t, c_t *as defined in the previous lemmas.*

Proof.

$$\nabla_\theta\, x'\Sigma_t^{-1}(\theta)M_t(\theta) = \nabla_\theta\, x'\Sigma_t^{-1}(\theta)\mu_t + \nabla_\theta\, x'\sigma_t^{-1}M_t(\theta). \tag{A.28}$$

Using the proof for lemma 3, ==>

$$\nabla_\theta\, x'\Sigma_t^{-1}(\theta)\mu_t = x\mu_t'c_t. \tag{A.29}$$

Moreover, using standard matrix calculus rules (see Movellan, 2006a),

$$\begin{aligned}\nabla_\theta\, x'\sigma_t^{-1}M_t(\theta) &= \nabla_\theta\, x'\sigma_t^{-1}(e^{-t\theta}\mu_o + (I - e^{-t\theta})\gamma)\\ &= tx\sigma_t^{-1}e^{-t\theta}(\gamma - \mu_0)'.\end{aligned} \tag{A.30}$$

Theorem 1.

$$\nabla_\theta\left(x'\sigma_t^{-1}\mu_t - \frac{1}{2}x'\sigma_t^{-1}x\right) = x\mu_t'c_t + tx\sigma_t^{-1}e^{-t\theta}(\gamma - \mu_0)' - \frac{1}{2}xx'c_t. \tag{A.31}$$

Theorem 2. *Let* θ *be a random parameter vector and*
the $p(\xi \mid \theta)$ is constant wrt θ. $\Psi_\xi(u) = \nabla_\theta\phi(u, \theta) = 0$.
Let $\{X_t : t \in \mathcal{R}+\}$ *be a collection of rvs with distribution*

$$p_t(x_t \mid \theta) \propto e^{\phi_t(x,\theta)}.$$

Then

$$\nabla_\theta\,(\mathbb{D}(\xi, X_\infty)) - \mathbb{D}(X_t, X_\infty)) = -(H_t + \tilde{R}_t),$$

where H_t *is Hinton's CD statistic,*

$$H_t \stackrel{def}{=} \mathbb{E}[\Psi_\infty(\xi)] - \mathbb{E}[\Psi_\infty(X_t)],$$

and the residual $\tilde{R}_t$ *is a covariance statistic,*

$$\tilde{R}_t \stackrel{def}{=} \mathbb{C}ov[\phi_t(X_t) - \phi_\infty(X_t)\,,\; \Psi_t(X_t)]. \tag{A.43}$$