Zhirong Yang and Jorma Laaksonen. 2008. Principal whitened gradient for information geometry. Neural Networks, volume 21, numbers 2-3, pages 232-240.

Available online at www.sciencedirect.com

ScienceDirect

Neural Networks 21 (2008) 232–240

Neural Networks

www.elsevier.com/locate/neunet

2008 Special Issue

Principal whitened gradient for information geometry[☆]

Zhirong Yang[*], Jorma Laaksonen

Department of Information and Computer Science, Helsinki University of Technology, P.O. Box 5400, FI-02015 TKK, Espoo, Finland

Received 8 August 2007; received in revised form 30 November 2007; accepted 11 December 2007

Abstract

We propose two strategies to improve the optimization in information geometry. First, a local Euclidean embedding is identified by whitening the tangent space, which leads to an additive parameter update sequence that approximates the geodesic flow to the optimal density model. Second, removal of the minor components of gradients enhances the estimation of the Fisher information matrix and reduces the computational cost. We also prove that dimensionality reduction is necessary for learning multidimensional linear transformations. The optimization based on the principal whitened gradients demonstrates faster and more robust convergence in simulations on unsupervised learning with synthetic data and on discriminant analysis of breast cancer data.

Keywords: Information geometry; Natural gradient; Whitening; Principal components; Riemannian manifold

1. Introduction

Denote $\mathcal{J}$ an objective function to be minimized and $\boldsymbol{\theta}$ its parameters. The steepest descent update rule

$$\boldsymbol{\theta}^{\text{new}} = \boldsymbol{\theta} - \eta \nabla \mathcal{J}(\boldsymbol{\theta}) \tag{1}$$

with η a positive learning rate is widely used for minimization tasks because it is easy to implement. However, this update rule performs poorly in machine learning problems where the parameter space is not Euclidean. It was pointed out by Amari that the geometry of the Riemannian space must be taken into account when calculating the learning directions (Amari, 1998). He suggested the use of *natural gradient* (NAT) updates in place of the ordinary gradient-based ones:

$$\boldsymbol{\theta}^{\text{new}} = \boldsymbol{\theta} - \eta \mathbf{G}(\boldsymbol{\theta})^{-1} \nabla \mathcal{J}(\boldsymbol{\theta}), \tag{2}$$

where $\mathbf{G}(\boldsymbol{\theta})$ is the Riemannian metric matrix. Optimization that employs natural gradients generally requires much less iterations than the conventional steepest gradient descent (ascend) method.

Information geometry is another important concept proposed by Amari and Nagaoka (2000), where the Riemannian metric tensor is defined as the Fisher information matrix. The application of the natural gradient to information geometry leads to substantial performance gains in Blind Source Separation (Amari, 1998), multilayer perceptrons (Amari, 1998; Amari, Park, & Fukumizu, 2000), and other engineering problems that deal with statistical information. Nevertheless, many of these applications are restricted by the additive Gaussian noise assumption. Little attention has been paid on incorporating the specific properties of information geometry to facilitate general optimization.

We propose here to improve the natural gradient by a novel additive update rule called *Principal Whitened Gradient* (PWG):

$$\boldsymbol{\theta}^{\text{new}} = \boldsymbol{\theta} - \eta \hat{\mathbf{G}}(\boldsymbol{\theta})^{-\frac{1}{2}} \nabla \mathcal{J}(\boldsymbol{\theta}). \tag{3}$$

The square root and hat symbols indicate two strategies we use, both of which are based on a crucial observation that the Fisher information is the covariance of gradient vectors. First, we identify a local Euclidean embedding in the parameter space by whitening the tangent space at the current estimate of $\boldsymbol{\theta}$. The additive update sequence with whitened gradients results in a better approximation to the geodesic flow. The choice of learning rates also becomes easier in the Euclidean embedding. Second, the whitening procedure is accompanied

[☆] An abbreviated version of some portions of this article appeared in Yang and Laaksonen (2007) as part of the IJCNN 2007 Conference Proceedings, published under IEE copyright.

[*] Corresponding author. Tel.: +358 9 451 5281, +358 41 486 9473 (Mobile); fax: +358 9 451 3277.

E-mail addresses: zhirong.yang@tkk.fi (Z. Yang), jorma.laaksonen@tkk.fi (J. Laaksonen).

0893-6080/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2007.12.016

with removal of minor components for computational efficiency and for better estimation of the Fisher information with finite data. We also prove that dimensionality reduction is necessary for learning a great variety of multidimensional linear transformations.

We demonstrate the advantage of PWG over NAT by three simulations, two for unsupervised learning and one for supervised. The first task in our experiments is to learn the variance of a multivariate Gaussian distribution. The second is to recover the component means of a Gaussian mixture model by the maximum likelihood method. The last one is to learn a matrix that maximizes discrimination of labeled breast cancer data. In all simulations, the updates with principal whitened gradients outperform the original natural gradient results in terms of efficiency and robustness.

The above innovation was preliminarily proposed in Yang and Laaksonen (2007) by the same authors. In this paper we include the following new contributions:

- We propose to interpret the Fisher information metric as the local change of the Kullback–Leibler divergence. A formal derivation based on Taylor series is also provided.
- We point out that the Fisher information matrix is the covariance of ordinary online gradients and this property is invariant of linear transformations. This result provides a new justification for the employed prewhitening procedure.
- We clarify the motivation for discarding the minor components as the principal ones correspond to the directions where the local information change is maximized.
- We use three different simulations. A new kind of simulation is added to demonstrate the learning of variance. The learning on Gaussian mixtures is extended to a two-dimensional case. We also use another data set for the discriminative learning experiment.

The remaining part of the paper is organized as follows. We first provide a brief of the natural gradient and information geometry, as well as the concept of geodesic updates in Section 2. In Section 3, we present two strategies to improve the optimization in information geometry. Next we demonstrate the performance of the proposed method with simulations on learning a Gaussian mixture model and discriminating breast cancer data in Section 4. Finally the conclusions are drawn in Section 5.

2. Background

2.1. Natural gradient and information geometry

A Riemannian metric is a generalization of the Euclidean one. In an m-dimensional Riemannian manifold, the inner product of two tangent vectors $\mathbf{u}$ and $\mathbf{v}$, at point $\boldsymbol{\theta}$, is defined as

$$\langle \mathbf{u}, \mathbf{v} \rangle_{\boldsymbol{\theta}} \equiv \sum_{i=1}^{m} \sum_{j=1}^{m} [\mathbf{G}(\boldsymbol{\theta})]_{ij} u_i v_j = \mathbf{u}^{\mathrm{T}} \mathbf{G}(\boldsymbol{\theta}) \mathbf{v}, \tag{4}$$

where $\mathbf{G}(\boldsymbol{\theta})$ is a positive definite matrix. A Riemannian metric reduces to Euclidean when $\mathbf{G}(\boldsymbol{\theta}) = \mathbf{I}$, the identity matrix.

The *natural gradient* (NAT) method proposed by Amari (1998) addresses the optimization in a Riemannian manifold. It mimics the steepest descent approach in the Euclidean space by seeking a vector $\mathbf{a}$ that minimizes the linear approximation of the updated objective (Amari, 1998)

$$\mathcal{J}(\boldsymbol{\theta}) + \nabla \mathcal{J}(\boldsymbol{\theta})^{\mathrm{T}} \mathbf{a}. \tag{5}$$

Natural gradient employs the constraint $\langle \mathbf{a}, \mathbf{a} \rangle_{\boldsymbol{\theta}} = \epsilon$ instead of the Euclidean norm, where ϵ is an infinitesimal constant. By introducing a Lagrange multiplier λ and setting the gradient with respect to $\mathbf{a}$ to zero, one obtains

$$\mathbf{a} = \frac{1}{2\lambda} \mathbf{G}(\boldsymbol{\theta})^{-1} \nabla \mathcal{J}(\boldsymbol{\theta}). \tag{6}$$

Observing $\mathbf{a}$ is proportional to $\mathbf{G}(\boldsymbol{\theta})^{-1} \nabla \mathcal{J}(\boldsymbol{\theta})$, Amari (1998) defined the *natural gradient*

$$\nabla_{\mathrm{NAT}} \mathcal{J}(\boldsymbol{\theta}) \equiv \mathbf{G}(\boldsymbol{\theta})^{-1} \nabla \mathcal{J}(\boldsymbol{\theta}) \tag{7}$$

and suggested the update rule

$$\boldsymbol{\theta}^{\mathrm{new}} = \boldsymbol{\theta} - \eta \mathbf{G}(\boldsymbol{\theta})^{-1} \nabla \mathcal{J}(\boldsymbol{\theta}) \tag{8}$$

with η a positive learning rate.

Many statistical inference problems can be reduced to probability density estimation. *Information geometry* proposed by Amari and Nagaoka (2000) studies a manifold of parametric probability densities $p(\mathbf{x}; \boldsymbol{\theta})$, where the Riemannian metric is defined as the Fisher information matrix

$$[\mathbf{G}(\boldsymbol{\theta})]_{ij} = E\left\{ \frac{\partial \ell(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \ell(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_j} \right\}. \tag{9}$$

Here $\ell(\mathbf{x}; \boldsymbol{\theta}) \equiv -\log p(\mathbf{x}; \boldsymbol{\theta})$. Amari also applied the natural gradient update rule for the optimization in the information geometry by using $\mathcal{J}(\boldsymbol{\theta}) = \ell(\mathbf{x}; \boldsymbol{\theta})$ as the online objective function, which is equivalent to the maximum likelihood approach (Amari, 1998). Similarly, the batch objective function can be defined to be the empirical mean of $\ell(\mathbf{x}; \boldsymbol{\theta})$ over $\mathbf{x}$.

It is worth to note that the natural gradient method has been applied to other Riemannian manifolds. For example, Amari (1998) proposed to use the natural gradient on the manifold of invertible matrices for the blind source separation problem. However, this is beyond the scope of information geometry. In this paper we only focus on the improvement in the manifolds defined by the Fisher information metric.

2.2. Geodesic updates

Geodesics in a Riemannian manifold generalize the concept of line segments in the Euclidean space. Given a Riemannian metric $\mathbf{G} \equiv \mathbf{G}(\boldsymbol{\theta})$ and $t \in \mathbb{R}$, a curve $\boldsymbol{\theta} \equiv \boldsymbol{\theta}(t)$ is a *geodesic* if and only if (Peterson, 1998)

$$\frac{\mathrm{d}^2\theta_k}{\mathrm{d}t^2} + \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{\mathrm{d}\theta_j}{\mathrm{d}t} \frac{\mathrm{d}\theta_i}{\mathrm{d}t} \Gamma_{ij}^{k} = 0, \quad \forall k = 1, \ldots, m, \tag{10}$$

where

$$\Gamma_{ij}^{k} \equiv \frac{1}{2} \sum_{l=1}^{m} \left(\mathbf{G}^{-1}\right)_{kl} \left(\frac{\partial G_{il}}{\partial \theta_j} + \frac{\partial G_{lj}}{\partial \theta_i} - \frac{\partial G_{ij}}{\partial \theta_l} \right) \tag{11}$$

are Riemannian connection coefficients. The geodesic with a starting point $\boldsymbol{\theta}(0)$ and a tangent vector $\mathbf{v}$ is denoted by $\boldsymbol{\theta}(t; \boldsymbol{\theta}(0), \mathbf{v})$. The *exponential map* of the starting point is then defined as

$$\exp_{\mathbf{v}}(\boldsymbol{\theta}(0)) \equiv \boldsymbol{\theta}(1; \boldsymbol{\theta}(0), \mathbf{v}). \tag{12}$$

It can be shown that the length along the geodesic between $\boldsymbol{\theta}(0)$ and $\exp_{\mathbf{v}}(\boldsymbol{\theta}(0))$ is $|\mathbf{v}|$ (Peterson, 1998).

The above concepts are appealing because a geodesic connects two points in the Riemannian manifold with the minimum length. Iterative application of exponential maps therefore forms an approximation of flows along the geodesic and the optimization can converge quickly.

Generally obtaining the exponential map (12) is not a trivial task. In most cases, the partial derivatives of the Riemannian tensor in (11) lead to rather complicated expression of Γ_{ij}^k and solving the differential equations (10) therefore becomes computationally infeasible. A simplified form of exponential maps without explicit matrix inversion can be obtained only in some special cases, for instance, when training multilayer perceptrons based on additive Gaussian noise assumptions (Amari, 1998; Amari et al., 2000). When the Riemannian metric is accompanied with left- and right-translation invariance, the exponential maps coincide with the ones used in Lie Group theory and can be accomplished by matrix exponentials (see e.g. Nishimori and Akaho (2005)). This property however does not hold for information geometry.

3. Principal whitened gradient

3.1. Whitening the gradient space

Most statistical inference techniques are based on the assumption that the data are independently and identically distributed (i.i.d.). The online gradients calculated by using individual data can be viewed as i.i.d. random variables. For clarity, we use the concise symbol $\nabla \equiv \nabla \mathcal{J}(\boldsymbol{\theta})$ in what follows. It is a crucial observation that the Fisher information matrix in (9) coincides with the covariance of these gradients. That is, the squared norm of a tangent vector $\mathbf{u}$ is given by

$$\|\mathbf{u}\|_{\mathbf{G}}^2 \equiv \mathbf{u}^{\mathrm{T}}\mathbf{G}\mathbf{u} = \mathbf{u}^{\mathrm{T}} E\{\nabla\nabla^{\mathrm{T}}\}\mathbf{u}. \tag{13}$$

The following theorem justifies that such a squared norm is proportional to the local information change along $\mathbf{u}$.

Theorem 1. *Let* $\mathbf{u}$ *be a tangent vector at* $\boldsymbol{\theta}$. *Denote* $p \equiv p(\mathbf{x}; \boldsymbol{\theta})$ *and* $p^t \equiv p(\mathbf{x}; \boldsymbol{\theta} + t\mathbf{u})$. *For information geometry,*

$$\|\mathbf{u}\|_{\mathbf{G}} \equiv \sqrt{\mathbf{u}^{\mathrm{T}}\mathbf{G}\mathbf{u}} = \sqrt{2} \lim_{t\to 0} \frac{\sqrt{D_{\mathrm{KL}}(p; p^t)}}{t}, \tag{14}$$

where D_{KL} is the Kullback–Leibler divergence. The proof can be found in Appendix A.

Now let us consider a linear transformation matrix $\mathbf{F}$ in the tangent space. From the derivation in Appendix A, it can be seen that the first term in the Taylor expansion always vanishes, and the second remains zero under linear transformations. When connecting to the norm, the terms of third or higher order can always be neglected if t is small enough. Thus the local information change in the transformed space is only measured by the second-order term.

Another crucial observation is that the Hessian $\mathbf{H}$ of the Kullback–Leibler divergence can be expressed in the form of gradient vectors instead of the second order derivative. Thus we have the following corollary:

Corollary 1. *Given a linear transformation* $\tilde{\mathbf{u}} = \mathbf{F}\mathbf{u}$, *the Hessian in the transformed space*

$$\tilde{\mathbf{H}} = \tilde{\mathbf{G}} \equiv E\{\tilde{\nabla}\tilde{\nabla}^{\mathrm{T}}\} = \mathbf{F}E\{\nabla\nabla^{\mathrm{T}}\}\mathbf{F}^{\mathrm{T}}, \tag{15}$$

and

$$\sqrt{2} \lim_{t\to 0} \frac{\sqrt{D_{\mathrm{KL}}(p, \tilde{p}^t)}}{t} = \sqrt{\tilde{\mathbf{u}} E\{\tilde{\nabla}\tilde{\nabla}^{\mathrm{T}}\}\tilde{\mathbf{u}}} = \|\tilde{\mathbf{u}}\|_{\tilde{\mathbf{G}}}, \tag{16}$$

where $\tilde{p}^t \equiv p(\mathbf{x}; \boldsymbol{\theta} + t\tilde{\mathbf{u}})$.

This allows us to look for a proper $\mathbf{F}$ for facilitating the optimization.

Since the Fisher information matrix $\mathbf{G}$ is positive semi-definite, one can always decompose such a matrix as $\mathbf{G} = \mathbf{E}\mathbf{D}\mathbf{E}^{\mathrm{T}}$ by a singular value decomposition, where $\mathbf{D}$ is a diagonal matrix and $\mathbf{E}^{\mathrm{T}}\mathbf{E} = \mathbf{I}$. Denote the *whitening matrix* $\mathbf{G}^{-\frac{1}{2}}$ for the gradient as

$$\mathbf{G}^{-\frac{1}{2}} = \mathbf{E}\mathbf{D}^{-\frac{1}{2}}\mathbf{E}^{\mathrm{T}}, \tag{17}$$

where

$$\left[\mathbf{D}^{-\frac{1}{2}}\right]_{ii} = \begin{cases} 1/\sqrt{D_{ii}} & \text{if } D_{ii} > 0 \\ 0 & \text{if } D_{ii} = 0. \end{cases} \tag{18}$$

With $\mathbf{F} = \mathbf{G}^{-\frac{1}{2}}$, the transformed Fisher information matrix becomes

$$E\{\tilde{\nabla}\tilde{\nabla}^{\mathrm{T}}\} = \mathbf{G}^{-\frac{1}{2}} E\{\nabla\nabla^{\mathrm{T}}\}\mathbf{G}^{-\frac{1}{2}} = \mathbf{I}. \tag{19}$$

That is, the whitening matrix locally transforms the Riemannian tangent space into its Euclidean embedding:

$$\|\tilde{\mathbf{u}}\|_{\tilde{\mathbf{G}}}^2 = \tilde{\mathbf{u}}^{\mathrm{T}} E\{\tilde{\nabla}\tilde{\nabla}^{\mathrm{T}}\}\tilde{\mathbf{u}} = \tilde{\mathbf{u}}^{\mathrm{T}}\tilde{\mathbf{u}} = \|\tilde{\mathbf{u}}\|^2. \tag{20}$$

The optimization problem in such an embedding then becomes seeking a tangent vector $\mathbf{a}$ that minimizes

$$\mathcal{J}(\boldsymbol{\theta}) + \left[\tilde{\nabla}\mathcal{J}(\boldsymbol{\theta})\right]^{\mathrm{T}} \mathbf{a} \tag{21}$$

under the constraint $\|\mathbf{a}\|^2 = \epsilon$. Again by using the Lagrangian method, one can obtain the solution

$$\mathbf{a} = -\frac{1}{2\lambda}\tilde{\nabla}\mathcal{J}(\boldsymbol{\theta}), \tag{22}$$

which leads to an ordinary steepest descent update rule in the whitened tangent space:

$$\boldsymbol{\theta}^{\mathrm{new}} = \boldsymbol{\theta} - \eta\tilde{\nabla}\mathcal{J}(\boldsymbol{\theta}). \tag{23}$$

By substituting $\tilde{\nabla} = \mathbf{G}^{-\frac{1}{2}}\nabla$, we get the *whitened gradient* update rule:

$$\boldsymbol{\theta}^{\text{new}} = \boldsymbol{\theta} - \eta\mathbf{G}(\boldsymbol{\theta})^{-\frac{1}{2}}\nabla\mathcal{J}(\boldsymbol{\theta}). \tag{24}$$

The update rule (24) has a form similar to the natural gradient one (2) except the square root operation on the eigenvalues of the Fisher information matrix at each iteration. We address that such a difference brings two distinguished advantages.

First, it is worth to notice that additive updates are equivalent to exponential maps when the Riemannian metric becomes Euclidean. In other words, the steepest descent updates with whitened gradients form approximated exponential maps and hence result in fast convergence. By contrast, the natural gradient updates (2) have no concrete geometric meaning and the updating sequence may therefore be far from the true geodesic. Moreover, our proposed objective (21) is the linear approximation of $\mathcal{J}(\boldsymbol{\theta} + t\mathbf{a})$, which is naturally defined in the Euclidean space. On the contrary, the underlying approximation (5) could perform poorly in a highly curved Riemannian manifold.

Second, we can see that the learning rate η in (2), or the corresponding Lagrange multiplier λ in (6), is a variable that depends not only on $\boldsymbol{\theta}$ and $\mathcal{J}$, but also on $\mathbf{G}(\boldsymbol{\theta})$. This complicates the choice of a proper learning rate. By contrast, selecting the learning rate for the whitened gradient updates is as easy as for the ordinary steepest descent approach because η in (24) is independent of the local Riemannian metric.

3.2. *Principal components of whitened gradient*

The singular value decomposition used in (17) is tightly connected to *Principal Component Analysis* (PCA) which is usually accompanied with dimensionality reduction for the following two reasons.

First, the Fisher information matrix is commonly estimated by the scatter matrix with $n < \infty$ samples $\mathbf{x}^{(i)}$:

$$\mathbf{G} \approx \frac{1}{n}\sum_{i=1}^{n}\nabla_{\boldsymbol{\theta}}\ell(\mathbf{x}^{(i)};\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}\ell(\mathbf{x}^{(i)};\boldsymbol{\theta})^{\mathrm{T}}. \tag{25}$$

However, the estimation accuracy could be poor because of sparse data in high-dimensional learning problems. Principal Component Analysis is a widely-applied method for reducing such kind of artifacts by reconstructing a positive semi-definite matrix from its low-dimensional linear embedding. In our case, the principal direction $\mathbf{w}$ maximizes the variance of the projected gradient:

$$\mathbf{w} = \arg\max_{\|\mathbf{u}\|=1} E\{(\mathbf{u}^{\mathrm{T}}\nabla)^2\}. \tag{26}$$

According to Theorem 1, $\mathbf{w}$ coincides with the direction that maximizes local information change:

$$E\{(\mathbf{u}^{\mathrm{T}}\nabla)^2\} = \mathbf{u}^{\mathrm{T}}E\{\nabla\nabla^{\mathrm{T}}\}\mathbf{u} = 2\left(\lim_{t\to 0}\frac{\sqrt{D_{\mathrm{KL}}(p;p^t)}}{t}\right)^2, \tag{27}$$

where $p \equiv p\left(\{\mathbf{x}^{(i)}\};\boldsymbol{\theta}\right)$, $p^t \equiv p\left(\{\mathbf{x}^{(i)}\};\boldsymbol{\theta} + t\mathbf{u}\right)$, and $\boldsymbol{\theta}$ is the current estimate. By this motivation, we preserve only the principal components and discard the minor ones that are probably irrelevant for the true learning direction. Moreover, the singular value decomposition (17) runs much faster than inverting the whole matrix $\mathbf{G}$ when the number of principal components is far less than the dimensionality of $\boldsymbol{\theta}$ (Golub & van Loan, 1989).

Second, dimensionality reduction is sometimes motivated by structural reasons. Consider a problem where one tries to minimize an objective of the form

$$\mathcal{J}(\mathbf{x};\mathbf{W}) = \mathcal{J}\left(\|\mathbf{W}^{\mathrm{T}}\mathbf{x}\|^2\right), \tag{28}$$

where $\mathbf{W}$ is an $m \times r$ matrix. Many objectives where Gaussian basis functions are used and evaluated in the linear transformed space can be reduced to this form. Examples can be found in Neighborhood Component Analysis (Goldberger, Roweis, Hinton, & Salakhutdinov, 2005) and in a variant of Independent Component Analysis (Hyvärinen & Hoyer, 2000). The following theorem justifies the necessity of dimensionality reduction for these kinds of problems.

Denote $\xi_i = \|\mathbf{W}^{\mathrm{T}}\mathbf{x}^{(i)}\|^2$ in (28) and $f_i = 2\partial\mathcal{J}(\xi_i)/\partial\xi_i$. The gradient of $\mathcal{J}(\mathbf{x}^{(i)};\mathbf{W})$ with respect to $\mathbf{W}$ is denoted by

$$\nabla^{(i)} \equiv \nabla_{\mathbf{W}}\mathcal{J}(\mathbf{x}^{(i)};\mathbf{W}) = f_i\mathbf{x}^{(i)}\left(\mathbf{x}^{(i)}\right)^{\mathrm{T}}\mathbf{W}. \tag{29}$$

The $m \times r$ matrix $\nabla^{(i)}$ can be represented as a row vector $\left(\boldsymbol{\psi}^{(i)}\right)^{\mathrm{T}}$ by concatenating the columns such that

$$\psi^{(i)}_{k+(l-1)m} = \nabla^{(i)}_{kl}, \quad k = 1,\ldots,m, l = 1,\ldots,r. \tag{30}$$

Piling up the rows $\left(\boldsymbol{\psi}^{(i)}\right)^{\mathrm{T}}$, $i = 1,\ldots,n$, yields an $n \times mr$ matrix $\boldsymbol{\Psi}$.

Theorem 2. *Suppose $m > r$. For any positive integer n, the column rank of $\boldsymbol{\Psi}$ is at most $mr - r(r-1)/2$.*

The proof can be found in Appendix B. In other words, no matter how many samples are available, the matrix $\boldsymbol{\Psi}$ is not full rank when $r > 1$, and neither is the approximated Fisher information matrix

$$\mathbf{G} \approx \frac{1}{n}\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi}. \tag{31}$$

That is, $\mathbf{G}$ is always singular for learning multidimensional linear transformations in (28). $\mathbf{G}^{-1}$ hence does not exist and one must employ dimensionality reduction before inverting the matrix.

Suppose $\hat{\mathbf{D}}$ is a diagonal matrix with the q largest eigenvalues of $\mathbf{G}$, and the corresponding eigenvectors form the columns of matrix $\hat{\mathbf{E}}$. The geodesic update rule with *Principal Whitened Gradient* (PWG) then becomes

$$\boldsymbol{\theta}^{\text{new}} = \boldsymbol{\theta} - \eta\hat{\mathbf{G}}^{-\frac{1}{2}}\boldsymbol{\psi}^{(i)}, \tag{32}$$

where $\hat{\mathbf{G}} = \hat{\mathbf{E}}\hat{\mathbf{D}}\hat{\mathbf{E}}^{\mathrm{T}}$. For learning problems (28), the new $\mathbf{W}$ is then obtained by reshaping $\boldsymbol{\theta}$ into an $m \times r$ matrix. The online gradient $\boldsymbol{\psi}^{(i)}$ in (32) can also be replaced by a batch gradient $E\{\boldsymbol{\psi}\}$.

4. Simulations

4.1. Learning of variances

First we applied the NAT update rule (2) and our PWG method (3) to the estimation of a 30-dimensional zero-mean Gaussian distribution

$$p(\mathbf{x};\mathbf{w}) = \frac{\prod_{d=1}^{30}|w_d|}{(2\pi)^{30/2}} \exp\left(-\frac{1}{2}\sum_{d=1}^{30}(w_d x_d)^2\right), \tag{33}$$

where the parameter w_d is the inverted variance of the dth dimension. Although there exist methods to obtain the estimate in a closed form, this simulation can serve as an illustrative example for comparing NAT and PWG.

We drew $n = 1,000$ samples from the above distribution. The negative log-likelihood of the ith sample is

$$\ell(\mathbf{x}^{(i)};\mathbf{w}) \equiv -\sum_{d=1}^{30}\log|w_d| + \frac{1}{2}\sum_{d=1}^{30}\left(w_d x_d^{(i)}\right)^2 + \frac{30}{2}\log 2\pi, \tag{34}$$

of which the partial derivative with respect to w_k is

$$\frac{\partial\ell(\mathbf{x}^{(i)};\mathbf{w})}{\partial w_k} = \left(w_k x_k^{(i)}\right)^2 - \frac{1}{w_k}. \tag{35}$$

With these online gradients, we approximated the Fisher information matrix by (25) and calculated the batch gradient

$$\nabla \equiv \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\ell(\mathbf{x}^{(i)};\mathbf{w})}{\partial\mathbf{w}}. \tag{36}$$

We first set the learning rate $\eta = 1$ in natural gradient updates. The whitening objective values in the iterative optimization are shown as the dot-dashed curve in Fig. 1. The learning proceeds slowly and after 1,000 iterations the objective is still very far from the optimal one. Increasing the learning to $\eta = 10$ can slightly speed up the optimization, but a too large learning rate, e.g. $\eta = 50$, leads to a jagged objective evolution. Although the objective keeps decreasing in the long run, the resulting variance estimates are no better than the ones obtained with $\eta = 10$.

By contrast, the optimization using PWG is much faster and more robust. The objective significantly decreases in the first two hundred iterations and smoothly converges to a nearly perfect value 43,814. For better comparison, we also plot the optimal objective 40,776, calculated with the true variance, as the dashed line in Fig. 1.

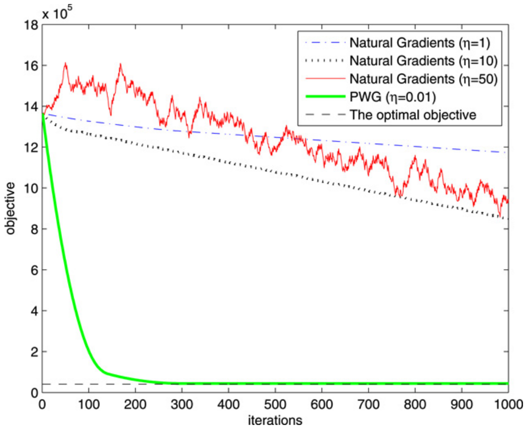

Fig. 1. Learning the variances of a multivariate Gaussian distribution.

4.2. Gaussian Mixtures

Next we tested the PWG method (3) and the NAT update rule (2) on synthetic data that were generated by a *Gaussian Mixture Model* (GMM) of ten two-dimensional normal distributions $\mathcal{N}(\boldsymbol{\mu}_*^{(k)},\mathbf{I})$, $k = 1,\ldots,10$. Here $\boldsymbol{\mu}_*^{(k)}$ were randomly chosen within $(0,1)\times(0,1)$. We drew 100 two-dimensional points from each Gaussian and obtained 1,000 samples in total. The true $\{\boldsymbol{\mu}_*^{(k)}\}$ values were unknown to the compared learning methods, and the learning task was to recover these means with the estimates $\{\boldsymbol{\mu}^{(k)}\}$ randomly initialized. We used the maximum likelihood objective, or equivalently the minimum negative log-likelihood

$$\mathcal{J}_{\text{GMM}}(\{\boldsymbol{\mu}^{(j)}\}) \equiv \sum_{i=1}^{1000}\ell(\mathbf{x}^{(i)};\{\boldsymbol{\mu}^{(j)}\}) = -\sum_{i=1}^{1000}\log\sum_{j=1}^{10}\exp\left(-\frac{\|\mathbf{x}^{(i)}-\boldsymbol{\mu}^{(j)}\|^2}{2}\right) + C, \tag{37}$$

where $C = 1000\log(2\pi)^{10/2}$ is a constant. We then computed the partial derivatives with respect to each mean:

$$\frac{\partial\ell(\mathbf{x}^{(i)};\{\boldsymbol{\mu}^{(j)}\})}{\partial\boldsymbol{\mu}^{(k)}} = -\sum_{i=1}^{1000}\frac{\left(\mathbf{x}^{(i)}-\boldsymbol{\mu}^{(k)}\right)\exp\left(-\frac{1}{2}\|\mathbf{x}^{(i)}-\boldsymbol{\mu}^{(k)}\|^2\right)}{\sum_{j=1}^{10}\exp\left(-\frac{1}{2}\|\mathbf{x}^{(i)}-\boldsymbol{\mu}^{(j)}\|^2\right)}. \tag{38}$$

Similarly, the batch gradient was the empirical mean of the online ones over i, and the Fisher information matrix was approximated by (25).

The dot-dashed line in Fig. 2 shows the evolution of the objective when using natural gradient (2) with the learning rate $\eta = 0.001$. Although the negative likelihood is decreasing most of the time, the optimization is severely hindered by some unexpected upward jumps. We found

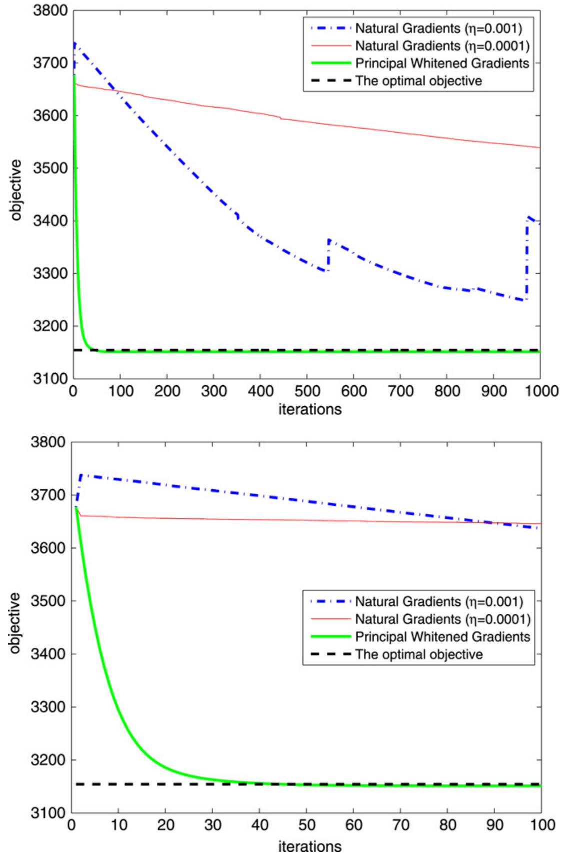

Fig. 2. Learning a GMM model by natural gradient and principal whitened gradient in 1000 rounds (top) and in the first 100 rounds (bottom).

17 rises in the curve, among which the most significant four take place in the 1st, 546th, 856th and 971st rounds. Consequently the objective cannot converge to a satisfactory result.

One may think that the $\eta = 0.001$ could be too large for natural gradient updates and turn to a smaller one, e.g. $\eta = 10^{-4}$. The training result is shown as the thin solid curve in Fig. 2, from which we can see that the objective decreases gradually but slowly. Although the unexpected rises are avoided, the optimization speed is sacrificed due to the excessively small learning rate.

The objectives by using principal whitened gradient are also shown in Fig. 2 (thick solid curve). Here we set $\eta = 0.01$ and seven principal components of the whitened gradient are used in (3). It can be seen that $\mathcal{J}_{\text{GMM}}$ decreases steadily and efficiently. Within 18 iterations, the loss becomes less than 3,200, which is much better than the loss levels obtained by the natural gradient updates. The final objective achieved by PWG updates is 3,151.40—a value very close to the global optimum 3,154.44 computed by using the true Gaussian means $\{\boldsymbol{\mu}_*^{(k)}\}$.

4.3. Wisconsin Diagnostic Breast Cancer Data

We then applied the compared methods on the real Wisconsin Diagnostic Breast Cancer (WDBC) dataset which is available in UCI Repository of machine learning databases (Newman, Hettich, Blake, & Merz, 1998). The dataset consists of $n = 569$ instances, each of which has 30 real numeric attributes. 357 samples are labeled as *benign* and the other 212 as *malignant*.

Denote $\{(\mathbf{x}^{(i)}, c_i)\}$, $\mathbf{x}^{(i)} \in \mathbb{R}^{30}$, $c_i \in \{\text{benign,malignant}\}$ for the labeled data pairs. We define here the *Unconstrained Informative Discriminant Analysis* (UIDA) which seeks a linear transformation matrix $\mathbf{W}$ of size 30×2 such that the negative discrimination

$$\mathcal{J}_{\text{UIDA}}(\mathbf{W}) \equiv -\sum_{i=1}^{n} \log p(c_i|\mathbf{y}^{(i)}) \tag{39}$$

is minimized in the transformed space where $\mathbf{y}^{(i)} = \mathbf{W}^{\mathrm{T}}\mathbf{x}^{(i)}$. Here the predictive density $p(c_i|\mathbf{y}^{(i)})$ is estimated by the Parzen window method:

$$p(c_i|\mathbf{y}^{(i)}) \propto \frac{\sum\limits_{j=1}^{n} \phi_{ij} e_{ij}}{\sum\limits_{j=1}^{n} e_{ij}}, \tag{40}$$

where

$$e_{ij} \equiv \begin{cases} \exp\left(-\dfrac{1}{2}\|\mathbf{y}^{(i)} - \mathbf{y}^{(j)}\|^2\right) & i \neq j \\ 0 & i = j \end{cases}, \tag{41}$$

and $\phi_{ij} = 1$ if $c_i = c_j$ and 0 otherwise.

UIDA is a variant of the *Informative Discriminant Analysis* (IDA) method which was first discussed in Peltonen and Kaski (2005). IDA extracts the discriminative components of data for visualization and recognition. The original IDA learning requires additional steps to select a proper Parzen window width, and the transformation matrix $\mathbf{W}$ is constrained to be orthonormal. In this illustrative example we aim at demonstrating the advantage of PWG over NAT updates in convergence. We hence remove the orthonormality constraint for simplicity. By this relaxation, the selection of the Parzen window width is absorbed into the learning of the transformation matrix.

PWG is suitable for the UIDA learning since UIDA's objective has the form (28). Let k denote a class index. First we compute the individual gradients

$$\nabla^{(i)} = \text{vectorize}\left(\frac{\sum\limits_{j=1}^{n} e_{ij}\mathbf{B}^{(ij)}}{\sum\limits_{j=1}^{n} e_{ij}}\right), \tag{42}$$

$$\nabla_k^{(i)} = \text{vectorize}\left(\frac{\sum\limits_{j:c_j=k} e_{ij}\mathbf{B}^{(ij)}}{\sum\limits_{j:c_j=k} e_{ij}}\right), \tag{43}$$

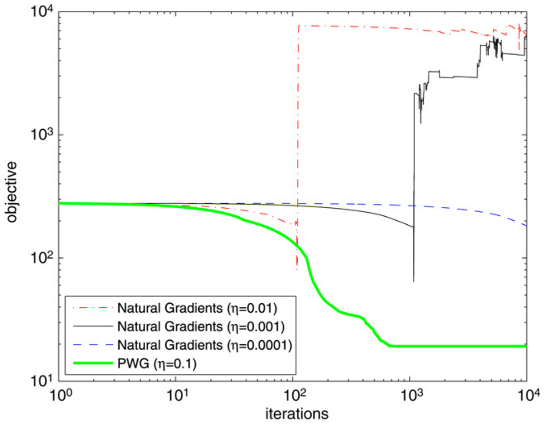

Fig. 3. Objective curves of discriminating WDBC data by using NAT and PWG.

where $\mathbf{B}^{(ij)} = \left(\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\right)\left(\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\right)^{\mathrm{T}} \mathbf{W}$. The batch gradient is

$$\tilde{\nabla} = \frac{1}{n}\sum_{i=1}^{n}\left(\nabla^{(i)} - \sum_{k} p(k)\nabla_k^{(i)}\right). \tag{44}$$

By applying Taylor expansion in analog to the derivation in Appendix A, we can obtain

$$\begin{aligned}&\mathcal{J}_{\mathrm{UIDA}}(\mathbf{W} + t\tilde{\nabla}) - \mathcal{J}_{\mathrm{UIDA}}(\mathbf{W}) \\ &\quad = \frac{1}{2}\left(t\tilde{\nabla}\right)^{\mathrm{T}}\left[\mathbf{G} - \sum_k p(k)\mathbf{G}_k\right]\left(t\tilde{\nabla}\right) + o(t^2),\end{aligned} \tag{45}$$

where

$$\mathbf{G} = \sum_{i=1}^{n}\nabla^{(i)}\left(\nabla^{(i)}\right)^{\mathrm{T}}, \tag{46}$$

$$\mathbf{G}_k = \sum_{i:c_i=k}\nabla_k^{(i)}\left(\nabla_k^{(i)}\right)^{\mathrm{T}}. \tag{47}$$

Then the learning direction is $\tilde{\nabla}$ left-multiplied by the principal whitened components of $\mathbf{G} - \sum_k p(k)\mathbf{G}_k$.

We tried three different learning rates in the natural gradient algorithm (2). The results in log-log scale are shown in Fig. 3. With $\eta = 0.01$, the natural gradient updates seem to work well in the first 110 iterations, but after the 111st iteration $\mathcal{J}_{\mathrm{UIDA}}$ soars up to the value 2004.7, which is much worse than the initial value. In the 112nd iteration, the loss even increases to 7012.9, and the subsequent natural gradient learning never returns an objective less than 4800. Such unexpected jumps cannot be avoided either by using $\eta = 0.001$, as an upward jump still occurs after the 1096th round. With an excessively small learning rate $\eta = 0.0001$, although there are no upward jumps, the learning speed becomes extremely slow and the UIDA objective is no less than 180 after 10,000 iterations.

By contrast, the proposed PWG method (3) demonstrates both efficiency and robustness in this learning task. From Fig. 3, it can be seen that the negative discrimination keeps decreasing by using PWG with $\eta = 0.1$. We obtained $\mathcal{J}_{\mathrm{UIDA}} < 200$ after 40 iterations, and the best objective achieved by using PWG is 19.24 after 10,000 iterations.

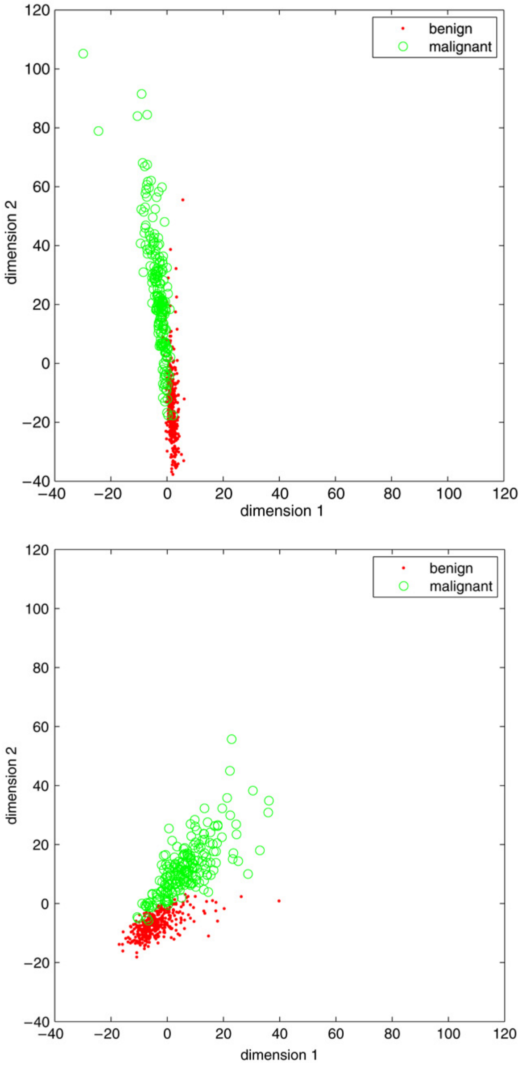

Fig. 4. WDBC data in the projected space. Top: 500 NAT updates with $\eta = 0.001$. Bottom: 500 PWG updates with $\eta = 0.1$.

The projected data are displayed in Fig. 4, where we used $\eta = 0.001$ for NAT and $\eta = 0.1$ for PWG. The results are examined after 500 iterations of both methods. We can see that the *benign* and *malignant* samples mainly distribute along a narrow line and are heavily mixed after the NAT updates, whereas the two classes are quite well separated by using the PWG method. The corresponding objective value is 229.04 by using NAT and 24.89 by PWG.

5. Conclusions

We have presented two strategies to improve natural gradient updates in information geometry. Whitening the tangent vectors with respect to the Fisher information matrix transforms the Riemannian space to be locally Euclidean. The resulting additive update sequence approximates the geodesic flow towards the optimal solution well. Calculating the learning direction with only principal components of the whitened gradients further enhances both efficiency and robustness of the optimization. We have also pointed out that dimensionality reduction is indispensable for learning multidimensional linear transformations. The proposed method has been validated by simulations in both unsupervised and supervised learning.

There exist several directions to extend the principal whitened gradient. In this paper we computed the whitening matrix and the batch gradient separately. Actually one may achieve a direct and faster method for computing their product, for example, by adopting online Principal Component Analysis (Yang, 1995). Another potential extension of the PWG update rule is to make it to accommodate additional constraints such as orthonormality or sparseness. Moreover, many conventional optimization techniques, such as the conjugate gradient, can be applied in the Euclidean embedding to further improve the convergence speed.

Acknowledgments

This work is supported by the Academy of Finland in the projects *Neural methods in information retrieval based on automatic content analysis and relevance feedback* and *Finnish Centre of Excellence in Adaptive Informatics Research.*

Appendix A. Proof of Theorem 1

The proof follows the second order Taylor expansion of the Kullback–Leibler divergence around $\boldsymbol{\theta}$:

$$\begin{aligned} D_{\mathrm{KL}}(p;p^t) &\equiv \int p(\mathbf{x};\boldsymbol{\theta}) \log \frac{p(\mathbf{x};\boldsymbol{\theta})}{p(\mathbf{x};\boldsymbol{\theta}+t\mathbf{u})} \mathrm{d}\mathbf{x} \\ &= D_{\mathrm{KL}}(p;p) + (t\mathbf{u})^{\mathrm{T}} \mathbf{g}(\boldsymbol{\theta}) \\ &\quad + \frac{1}{2}(t\mathbf{u})^{\mathrm{T}} \mathbf{H}(\boldsymbol{\theta})(t\mathbf{u}) + o(t^2), \end{aligned} \tag{A.1}$$

where

$$\mathbf{g}(\boldsymbol{\theta}) = \left. \frac{\partial D_{\mathrm{KL}}(p;p^t)}{\partial \boldsymbol{\theta}^t} \right|_{\boldsymbol{\theta}^t=\boldsymbol{\theta}} \tag{A.2}$$

and

$$\mathbf{H}(\boldsymbol{\theta}) = \left. \frac{\partial^2 D_{\mathrm{KL}}(p;p^t)}{(\partial \boldsymbol{\theta}^t)^2} \right|_{\boldsymbol{\theta}^t=\boldsymbol{\theta}}, \tag{A.3}$$

with $\boldsymbol{\theta}^t = \boldsymbol{\theta} + t\mathbf{u}$.

The first term equals zero by the definition of the divergence. Suppose the density function fulfills the mild regularity conditions:

$$\int \frac{\partial^k}{\partial \boldsymbol{\theta}^k} p(\mathbf{x};\boldsymbol{\theta}) \mathrm{d}\mathbf{x} = 0, \quad k \in \{1,2\}. \tag{A.4}$$

The second term also vanishes because

$$\begin{aligned} \mathbf{g}(\boldsymbol{\theta}) &= -\int p(\mathbf{x};\boldsymbol{\theta}) \frac{\partial \log p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathrm{d}\mathbf{x} \\ &= -\int \frac{\partial}{\partial \boldsymbol{\theta}} p(\mathbf{x};\boldsymbol{\theta}) \mathrm{d}\mathbf{x} = 0. \end{aligned} \tag{A.5}$$

Furthermore,

$$\begin{aligned} \mathbf{H}(\boldsymbol{\theta}) &= -\int p(\mathbf{x};\boldsymbol{\theta}) \frac{\partial^2}{\partial^2 \boldsymbol{\theta}} \log p(\mathbf{x};\boldsymbol{\theta}) \mathrm{d}\mathbf{x} \\ &= -\int p(\mathbf{x};\boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\theta}} \left(\frac{\partial p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \bigg/ p(\mathbf{x};\boldsymbol{\theta}) \right) \mathrm{d}\mathbf{x} \\ &= -\int p(\mathbf{x};\boldsymbol{\theta}) \left[\frac{\partial^2 p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \frac{1}{p(\mathbf{x};\boldsymbol{\theta})} \right. \\ &\qquad \left. + \frac{\partial p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left(\frac{\partial \left(\frac{1}{p(\mathbf{x};\boldsymbol{\theta})} \right)}{\partial \boldsymbol{\theta}} \right)^{\mathrm{T}} \right] \mathrm{d}\mathbf{x} \\ &= -\int \frac{\partial^2}{\partial \boldsymbol{\theta}^2} p(\mathbf{x};\boldsymbol{\theta}) \mathrm{d}\mathbf{x} \\ &\quad + \int p(\mathbf{x};\boldsymbol{\theta}) \frac{\left(\frac{\partial \log p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \log p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^{\mathrm{T}}}{p^2(\mathbf{x};\boldsymbol{\theta})} \mathrm{d}\mathbf{x} \\ &= E\left\{ \frac{\partial \log p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left(\frac{\partial \log p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^{\mathrm{T}} \right\} \\ &= \mathbf{G}(\boldsymbol{\theta}). \end{aligned} \tag{A.6}$$

Therefore $D_{\mathrm{KL}}(p;p^t) = \frac{1}{2}t^2\mathbf{u}^{\mathrm{T}}\mathbf{G}\mathbf{u} + o(t^2)$. We thus obtain

$$\begin{aligned} \sqrt{2} \lim_{t\to 0} \frac{\sqrt{D_{\mathrm{KL}}(p;p^t)}}{t} &= \sqrt{\mathbf{u}^{\mathrm{T}}\mathbf{G}\mathbf{u}} + \lim_{t\to 0} o(t) \\ &= \|\mathbf{u}\|_{\mathbf{G}}. \quad \square \end{aligned} \tag{A.7}$$

Appendix B. Proof of Theorem 2

We apply the induction method on r. When $r = 1$, the number of columns of $\boldsymbol{\Psi}$ is m. Therefore the column rank of $\boldsymbol{\Psi}$, $\mathrm{rank}_{\mathrm{col}}(\boldsymbol{\Psi})$, is obvious no greater than $m \times r - r \times (r-1)/2 = m$.

Suppose $\mathrm{rank}_{\mathrm{col}}(\boldsymbol{\Psi}^{(k-1)}) \le m(k-1) - (k-1)(k-2)/2$ holds for $r = k-1$, $k \in \mathbb{Z}^+$. Denote $\mathbf{w}^{(j)}$ the jth column of $\mathbf{W}$. Then we can write $\boldsymbol{\Psi}^{(k-1)} = \left(\mathbf{B}^{(1)}, \ldots, \mathbf{B}^{(k-1)}\right)$ with block matrix representation $\mathbf{B}^{(j)}$ which equals

$$\begin{pmatrix} f_1 x_1^{(1)} \sum_{d=1}^m x_d^{(1)} w_d^{(j)} & \ldots & f_1 x_m^{(1)} \sum_{d=1}^m x_d^{(1)} w_d^{(j)} \\ \vdots & \ddots & \vdots \\ f_n x_1^{(n)} \sum_{d=1}^m x_d^{(n)} w_d^{(j)} & \ldots & f_n x_m^{(n)} \sum_{d=1}^m x_d^{(n)} w_d^{(j)} \end{pmatrix}. \tag{B.1}$$

Now consider each of the matrices

$$\tilde{\mathbf{B}}^{(jk)} \equiv \left(\mathbf{B}^{(j)}\mathbf{B}^{(k)}\right), \quad j = 1, \ldots, k-1. \tag{B.2}$$

Notice that the coefficients

$$\boldsymbol{\rho} \equiv \left(w_1^{(k)}, \ldots, w_m^{(k)}, -w_1^{(j)}, \ldots, -w_m^{(j)} \right) \tag{B.3}$$

fulfill

$$\boldsymbol{\rho}^{\mathrm{T}} \tilde{\mathbf{B}}^{(jk)} = 0, \quad j = 1, \ldots, k-1. \tag{B.4}$$

Treating the columns as symbolic objects, one can solve the $k-1$ equations (B.4) by for example Gaussian elimination and then write out the last $k-1$ columns of $\boldsymbol{\Psi}^{(k)}$ as linear combinations of the first $mr-(k-1)$ columns. That is, at most $m-(k-1)$ linearly independent dimensions can be added when $\mathbf{w}^{(k)}$ is appended. The resulting column rank of $\boldsymbol{\Psi}^{(k)}$ is therefore no greater than

$$\begin{aligned} & m(k-1) - \frac{(k-1)(k-2)}{2} + m - (k-1) \\ & \quad = mk - \frac{k(k-1)}{2}. \quad \square \end{aligned} \tag{B.5}$$

References

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation*, *10*(2), 251–276.

Amari, S., & Nagaoka, H. (2000). *Methods of information geometry*. Oxford: University Press.

Amari, S., Park, H., & Fukumizu, K. (2000). Adaptive method of realizing natural gradient learning for multilayer perceptrons. *Neural Computation*, *12*(6), 1399–1409.

Goldberger, J., Roweis, S., Hinton, G., & Salakhutdinov, R. (2005). Neighbourhood components analysis. *Advances in Neural Information Processing*, *17*, 513–520.

Golub, G. H., & van Loan, C. F. (1989). *Matrix Computations* (2nd edition). The Johns Hopkins University Press.

Hyvärinen, A., & Hoyer, P. (2000). Emergence of phase- and shift- invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, *12*(7), 1705–1720.

Newman, D., Hettich, S., Blake, C., & Merz, C. (1998). UCI repository of machine learning databases. http://www.ics.uci.edu/~mlearn/MLRepository.html.

Nishimori, Y., & Akaho, S. (2005). Learning algorithms utilizing quasi-geodesic flows on the Stiefel manifold. *Neurocomputing*, *67*, 106–135.

Peltonen, J., & Kaski, S. (2005). Discriminative components of data. *IEEE Transactions on Neural Networks*, *16*(1), 68–83.

Peterson, P. (1998). *Riemannian geometry*. New York: Springer.

Yang, B. (1995). Projection approximation subspace tracking. *IEEE Transactions on Signal Processing*, *43*(1), 95–107.

Yang, Z., & Laaksonen, J. (2007). Approximated geodesic updates with principal natural gradients. In: *Proceedings of the 2007 international joint conference on neural networks* (pp. 1320–1325).