Key Diagonal Blocks of the Fisher Information Matrix on Neural Manifold of Full-Parametrised Multilayer Perceptrons

Xiongzi Chen
Mathematical College, Sichuan University, Chengdu 610064, P. R. China

Abstract: It’s well known the natural gradient learning (NGL) ([1]) may avoid global optima or phenomena of plateau in the training process since it takes into consideration the intrinsic geometric structure of the parameter space. But, natural gradient ([1]) is itself induced by Fisher information matrix (FIM) ([2]) defined on the 1-form tangent space ([3]), therefore calculation of relevant FIM is key to the realization of NGL. This paper gives explicit derivation and compact matrix representation of the diagonal blocks, and their inverses as well, of the FIM based Riemannian metric on neural manifold of full-parametrised multilayer perceptrons (MLP), thus extending and complementing the results partially given in [1] and [3].

Keywords: Neural manifold, Natural gradient learning, Fisher information matrix (FIM), Full-parametrised Multilayer perceptrons (FP-MLPs).

1 Introduction

The simplest neural manifold is the space of all single feedforward neurons (i.e., perceptrons) with input nodes and single output node. We use the same settings and notations proposed in [1] without restating the detailed assumptions on the variables that occur in the model

\[ y = f(w \cdot x) + \varepsilon, \]

(1)

where \( x \sim N(0, I) \) and \( \varepsilon \sim N(0, \sigma^2) \) and \( x, \varepsilon \) are statistically independent, \( w = (w_1, \ldots, w_n)^T \in \mathbb{R}^n \) and, \( |w| = \left( \sum_{i=1}^n w_i^2 \right)^{1/2} \) which in this paper will denote its Euclidean norm. It’s obtained in [1] that the FIM is

\[ G(w) = |w|^2 c_1(w)I + |c_2(w) - c_1(w)|ww^T. \]

(1)

However, the mathematical details on how (1) is derived was not fully shown.

To investigate the FIM on neural manifolds \( \mathcal{M} \) of full-parametrised MLPs (which are MLPs with additionally thresholds variables), paper [4] shows the block form of this FIM:

For neural manifold \( \mathcal{M} \) of all fully connected \( n-p-1 \) type MLPs whose input-output relation is \( y = v \cdot f(x) + \varepsilon = \sum_{i=1}^p v_i f_i(w_i \cdot x + \theta) + \varepsilon \), where \( x \) and \( \varepsilon \) are mutually independent and normally distributed respectively as \( N(0, I) \) and \( N(0, \sigma^2) \), the Fisher information matrix \( G(W, \theta, v) \) in blocks is

\[ G(W, \theta, v) = \frac{1}{\sigma^2} E \left[ \begin{pmatrix} \tilde{V}_w \ln p(\cdot) \otimes (\varepsilon \cdot \tilde{V}_w \ln p(\cdot)) & [\tilde{V}_w \ln p(\cdot)]f^T & [\tilde{V}_w \ln p(\cdot)](v \ast \tilde{L}_w f)^T & (v \ast \tilde{L}_w f) \otimes (v \ast \tilde{L}_w f)^T \\ f[\tilde{V}_w \ln p(\cdot)]^T & \ast(v \ast \tilde{L}_w f) \otimes f^T & (v \ast \tilde{L}_w f) \otimes (v \ast \tilde{L}_w f)^T & \ast \otimes f^T \end{pmatrix} \right] \]

(2)

where \( \tilde{V}_w \ln p(\cdot) = (\tilde{V}_w \ln p(\cdot), \ldots, \tilde{V}_w \ln p(\cdot))^T \in \mathbb{R}^{kn} \).

Though [1], [3] partially touches the blocks of the diagonal block \( G_{wp} = [\tilde{V}_w \ln p(\cdot)] \otimes [\tilde{V}_w \ln p(\cdot)]^T \), the details and compact representations are missing, hereunder we’ll provide the explicit representations of these diagonal blocks and their inverses.

2 Key Diagonal Blocks of the Riemannian Metric on \( \mathcal{M} \)

To provide a concise mathematical statement here in, we use almost the same settings given in [4] unless otherwise stated or supplemented. Here it’s supposed that \( f_i' (\tau) \in L_2(\mathbb{R}, d\tau) \), i.e., \( f_i' \) belongs to the family of square integrable functions with respect to the Lebesgue measure \( d\tau \) on the Borel \( \sigma \)-algebra over the real line \( \mathbb{R} \) and that the probability densities appearing here are regular in the sense specified in [3].
First, through theorem 1, we provide full details on how formula (1) was derived and provide its compact representation in matrix notations, then we proceed on to obtain the main targets of this paper.

**Theorem 1**

For the model \( y = f(w \cdot x) + \varepsilon \) under the setting given in [1], the Fisher information matrix and its inverse at \( w \) are respectively

\[
G(w) = \begin{pmatrix} |w|^2 & c_2 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} w \\ T \end{pmatrix}, \quad G^{-1}(w) = \begin{pmatrix} |w|^2 & c_2 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} w \\ T \end{pmatrix}^T.
\]

where the symbols bear the meanings given in the following proof.

**Proof:**

Suppose \( \{\tilde{w}, v_1, \cdots, v_{n-1}\} \), where \( \tilde{w} = w/|w| \), forms the orthonormal basis for \( \mathbb{R}^n \). Then \( \{\tilde{w} \cdot x, v_1 \cdot x, \cdots, v_{n-1} \cdot x\} \) are mutually statistically independent with \( |w_0| = \sqrt{w \cdot w_0} \sim N(0,1) \) and \( v_j \cdot x = \tilde{x} \sim N(0,1) \), and consequently

\[
\begin{align*}
\gamma_{ij} &= \tilde{w}^T G(w) \tilde{w} = \frac{1}{\sigma^2} E(|f'(w \cdot x)|^2 (w \cdot x)^2) = \frac{1}{\sigma^2} \int \tilde{x}^2 f'(\tilde{w} \cdot \tilde{x})^2 \exp\{-\frac{\tilde{x}^2}{2}\} d\tilde{x} \\
c_{ij} &= (v_j)^T G(w)v_j = \frac{1}{\sigma^2} E(|f'(w \cdot x)|^2 (v_j \cdot x)^2) = \frac{1}{\sigma^2} \int f'(\tilde{w} \cdot \tilde{x})^2 \exp\{-\frac{\tilde{x}^2}{2}\} d\tilde{x} \\
c_{ij} &= (v_i)^T G(w)v_j = \frac{1}{\sigma^2} E(|f'(w \cdot x)|^2 (v_i \cdot x)(v_j \cdot x)) = 0, i \neq j \\
c_{ij} &= (w_0)^T G(w)v_i = \frac{1}{\sigma^2} E(|f'(w \cdot x)|^2 (w_0 \cdot x)(v_i \cdot x)) = 0
\end{align*}
\]

Set \( |w|^2 c_2 = \gamma_{11} \), \( |w|^2 c_i = c_{ii} \); \( T = (v_1 \cdots v_{n-1}) \); \( C = \begin{pmatrix} c_{11} & \cdots & c_{n-2,n-1} \end{pmatrix} \); \( C = \begin{pmatrix} |w|^2 c_2, |w|^2 c_I \end{pmatrix} \). Then, from the orthogonality of \( T \), it’s inferred that

\[
G(w) = \begin{pmatrix} w \\ T \end{pmatrix}^T C T = \begin{pmatrix} w \\ T \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma^2} |w|^2 & c_2 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} w \\ T \end{pmatrix} = |w|^2 c_i T T^T + c_2 w w^T.
\]

Since \( I = T T^T = \begin{pmatrix} \tilde{w} \\ T \end{pmatrix} \begin{pmatrix} \tilde{w}^T \\ T^T \end{pmatrix} = \tilde{w}^T \tilde{w} + T T^T \), then \( T T^T = I - \tilde{w}^T \tilde{w} \) and hence

\[
G(w) = |w|^2 c_i (I - \tilde{w}^T \tilde{w}) + c_2 w w^T = |w|^2 c_i w + |w|^4 (c_i^2 - c_2^2) w w^T.
\]

Further, the inverse \( G^{-1}(w) \) is

\[
G^{-1}(w) = \begin{pmatrix} w \\ T \end{pmatrix}^T C^{-1} T = \begin{pmatrix} w \\ T \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma^2} |w|^2 & c_2 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} w \\ T \end{pmatrix} = |w|^2 c_i^{-1} I + |w|^4 (c_i^2 - c_2^2) w w^T,
\]

which completes the proof.

2.2 Now, let’s consider full-parameters MLPs proposed in [4] as:

\[
y = v \cdot f(x) + \varepsilon = \sum_{i=1}^k v_i f_i(w_i \cdot x + \theta_i) + \varepsilon
\]

(II)

where \( x \sim N(0,1) \), \( \varepsilon \sim N(0,\sigma^2) \). For model (II), the joint probability density function is

\[
\ln p(x,y|W,v,\theta) = \ln q(x) \ln p(y|x,W,v,\theta) = \ln q(x) - \ln(\sqrt{2\pi\sigma}) - \frac{1}{2\sigma^2} [y - v \cdot f(x)]^2,
\]

so that its gradients are
where \( p(\cdot) = p(x, y; W, v, \theta) \) satisfies regularity conditions specified in [5] and \( p(\cdot) = p(y \mid x; W, v, \theta) \).

By the definition given in [2], for model (II) the corresponding FIM should be

\[
G(p) = G(W, \theta, v) = E\left( \nabla\ln p(\cdot) \nabla\ln p(\cdot)^T \right).
\] 

(6)

For simplicity of notations, we specifically set \( G_{w} = (G_{w, w})_{i=i} = G_{w, w} \otimes G_{w, w} \), where \( \otimes \) is the symbol of tensor product and

\[
G_{w, w} = E\left( \nabla_{w} \ln p(\cdot) \nabla_{w} \ln p(\cdot)^T \right) = \frac{v_{y}}{\sigma^{2}} E\left( f'_{y}(w_{x} + \theta) f'_{y}(w_{x} + \theta) xx^{T} \right).
\]

Hereunder we’ll obtain explicitly the diagonal blocks and their inverses of \( G_{w} \), whose compact representations are also shown.

2.2.1 For the key Blocks \( G_{w, w}, i \neq j \) of \( G_{w} = (G_{w, w})_{i=i} \) and their inversions, we have

**Theorem 2** Suppose \( \text{rank} W = k \), then for the model \( y = v \cdot f(x) + \varepsilon = \sum_{i=1}^{k} v_{i} f(w_{i} + \theta) + \varepsilon \), the blocks \( G_{w, w}, i \neq j \) of the FIM and their inverse are respectively

\[
G_{w, w} = (T_{i}, T_{i})\left( A_{y} \ 0 \ \beta_{y} I \right)\begin{pmatrix} T_{i}^{T} \ T_{i}^{T} \end{pmatrix} = \beta_{y}^{-1} I + (w_{i}, w_{j})D_{y} (w_{i}, w_{j})^{T},
\]

(7)

\[
G_{w, w}^{-1} = (T_{i}, T_{i})\left( A_{y}^{-1} \ 0 \ \beta_{y}^{-1} I \right)\begin{pmatrix} T_{i}^{T} \ T_{i}^{T} \end{pmatrix} = \beta_{y}^{-1} I + (w_{i}, w_{j})H_{y} (w_{i}, w_{j})^{T},
\]

(8)

where the meanings of the symbols and values are given in the proof.

**Proof:** Since for any pair of \( \{w_{i}, w_{j}\} \), there exists an orthonormal basis \( \{v_{1}, \ldots, v_{n}\} \) for \( \mathbb{R}^{n} \) such that \( L_{i,j} = \text{span}\{v_{i}, v_{j}\} = \text{span}\{w_{i}, w_{j}\} \) and \( L_{i,j} \perp L_{i} \) in the Euclidian sense, where \( L_{i} = \text{span}\{v_{1}, \ldots, v_{k}\} \), so all the vectors \( \{v_{1}, x, \ldots, v_{n}, x\} \) are statistically independent and distributed normally as \( N(0,1) \).

We set

\[
\begin{pmatrix} w_{i} \ w_{j} \end{pmatrix}^{T} = K (v_{1} \ v_{2})^{T}, K = \begin{pmatrix} k_{i1} & k_{j1} \\ k_{i2} & k_{j2} \end{pmatrix},
\]

\[
\begin{pmatrix} v_{1} \ v_{2} \end{pmatrix}^{T} = P (w_{i} \ w_{j})^{T}, K^{-1} = P = \begin{pmatrix} p_{i1} & p_{i2} \\ p_{j1} & p_{j2} \end{pmatrix} = \frac{1}{k_{i1}k_{j2} - k_{i2}k_{j1}} \begin{pmatrix} k_{j2} & -k_{i2} \\ -k_{j1} & k_{i1} \end{pmatrix}
\]

(2-2)

First of all, we could obtain two groups of four independent values relevant respectively to \( w_{i}, w_{j} \) and \( v_{1}, v_{2} \) (see (A.1), (A.2) of the Appendix 1 for more details), which are

\[
A_{y} = (v_{1} \ v_{2})^{T} G_{w, w} (v_{1} \ v_{2}) = \begin{pmatrix} v_{1}^{T} G_{w, w} v_{1} & v_{1}^{T} G_{w, w} v_{2} \\ v_{2}^{T} G_{w, w} v_{1} & v_{2}^{T} G_{w, w} v_{2} \end{pmatrix} = \begin{pmatrix} \lambda_{y}^{11} & \lambda_{y}^{12} \\ \lambda_{y}^{21} & \lambda_{y}^{22} \end{pmatrix},
\]

\[
\Gamma_{y} = (w_{i} \ w_{j})^{T} G_{w, w} (w_{i} \ w_{j}) = \begin{pmatrix} w_{i}^{T} G_{w, w} w_{i} & w_{i}^{T} G_{w, w} w_{j} \\ w_{j}^{T} G_{w, w} w_{i} & w_{j}^{T} G_{w, w} w_{j} \end{pmatrix} = \begin{pmatrix} \gamma_{y}^{11} & \gamma_{y}^{12} \\ \gamma_{y}^{21} & \gamma_{y}^{22} \end{pmatrix}
\]

(2-3)

Obviously, \( \Gamma_{y} = KA_{y}K^{T} \), or equivalently,
\[ A_y = P \Gamma_y P^T. \] (9)

For \( v_i, 2 < l \leq n \), we immediately have

\[
\begin{align*}
(2-4) & \quad \beta_{ij} = v_i^T G_{w, y} v_i = \frac{v_i^T G_{w, y} v_i}{\sigma^2} E \left[ f'_y (w_i \cdot x + \theta_j) f'_y (w_j \cdot x + \theta_j) \right], \quad 2 < l \leq n \\
& \quad A_{lm} = v_i^T G_{w, y} v_m = 0, \quad 2 < l, m \leq n, l \neq m
\end{align*}
\]

By setting \( R = (w_1 \quad w_j) \), \( T_1 = (v_1 \quad v_2) \), \( T_2 = (v_3 \ldots v_n) \), \( T_y = (T_1 \quad T_2)^T \), \( C_1 = \beta_{y} I_{(n-2)(n-2)} \), \( C_2 = \text{diag} \{A_y, C_1\} = \text{diag} \{A_y, \beta_{y} I_{(n-2)(n-2)}\} \). Then \( T_y^T = P R^T \), we could obtain from (2-3), (2-4) that \( T G_{w, y} T^T = C \), which induces compactly

\[
G_{w, y} = T_y^T C_y T_y = (T_1 \quad T_2) \begin{pmatrix} A_y & 0 \\ 0 & \beta_{y} I_{(n-2)(n-2)} \end{pmatrix} \begin{pmatrix} T_1^T \\ T_2^T \end{pmatrix}. \tag{10}
\]

Expanding (10) into

\[
G_{w, y} = (T_1 \quad T_2) \begin{pmatrix} A_y & 0 \\ 0 & \beta_{y} I_{(n-2)(n-2)} \end{pmatrix} \begin{pmatrix} T_1^T \\ T_2^T \end{pmatrix} = T A_y T^T + \beta_y T T_y^T,
\]

and inferring from (2-2), we have

\[
T A_y T^T = I - v_1 v_1^T - v_2 v_2^T. \tag{12}
\]

(The exact expressions of (11) and (12) are given by (A-3), (A-4) of Appendix 1.)

Hence, we have

\[
G_{w, y} = \lambda_{y}^{11} v_1 v_1^T + \lambda_{y}^{12} v_1 v_2^T + \lambda_{y}^{22} v_2 v_2^T + \beta_y (I - v_1 v_1^T - v_2 v_2^T), \tag{13}
\]

which is

\[
G_{w, y} = \frac{v_i^T}{\sigma^2} E \left[ f'_y (w_i \cdot x + \theta_j) f'_y (w_j \cdot x + \theta_j) xx^T \right] = \beta_{y} I + \left( w_i \quad w_j \right) D_y \begin{pmatrix} w_i \\ w_j \end{pmatrix}^T, \tag{14}
\]

where \( D_y = P^T A_y P - \beta_y P^T P = \begin{pmatrix} d_{ii} & d_{ij} \\ d_{ji} & d_{jj} \end{pmatrix} \) and \( \beta_y \) are given in (A.4), (A.5) of the Appendix 1.

As for the inverse of \( G_{w, y} \), we have simply the compact form

\[
G_{w, y}^{-1} = (T_1 \quad T_2) \begin{pmatrix} A_y^{-1} & 0 \\ 0 & \beta_{y}^{-1} I \end{pmatrix} \begin{pmatrix} T_1^T \\ T_2^T \end{pmatrix} = T A_y^{-1} T + \beta_y^{-1} T T_y^T. \tag{15}
\]

Setting \( |A_y| = \det A_y^{-1} \), we then have \( A_y^{-1} = \frac{1}{|A_y|} \begin{pmatrix} \lambda_{y}^{22} & -\lambda_{y}^{12} \\ -\lambda_{y}^{21} & \lambda_{y}^{11} \end{pmatrix} \) and consequently

\[
T A_y^{-1} T^T = \frac{1}{|A_y|} \begin{pmatrix} 1 & v_2 \\ v_1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_{y}^{22} & -\lambda_{y}^{12} \\ -\lambda_{y}^{21} & \lambda_{y}^{11} \end{pmatrix} \begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix} = \frac{\lambda_{y}^{22} v_1 v_1^T + (\lambda_{y}^{21}) v_1 v_2^T + (\lambda_{y}^{12}) v_2 v_2^T + \lambda_{y}^{11} v_2 v_2^T}{|A_y|} \tag{16}
\]

and
\[
G_{w_{ij}}^{-1} = (T_1 \ T_2) \left( \begin{array}{cc}
A_{ij}^{-1} & 0 \\
0 & \beta_{ij}^{-1}I
\end{array} \right) \left( \begin{array}{c}
T_1^T \\
T_2^T
\end{array} \right) = \beta_{ij}^{-1}I + \left( \begin{array}{c}
w_i \\
w_j
\end{array} \right) H_{\beta} \left( \begin{array}{c}
w_i \\
w_j
\end{array} \right)^T,
\]

where \( H_{\beta} = P^T A_{ij}^{-1}P - \beta_{ij}^{-1}P^TP = \left( \begin{array}{cc}
h_{ij} & h_{i,j} \\
h_{j,i} & h_{jj}
\end{array} \right) \) is defined by (A.6) in Appendix 1.

For the blocks \( G_{w_{ii}} \) of \( G_W \), we have similarly
\[
G(w_i) = w_i^2 c_{i,j} + c_{i,j}^2 w_i w_j^T \quad \text{and} \quad G(w_i) = w_i^2 c_{i,j}^{-1}I + |w_i|^2 c_{i,j}^{-1}w_i w_j^T,
\]
where \( c_{i,j}, c_{i,j}^2 \) are obtained by the same method used in theorem 1.

### 3 Concluding Remarks

In all the sections above, we have struggled to give a detailed and compact representation of the diagonal blocks \( G_{w_{ij}} \) of the FIM \( G(\rho) \) on neural manifold \( M \) of full-parametrised MLPs, thus extending results of similar blocks in simpler neural manifolds under more versatile settings. Obviously, the increased complexities in calculations of FIM actually in most cases only permits a block matrix representation of the desired metric, since the activation functions are of vast variety and the independence assumptions needed to explicitly obtain the entries of \( G(\rho) \) and its inverse are too restrictive. Despite of the tantalization of the entry-wise presentation of \( G(\rho) \) and \( G^{-1}(\rho) \), we’ve nonetheless went a step further in calculating some key blocks and we’ll probably resort to the eigenstructure of the square blocks of \( G(\rho) \) under stringent conditions and try to find out more if feasible.

### 4 Appendix 1

To make the paper more concentrated on its core contents, we have exclusively moved the complicated computations here with reference remarks given at the appropriate places in previous sections.

#### 4.1 Lemma 1

For any orthogonal matrix \( T_{n \times n} = (v_1 \ldots v_n) = (T_1 \ T_2) \), where \( T_1 = (v_1 \ldots v_i) \), \( T_2 = (v_{i+1} \ldots v_n) \), \( 1 \leq i, i_n \leq n \), the identity holds
\[
T_1 T_2^T = I - \sum_{i=1}^{\frac{n}{2}} v_i v_i^T.
\]

**proof:** From the orthogonality of \( T \), it’s obtained that
\[
I = T_{n \times n} T_{n \times n}^T = (v_1 \ldots v_n T_1)(v_1 \ldots v_n T_2)^T = T_1 T_2^T + \sum_{i=1}^{\frac{n}{2}} v_i v_i^T,
\]
hence \( T_2 T_2^T = I - \sum_{i=1}^{\frac{n}{2}} v_i v_i^T. \)

#### 4.2 Exact expressions of \( A_g, \Gamma_g \):

\[
\begin{align*}
\gamma_{g1}^{1,1} &= w_i^T G_{w_{ij}} w_i = \frac{\nu \nu_j}{\sigma^2} E \left[ f'_i(w_i \cdot x + \theta) f'_j(w_j \cdot x + \theta) (w_i \cdot x)^2 \right] \\
\gamma_{g1}^{1,2} &= w_i^T G_{w_{ij}} w_j = \frac{\nu \nu_j}{\sigma^2} E \left[ f'_i(w_i \cdot x + \theta) f'_j(w_j \cdot x + \theta) (w_j \cdot x)^2 \right] \\
\gamma_{g1}^{2,1} &= w_j^T G_{w_{ij}} w_i = \frac{\nu \nu_j}{\sigma^2} E \left[ f'_i(w_i \cdot x + \theta) f'_j(w_j \cdot x + \theta) (w_i \cdot x) \right] \\
\gamma_{g1}^{2,2} &= w_j^T G_{w_{ij}} w_j = \frac{\nu \nu_j}{\sigma^2} E \left[ f'_i(w_i \cdot x + \theta) f'_j(w_j \cdot x + \theta) (w_j \cdot x) \right]
\end{align*}
\]
\[
\begin{align*}
\lambda_{ij} &= v_i^T G_{ij} v_i = \frac{v_i^T v_j}{\sigma^2} E \left\{ f_i'(w_i \cdot x + \theta) f_j'(w_j \cdot x + \theta)(v_i \cdot x) \right\} \\
\lambda_{ij}^{22} &= v_i^T G_{ij} v_j = \frac{v_i^T v_j}{\sigma^2} E \left\{ f_i'(w_i \cdot x + \theta) f_j'(w_j \cdot x + \theta)(v_i \cdot x) \right\} \\
\lambda_{ij}^{12} &= v_i^T G_{ij} v_j = \frac{v_i^T v_j}{\sigma^2} E \left\{ f_i'(w_i \cdot x + \theta) f_j'(w_j \cdot x + \theta)(v_i \cdot x)(v_j \cdot x) \right\} \\
\lambda_{ij}^{21} &= v_i^T G_{ij} v_j = \frac{v_i^T v_j}{\sigma^2} E \left\{ f_i'(w_i \cdot x + \theta) f_j'(w_j \cdot x + \theta)(v_j \cdot x) \right\} \\
\lambda_{ij}^{11} &= v_i^T G_{ij} v_j = \lambda_{ij}^{22} \\
\end{align*}
\]

(A.2)

4.3 The exact expressions of $\beta_i, d_{ij},$ etc.:

we set $K_{ij}^{-1} = \det K_{ij}^{-1}$, then

\[
\begin{align*}
T_{ij}^T &= I \frac{1}{\det K_{ij}} \left\{ (k_{ij}^2 + k_{ij}) w_i w_j^T - (k_{ij}^2 + k_{ij}^2) w_j w_j^T - (k_{ij}^2 + k_{ij}) w_j w_j^T + (k_{ij}^2 + k_{ij}) w_j w_j^T \right\} \\
\end{align*}
\]

(A.3)

\[
\begin{align*}
v_{ij} &= \frac{1}{\det K_{ij}} \left\{ (k_{ij}^2 + k_{ij}) w_i w_j^T + (k_{ij}^2 + k_{ij}) w_j w_i^T + (k_{ij}) w_j w_i^T \right\} \\
v_{ij} &= \frac{1}{\det K_{ij}} \left\{ (k_{ij}^2 + k_{ij}) w_i w_j^T + (k_{ij}) w_j w_i^T \right\} \\
v_{ij} &= \frac{1}{\det K_{ij}} \left\{ (k_{ij}^2 + k_{ij}) w_i w_j^T + (k_{ij}) w_j w_i^T \right\} \\
\end{align*}
\]

(A.4)

\[
\begin{align*}
d_{ij} &= \frac{1}{\det K_{ij}} \left\{ \lambda_{ij}^{11} k_{ij}^2 + \lambda_{ij}^{12} k_{ij} + \lambda_{ij}^{21} k_{ij} + \lambda_{ij}^{22} k_{ij}^2 - (k_{ij}^2 + k_{ij}) \right\} \\
d_{ij} &= \frac{1}{\det K_{ij}} \left\{ \lambda_{ij}^{11} k_{ij}^2 + \lambda_{ij}^{12} k_{ij} + \lambda_{ij}^{21} k_{ij} + \lambda_{ij}^{22} k_{ij}^2 - (k_{ij}^2 + k_{ij}) \right\} \\
d_{ij} &= \frac{1}{\det K_{ij}} \left\{ \lambda_{ij}^{11} k_{ij}^2 + \lambda_{ij}^{12} k_{ij} + \lambda_{ij}^{21} k_{ij} + \lambda_{ij}^{22} k_{ij}^2 - (k_{ij}^2 + k_{ij}) \right\} \\
\end{align*}
\]

(A.5)

\[
\begin{align*}
h_{ij} &= \frac{1}{\det A_{ij}^{-1}} \left\{ \lambda_{ij}^{11} k_{ij}^2 + \lambda_{ij}^{12} k_{ij} + \lambda_{ij}^{21} k_{ij} + \lambda_{ij}^{22} k_{ij}^2 - (k_{ij}^2 + k_{ij}) \right\} \\
h_{ij} &= \frac{1}{\det A_{ij}^{-1}} \left\{ \lambda_{ij}^{11} k_{ij}^2 + \lambda_{ij}^{12} k_{ij} + \lambda_{ij}^{21} k_{ij} + \lambda_{ij}^{22} k_{ij}^2 - (k_{ij}^2 + k_{ij}) \right\} \\
h_{ij} &= \frac{1}{\det A_{ij}^{-1}} \left\{ \lambda_{ij}^{11} k_{ij}^2 + \lambda_{ij}^{12} k_{ij} + \lambda_{ij}^{21} k_{ij} + \lambda_{ij}^{22} k_{ij}^2 - (k_{ij}^2 + k_{ij}) \right\} \\
\end{align*}
\]

(A.6)

5 Acknowledgements

The author would like to thank Mr. Guojun Zhao for discovering the relationship in Lemma 1 of Appendix 1, which leads to the final detail of (3).

References

多层感知器流形上的 Fisher 信息矩阵的关键分块子矩阵

陈雄志，蔡长林
四川大学数学学院，成都 610064

摘要：众所周知，自然梯度学习法 (1) 能够在训练过程中避免局部最优或高原现象，因为它考虑到了参数空间的内禀几何性质。但是，自然梯度 (1) 本身是由定义在 1-形式切空间上的 Fisher 信息矩阵 (FIM) 来诱导的，所以对 FIM 的计算就成为实现自然梯度学习的关键环节。本文给出了全参数 MLP 流形上的 FIM 的关键对角子块及其逆的显式表达和紧凑矩阵表达。如此，本文就将 (1) 和 (3) 中部分给出的结果做了推广和补充。

关键词：神经流形，自然梯度学习，Fisher 信息矩阵 (FIM)，全参数多层感知器 (Full-parametrised MLP)