RESEARCH PAPER

Invariance properties of the natural gradient in overparametrised systems

Jesse van Oostrum[1] · Johannes Müller[2] · Nihat Ay[1,3,4]

Abstract

The natural gradient field is a vector field that lives on a model equipped with a distinguished Riemannian metric, e.g. the Fisher–Rao metric, and represents the direction of steepest ascent of an objective function on the model with respect to this metric. In practice, one tries to obtain the corresponding direction on the parameter space by multiplying the ordinary gradient by the inverse of the Gram matrix associated with the metric. We refer to this vector on the parameter space as the natural parameter gradient. In this paper we study when the pushforward of the natural parameter gradient is equal to the natural gradient. Furthermore we investigate the invariance properties of the natural parameter gradient. Both questions are addressed in an overparametrised setting.

Keywords Natural gradient · Riemannian metric · Deep learning · Information geometry

1 Introduction

Within the field of deep learning, gradient methods have become ubiquitous tools for parameter optimisation. Standard gradient optimisation procedures use the

A previous version of this paper [14] was presented at the International Conference on Geometric Science of Information 2021 in Paris. Here invariance properties for proper parametrisations were discussed. The current paper treats invariance for both proper and non-proper parametrisations.

✉ Jesse van Oostrum
    jesse.van@tuhh.de

1   Hamburg University of Technology, Hamburg, Germany

2   Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany

3   Leipzig University, Leipzig, Germany

4   Santa Fe Institute, Santa Fe, USA

&#x2A themed; Springer

vector of coordinate derivatives of the objective function as the update direction of the parameters. This is implicitly assuming a Euclidean geometry on the space of parameters. It can be argued that this is not always the most natural choice of geometry. Instead one can choose a more natural geometry for the problem at hand and then determine the Riemannian gradient of the objective function for this natural geometry, resulting in the so called *natural gradient*. The *natural gradient method* is the optimisation algorithm that performs discrete parameter updates in the direction of the natural gradient. This method was first proposed by Amari [1] using the geometry induced by the Fisher–Rao metric. It is an active field of study within information geometry [3, 6, 10] and has been shown extremely effective in many applications [4, 16, 17]. More recently, also other geometries on the model have been studied, such as the Wasserstein geometry [9, 11]. The natural gradient is defined independently of a specific parametrisation. Although it is an open problem, there is work supporting the idea that the efficiency of learning of the method is due to this invariance [18].

In practice, the update direction of the parameters is given by the ordinary gradient multiplied by the inverse of the Gram matrix associated with the metric on the model. We will refer to this vector on the parameter space as the *natural parameter gradient*. In order to determine whether this direction is desired we have to map this vector to the model, since it is the location on the model, not on the parameter space, that determines the performance of the model. In non-overparametrised systems it can be shown that in a non-singular point on the model, the pushforward of the natural parameter gradient is equal to the natural gradient. Furthermore the natural parameter gradient can be called parametrisation invariant in this case [12]. In many practical applications of machine learning, and in particular deep learning, one deals however with overparametrised models, in which different directions on the parameter space correspond to a single direction on the model. In this case, the Gram matrix is degenerate and we use a generalised inverse to calculate the natural parameter gradient. In this paper, we will investigate whether the pushforward of the natural parameter gradient remains equal to the natural gradient in the overparametrised setting.

The Moore-Penrose (MP) inverse is the canonical choice of generalised inverse for the natural parameter gradient [5]. The definition of the MP inverse is based on the Euclidean inner product defined on the parameter space. Using the MP inverse is therefore thought to affect the parametrisation invariance of the natural parameter gradient [13], and thus potentially the performance of the natural gradient method. In this paper we propose two different notions of invariance. The first evaluates the invariance of the natural parameter gradient by examining the behaviour of its pushforward on the model. The second looks at the behaviour on the parameter space itself. Since the location and direction on the model is what matters, we argue that the former is of greater importance.
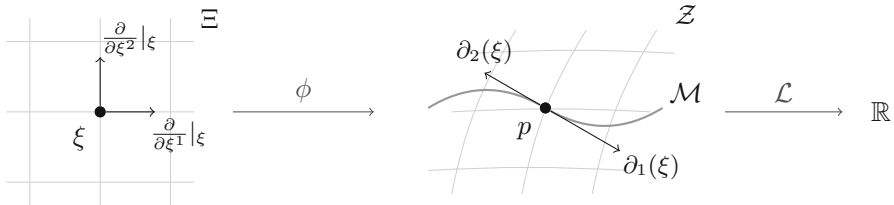
Fig. 1 Parametrisation and objective function

2 The natural gradient

Let $(\mathcal{Z}, g)$ be a Riemannian manifold, Ξ be a parameter space that we assume to be an open subset of $\mathbb{R}^d$, $\phi\colon \Xi \to \mathcal{Z}$ a smooth map (taking the role of a parametrisation[1]), $\mathcal{M} := \phi(\Xi) \subset \mathcal{Z}$ a model[2], and $\mathcal{L}\colon \mathcal{Z} \to \mathbb{R}$ a smooth (objective) function (see Fig. 1). We call $p \in \mathcal{M}$ *non-singular* if $\mathcal{M}$ is locally an embedded submanifold of $\mathcal{Z}$ around p and we denote the set of non-singular points with Smooth($\mathcal{M}$). A point p is called *singular* if it is not non-singular. The Riemannian gradient of $\mathcal{L}$ on $\mathcal{Z}$ is defined implicitly as follows:

$$g_p \left(\mathrm{grad}_p^{\mathcal{Z}} \mathcal{L}, \cdot \right) = d\mathcal{L}_p(\cdot). \tag{1}$$

By the Riesz representation theorem, this defines the gradient uniquely.

Definition 1 (Natural gradient) For $p \in$ Smooth($\mathcal{M}$) the Riemannian gradient of $\mathcal{L}|_{\mathcal{M}}$ on the model $\mathcal{M}$ is called the *natural gradient* and is denoted $\mathrm{grad}_p^{\mathcal{M}} \mathcal{L}$.

It is easy to show that:

$$\mathrm{grad}_p^{\mathcal{M}} \mathcal{L} = \Pi_p(\mathrm{grad}_p^{\mathcal{Z}} \mathcal{L}), \tag{2}$$

where Π_p is the projection onto $T_p\mathcal{M}$. We define the pushforward of the tangent vector on the parameter space through the parametrisation as $\partial_i(\xi) := d\phi_\xi \left(\left. \frac{\partial}{\partial \xi^i} \right|_\xi \right)$, and the Gram matrix $G(\xi)$ by $G_{ij}(\xi) := g_{\phi(\xi)} \left(\partial_i(\xi), \partial_j(\xi) \right)$. We denote the vector of coordinate derivatives with $\nabla_\xi \mathcal{L} := (\partial_1(\xi)\mathcal{L}, ..., \partial_d(\xi)\mathcal{L}) = \left(\frac{\partial \mathcal{L} \circ \phi}{\partial \xi^1}(\xi), ..., \frac{\partial \mathcal{L} \circ \phi}{\partial \xi^d}(\xi) \right) \in \mathbb{R}^d$. Let ξ be such that $\phi(\xi) \in$ Smooth($\mathcal{M}$). We say that a parametrisation is *proper* in ξ when: span $(\{\partial_1(\xi), ..., \partial_d(\xi)\}) = T_{\phi(\xi)}\mathcal{M}$. Furthermore, following the Einstein summation convention, we write $a^i b_i$ for the sum $\sum_i a^i b_i$.

---

[1] Within the context of differential geometry, a parametrisation usually means a local diffeomorphism onto its image. This is also referred to as a coordinate system. In our context, we use a more general definition of parametrisation, where $\phi\colon \Xi \to \mathcal{Z}$ no longer needs to be a diffeomorphism onto its image but only smooth.

[2] Note that in the literature this is also called a parametrised model, statistical manifold, or in the context of machine learning, a neuromanifold. We choose the word model, however, to emphasize that we do not assume a manifold structure on $\mathcal{M}$.

Definition 2 (Generalised inverse) A *generalised inverse* of an $n \times m$ matrix A, denoted A^+, is an $m \times n$ matrix satisfying the following property:

$$AA^+A = A. \tag{3}$$

Note that this definition implies that for $w \in \mathbb{R}^n$ in the image of A, i.e. $w = Av$ for some $v \in \mathbb{R}^m$, we have:

$$AA^+w = AA^+Av = Av = w. \tag{4}$$

This shows that AA^+ is the identity operator on the image of A.

Definition 3 (Natural parameter gradient) We define the *natural parameter gradient* to be the following vector on the parameter space:

$$\widetilde{\nabla}_\xi \mathcal{L} := \left(G^+(\xi)\nabla_\xi \mathcal{L}\right)^i \left.\frac{\partial}{\partial \xi^i}\right|_\xi. \tag{5}$$

The pushforward of this vector, given by:

$$d\phi_\xi \widetilde{\nabla}_\xi \mathcal{L} = \left(G^+(\xi)\nabla_\xi \mathcal{L}\right)^i \partial_i, \tag{6}$$

is called the *natural parameter gradient on* $\mathcal{M}$.

Often, the natural parameter gradient is denoted with matrix notation as follows:

$$\widetilde{\nabla}_\xi \mathcal{L} := G^+(\xi)\nabla_\xi \mathcal{L}, \tag{7}$$

where an identification between the canonical basis of $\mathbb{R}^d$ and the vectors $\left.\frac{\partial}{\partial \xi^i}\right|_\xi$ is made implicitly.

We are now in the position to state the main result of the paper:

Theorem 1 *Let* $\xi \in \Xi$ *and* $p = \phi(\xi) \in \mathcal{M}$. *We have:*

$$d\phi_\xi \widetilde{\nabla}_\xi \mathcal{L} = \Pi_\xi \left(\mathrm{grad}_p^{\mathcal{Z}} \mathcal{L}\right), \tag{8}$$

where Π_ξ *is the projection onto* $\mathrm{span}\{\partial_i(\xi)\}_i$. *In particular, when* $\phi(\xi)$ *is non-singular and* $\mathrm{span}\{\partial_i(\xi)\}_i = T_p\mathcal{M}$ *we have:*

$$d\phi_\xi \widetilde{\nabla}_\xi \mathcal{L} = \mathrm{grad}_p^{\mathcal{M}} \mathcal{L}. \tag{9}$$

This theorem implies that under certain conditions the pushforward of the natural parameter gradient is equal to the natural gradient. Furthermore we see that in general the natural parameter gradient on $\mathcal{M}$ is dependent on the choice of parametrisation through Π_ξ, but becomes invariant when the coordinate vectors span the full tangent

space of $\mathcal{M}$. In the next section we will study in more detail the invariance properties of the natural parameter gradient.

The proof of Theorem 1 will be based on the following result from linear algebra:

Lemma 1 *Let $(V, \langle \cdot, \cdot \rangle)$ be a finite-dimensional inner product space and V^* its dual space. Let $\{e_i\}_{i \in \{1,\ldots,d\}} \subset V$ (not necessarily linearly independent), G the matrix defined by $G_{ij} = \langle e_i, e_j \rangle$, $\omega \in V^*$, v such that $\langle v, \cdot \rangle = \omega(\cdot)$ and Π the projection on the space $\mathrm{span}\{e_i\}_i$. Then,*

$$\Pi(v) = \left(G^+\right)^{ij} \omega(e_j) e_i. \tag{10}$$

Proof Start by noting $\Pi(v)$ is uniquely defined by the fact that $\langle \Pi(v), w \rangle = \omega(w)$, for $w \in \mathrm{span}\{e_i\}_i$ and $\langle \Pi(v), w \rangle = 0$ for $w \in (\mathrm{span}\{e_i\}_i)^\perp$. Since the RHS of (10) lies in the span of $\{e_i\}_i$, it remains to show that for an arbitrary vector $w = w^i e_i \in \mathrm{span}\{e_i\}_i$ we have:

$$\langle \left(G^+\right)^{ij} \omega(e_j) e_i, w^k e_k \rangle = \omega(w). \tag{11}$$

Working out the LHS gives:

$$\langle \left(G^+\right)^{ij} \omega(e_j) e_i, w^k e_k \rangle = G_{ik} \left(G^+\right)^{ij} \omega(e_j) w^k \tag{12}$$

$$= w^k G_{ki} \left(G^+\right)^{ij} G_{jl} v^l \tag{13}$$

$$= w^k G_{kl} v^l \tag{14}$$

$$= w^k \omega(e_k) \tag{15}$$

$$= \omega(w), \tag{16}$$

where we use the fact that: $\omega(e_i) = G_{ij} v^j$ and the symmetry of G in the second equality.

$\square$

Proof (Proof of Theorem 1) We now let $T_p\mathcal{M}$ take the role of V, $d\mathcal{L}_p$ the role of ω, $\partial_i(\xi)$ the role of e_i, and $\mathrm{grad}_p \mathcal{L}$ the role of v. Equation (8) now follows immediately. When the tangent vectors $\{\partial_i(\xi)\}_i$ span the whole tangent space of $\mathcal{M}$ at p, Π_ξ becomes the identity on $T_p\mathcal{M}$. This gives Eq. (9). $\square$

3 Invariance properties of the natural parameter gradient

In this section we study the invariance properties of the natural parameter gradient by using an alternative parametrisation of $\mathcal{M}$ given by:

$$\psi : \Theta \ni \theta \mapsto \psi(\theta) \in \mathcal{Z}. \tag{17}$$

Note that $G^+(\xi)$, $\nabla_\xi \mathcal{L}$ and $\partial_i(\xi)$ in the definition of $d\phi_\xi \widetilde{\nabla}_\xi \mathcal{L}$ all implicitly depend on the parametrisation ϕ. For an alternative parametrisation ψ we will there-
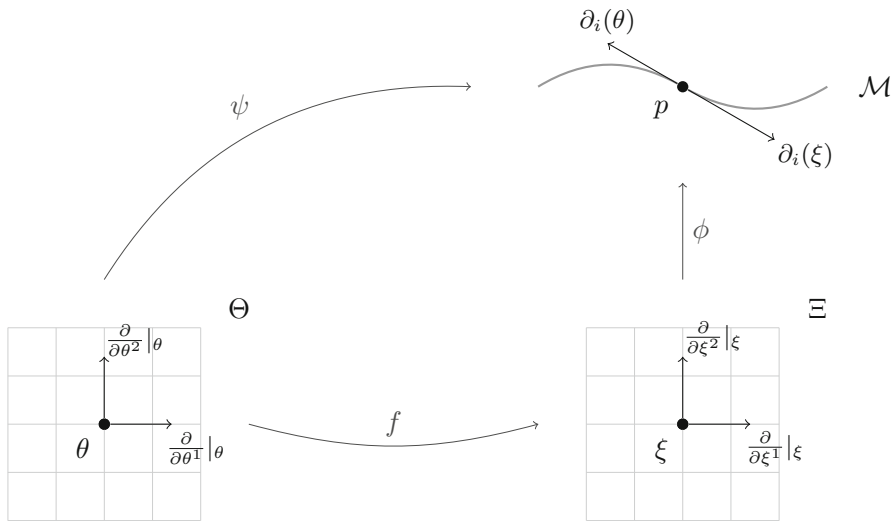
Fig. 2 Two parametrisations of $\mathcal{M}$

fore write: $\partial_i(\theta) := d\psi_\theta\left(\frac{\partial}{\partial\theta}|_\theta\right)$, $G_{ij}(\theta) := g_{\psi(\theta)}\left(\partial_i(\theta), \partial_j(\theta)\right)$, and $\nabla_\theta\mathcal{L} := (\partial_1(\theta)\mathcal{L}, ..., \partial_d(\theta)\mathcal{L})$ (see Fig. 2).

The invariance properties can be studied from the perspective of the model and from the perspective of the parameter space itself. Since the former is of more importance, we will start with this one.

3.1 Parametrisation dependence and reparametrisation invariance on the model

A parametrisation can be used to represent tangent vectors on the model space by elements of $\mathbb{R}^d$. A representation (of vectors on $\mathcal{M}$) can be interpreted as the map $\mathcal{O}\colon (\phi, \xi) \mapsto \mathcal{O}(\phi, \xi) \in T_\xi\Xi\ (\cong \mathbb{R}^d)$ that takes a parametrisation-coordinate pair and assigns a tangent vector on the parameter space to it. The natural parameter gradient defined by $\widetilde{\nabla}_\xi\mathcal{L} = \left(G^+(\xi)\nabla_\xi\mathcal{L}\right)^i \left.\frac{\partial}{\partial\xi^i}\right|_\xi$ in Eq. (5) is an example of a representation, where the dependence on ϕ on the RHS is implicit. Naively, one could define invariance of a representation in the following way:

Definition 4 (Parametrisation independence) Let $\mathcal{M}$ be a model. A representation $\mathcal{O}(\cdot, \cdot)$ is called *parametrisation independent* if for any pair ϕ, ψ of parametrisations of $\mathcal{M}$, and coordinates ξ, θ such that $\psi(\theta) = \phi(\xi)$, the following holds:

$$d\psi_\theta\mathcal{O}(\psi, \theta) = d\phi_\xi\mathcal{O}(\phi, \xi). \tag{18}$$

It turns out that this is not a very useful definition. As we will see, no non-trivial representation can be parametrisation independent in the sense of this definition. We will illustrate this in Example 1 and 2 below for the natural parameter gradient on specific models. A formal proof can be found in Appendix A.1.
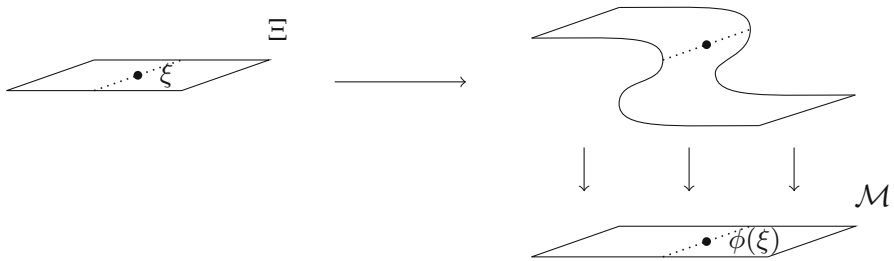
Fig. 3 Parametrisation with a non-surjective span of the parameter vectors

In order to overcome the limitation of Definition 4, we propose the following more suitable definition of invariance of a representation:

Definition 5 (Reparametrisation invariance) Let $\mathcal{M}$ be a model. A representation $\mathcal{O}(\cdot, \cdot)$ is called *reparametrisation invariant* if for any pair ϕ, ψ of parametrisations of $\mathcal{M}$, such that $\psi = \phi \circ f$ for a diffeomorphism $f : \Theta \to \Xi$, and coordinates ξ, θ such that $\theta = f^{-1}(\xi)$, the equality (18) holds.

Due to the extra requirement of the existence of the reparametrisation function f in Definition 5, we get the following central result of this paper:

Theorem 2 *The natural parameter gradient is reparametrisation invariant.*

Proof By Definition 5, we need to show that for $\psi = \phi \circ f$ and $\theta = f^{-1}(\xi)$ we have:

$$d\psi_\theta \widetilde{\nabla}_\theta \mathcal{L} = d\phi_\xi \widetilde{\nabla}_\xi \mathcal{L}. \tag{19}$$

Since the differential df_θ is surjective, we have that $\mathrm{span}\{\partial_i(\xi)\}_i = \mathrm{span}\{\partial_j(\theta)\}_j$. Therefore, by using Eq. (8) of Theorem 1, we get:

$$d\psi_\theta \widetilde{\nabla}_\theta \mathcal{L} = \Pi_\theta \left(\mathrm{grad}_{\psi(\theta)}^{\mathcal{Z}} \mathcal{L} \right) = \Pi_\xi \left(\mathrm{grad}_{\phi(\xi)}^{\mathcal{Z}} \mathcal{L} \right) = d\phi_\xi \widetilde{\nabla}_\xi \mathcal{L}, \tag{20}$$

which is what we wanted to show. □

Remark 1 Note that under the extra assumptions that $\mathcal{M}$ is a smooth manifold and all parametrisations are required to be diffeomorphisms Definitions 4 and 5 are equivalent, since the composition $f = \psi \circ \phi^{-1}$ is a diffeomorphism. These assumptions are often implicitly made when referring to the invariance of the natural gradient. However, as we will see below, this is no longer the case in our more general setting.

Example 1

This example will be of a graphical nature. Consider the parametrisation ϕ that is the composition of the 2 maps in Fig. 3. Now let $\psi : \Xi \to \mathcal{M}$ be this parametrisation but with a 90 degree rotation around $\phi(\xi)$ applied before projecting down to $\mathcal{M}$.

Note that the spans of the parameter vectors have trivial intersection as depicted in Fig. 4a. This immediately implies that for any representation we have: $d\psi_\theta \mathcal{O}(\psi, \theta) \neq$
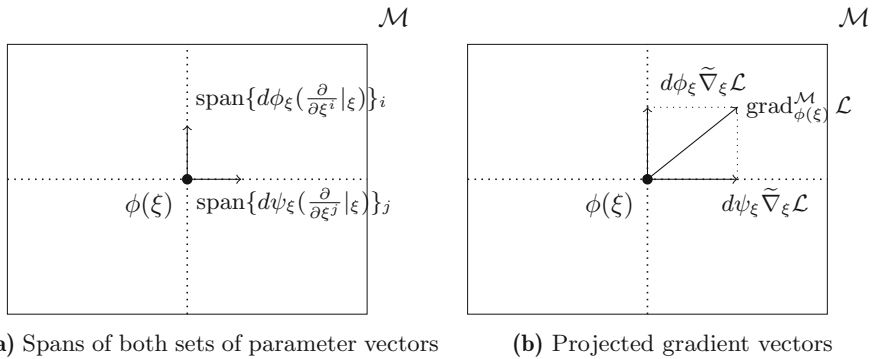
(a) Spans of both sets of parameter vectors   **(b)** Projected gradient vectors

Fig. 4 Spans and gradient vectors on the model of both parametrisations

$d\phi_\xi \mathcal{O}(\phi, \xi)$ except when both sides are equal to zero. In particular, we can let the natural gradient be as in Fig. 4b. We know from Theorem 1 that the natural parameter gradients on $\mathcal{M}$ will be the projections of the natural gradient onto the respective spans of the parameter vectors as depicted in the figure. Note that the projection should be orthogonal with respect to the inner product g_p, which we have chosen here to be Euclidean for ease of illustration.

This example shows that for non-singular points, we can construct two parametrisations that give different natural parameter gradients on the same point of the model. Note that this is not in violation of Theorem 2 since there does not exist a diffeomorphism f such that $\psi = \phi \circ f$.

Example 2

Let us consider the case in which ϕ is a smooth map from an interval on the real line to $\mathbb{R}^2$ as depicted in Fig. 5. We have that ξ_1 and ξ_2 are both mapped to the same point p in $\mathbb{R}^2$. Note that $\mathcal{M}$ is in this case not a locally embedded submanifold around p and thus p is a singular point. Note that $G(\xi_1)$ is a real number different from zero and therefore non-degenerate. Calculating the natural parameter gradient on $\mathcal{M}$ for $\xi = \xi_1$ gives:

$$d\phi_{\xi_1} \widetilde{\nabla}_{\xi_1} \mathcal{L} = G^+(\xi_1) \nabla_\xi \mathcal{L} \, \partial(\xi_1) \tag{21}$$

$$= G^{-1}(\xi_1) \frac{\partial \mathcal{L} \circ \phi}{\partial \xi}(\xi_1) \partial(\xi_1). \tag{22}$$

Since $G^{-1}(\xi_1) \frac{\partial \mathcal{L} \circ \phi}{\partial \xi}(\xi_1)$ is a scalar, the resulting vector will lie in the span of $\partial(\xi_1)$ illustrated by the blue arrows in the figure.

Now let $f \colon \Theta \to \Xi$ be a diffeomorphism such that $f(\theta_1) = \xi_2$. An alternative parametrisation of $\mathcal{M}$ is given by:
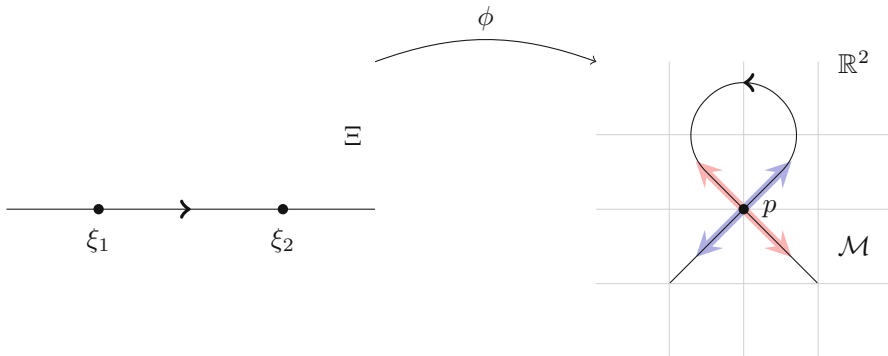
$$\psi = \phi \circ f. \tag{23}$$

Fig. 5 Parametrisation that contains a singular point

Calculating the natural parameter gradient at θ_1 for this parametrisation gives:

$$d\psi_{\theta_1} \widetilde{\nabla}_{\theta_1} \mathcal{L} = G^{-1}(\theta_1) \frac{\partial \mathcal{L} \circ \psi}{\partial \theta}(\theta_1) \partial(\theta_1). \tag{24}$$

Note that this vector is in the span of $\partial(\theta_1)$ denoted by the red arrows in the figure and therefore in general different from (22). This shows that when $\text{span}\{\partial_i(\xi)\}_i \neq \text{span}\{\partial_j(\theta)\}_j$ the outcome of $\left(G^+(\xi)\nabla_\xi \mathcal{L}\right)^i \partial_i(\xi)$ can be dependent on the choice of parametrisation and therefore the natural parameter gradient is not parametrisation independent. Note however that this result is not in contradiction with Theorem 2 since we do not have $\theta_1 = f^{-1}(\xi_1)$. See Appendix A.2 for a worked-out example of this.

A.2 Reparametrisation (in)variance on the parameter space

In the previous section we have looked at the invariance properties of the natural parameter gradient from the perspective of the model. One can also study the invariance properties from the perspective of the parameter space as is done for example in Section 12 of [12]. Translating the definition of invariance given there to our notation gives the following:

Definition 6 (Reparametrisation invariance on the parameter space) A representation $\mathcal{O}(\cdot, \cdot)$ is called *reparametrisation invariant on the parameter space* if for any pair of parametrisations ϕ, ψ such that $\psi = \phi \circ f$ for a diffeomorphism $f \colon \Theta \to \Xi$, and coordinates ξ, θ such that $\theta = f^{-1}(\xi)$, we have:

$$df_\theta \mathcal{O}(\psi, \theta) = \mathcal{O}(\phi, \xi). \tag{25}$$

Note that reparametrisation invariance on the parameter space implies reparametrisation invariance on the model as defined in Definition 5. Furthermore, it can be shown that when $\mathcal{M}$ is a smooth manifold and all parametrisations are required to be diffeomorphisms, like in Remark 1, this definition is equivalent to Definitions 4 and 5. In that
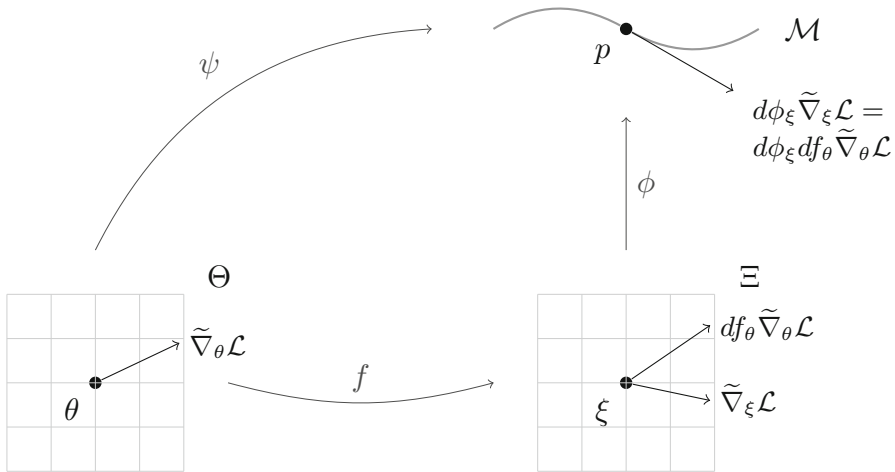
Fig. 6 Two parametrisations of $\mathcal{M}$ with different gradient vectors on the parameter space

case, the natural parameter gradient satisfies Eq. (25). As we will see below, this is not true for general ϕ. We would like to argue, however, that this is not a suitable definition of invariance, since multiple vectors on the parameter space can be mapped to the same vector on the model. Therefore inequality on the parameter space does not have to imply inequality on the model.

We will now make the above explicit. Let us choose the MP inverse as generalised inverse for $\widetilde{\nabla}_\xi \mathcal{L}$ and consider an alternative parametrisation $\psi = \phi \circ f$ for a diffeomorphism $f \colon \Theta \to \Xi$ (see Fig. 6). We denote the matrix of partial derivatives of f at θ with $F_i^j(\theta) = \frac{\partial f^j}{\partial \theta^i}(\theta)$. For $\xi = f(\theta)$ we get the following relations:

$$\partial_i(\theta) = F_i^j(\theta)\, \partial_j(\xi) \tag{26}$$

$$\nabla_\theta \mathcal{L} = F(\theta)\, \nabla_\xi \mathcal{L} \tag{27}$$

$$G(\theta) = F(\theta)\, G(\xi)\, F^T(\theta). \tag{28}$$

We map $\widetilde{\nabla}_\theta \mathcal{L}$ to $T_\xi\, \Xi$ through df_θ and get:

$$df_\theta \widetilde{\nabla}_\theta \mathcal{L} = df_\theta \left(\left(G^+(\theta)\nabla_\theta\mathcal{L}\right)^i \left.\frac{\partial}{\partial\theta^i}\right|_\theta \right) \tag{29}$$

$$= \left(\left(F(\theta)G(\xi)F^T(\theta)\right)^+ F(\theta)\, \nabla_\xi\mathcal{L}\right)^i F_i^j(\theta) \left.\frac{\partial}{\partial\xi^j}\right|_\xi \tag{30}$$

$$= \left(F^T(\theta) \left(F(\theta)G(\xi)F^T(\theta)\right)^+ F(\theta)\, \nabla_\xi\mathcal{L}\right)^j \left.\frac{\partial}{\partial\xi^j}\right|_\xi. \tag{31}$$

We will write y_Ξ, y_Θ for the coefficients of $\widetilde{\nabla}_\xi\mathcal{L}$ and $df_\theta\widetilde{\nabla}_\theta\mathcal{L}$ respectively. From Theorem 1 and the fact that $F(\theta)$ is of full rank we know that $F(\theta)\, \nabla_\xi\mathcal{L}$ lies in the

image of $F(\theta)G(\xi)F^T(\theta)$. Therefore, by the definition of the MP inverse, we have that $\left(F(\theta)G(\xi)F^T(\theta)\right)^+ F(\theta)\nabla_\xi\mathcal{L} = \arg\min_x\{||x|| : F(\theta)G(\xi)F^T(\theta)x = F(\theta)\nabla_\xi\mathcal{L}\}$, where $||\cdot||$ is the Euclidean norm on $\mathbb{R}^d$. The coefficients in (31) become:

$$y_\Theta = F^T(\theta)\left(F(\theta)G(\xi)F^T(\theta)\right)^+ F(\theta)\nabla_\xi\mathcal{L} \tag{32}$$

$$= \arg\min_y\left\{||\left(F^T(\theta)\right)^{-1} y|| : G(\xi)y = \nabla_\xi\mathcal{L}\right\}, \tag{33}$$

where we substitute $y = F^T(\theta)x$ in the last line.

Remark 2 Note that $||\left(F^T\right)^{-1}(\cdot)||$ is the pushforward of the norm on Θ through f. This shows nicely the equivalence of the gradient for on the one hand constructing a different parametrisation (ψ), and on the other hand defining a different inner product $\left(||\left(F^T\right)^{-1}(\cdot)||\right)$ for the existing parametrisation (ϕ).

Comparing the result to the natural parameter gradient on Ξ gives:

$$\widetilde{\nabla}_\xi\mathcal{L} = \left(G^+(\xi)\nabla_\xi\mathcal{L}\right)^i \left.\frac{\partial}{\partial\xi^i}\right|_\xi \tag{34}$$

$$= (y_\Xi)^i \left.\frac{\partial}{\partial\xi^i}\right|_\xi \tag{35}$$

$$y_\Xi = \arg\min_y\{||y|| : G(\xi)y = \nabla_\xi\mathcal{L}\}. \tag{36}$$

Because the norms in (33) and (36) are different, generally $y_\Theta \neq y_\Xi$. However, both satisfy $G(\xi)y = \nabla_\xi\mathcal{L}$ and therefore $G(\xi)(y_\Theta - y_\Xi) = 0$. This implies:

$$d\phi_\xi\left(df_\theta\widetilde{\nabla}_\theta\mathcal{L} - \widetilde{\nabla}_\xi\mathcal{L}\right) = (y_\Theta - y_\Xi)^i\,\partial_i(\xi) \tag{37}$$

$$= 0, \tag{38}$$

where the last equality can be verified by taking the norm on the RHS of (37) using that it is non-degenerate, like so:

$$||(y_\Theta - y_\Xi)^i\,\partial_i(\xi)||_g^2 = (y_\Theta - y_\Xi)^T G(\xi)(y_\Theta - y_\Xi) = 0. \tag{39}$$

This shows that for overparametrised systems the natural parameter gradient is not reparametrisation invariant on the parameter space. However, as implied by Theorem 1, the dependency on the parametrisation disappears when the gradient is mapped to the model. See Appendix A.3 for a worked-out example of the above discussion.

4 Practical considerations for the natural gradient method

The natural gradient method is performed by updating the current parameter vector $\xi \in \Xi$ in the direction of the vector $G^+(\xi)\nabla_\xi\mathcal{L} \in \mathbb{R}^d$. In case of a constrained

parameter space, i.e. Ξ is not the full space $\mathbb{R}^d$, such as the space of covariance matrices, one runs the risk of stepping outside the parameter space, see also [2]. This is called a constraint violation. One can use backprojection [8], addition of a penalty, and weight clipping [7] to avoid these violations. Note that for a variety of neural network applications, including many supervised learning tasks, the parameter space is unconstrained. For these models however, the generalised inverse is often hard to compute due to the high number of parameters. In this context, the Woodbury matrix identity with damping is often used instead [15]. Investigating these topics further falls outside the scope of this paper.

Reparametrisation (in)variance of the natural gradient method trajectory

We have shown in Theorem 2 that from the perspective of the model, the natural parameter gradient is reparametrisation invariant. That is, for two parametrisations ϕ, ψ for which $\psi = \phi \circ f$ for a diffeomorphism f and $\theta = f^{-1}(\xi)$, we have that,

$$d\phi_\xi \widetilde{\nabla}_\xi \mathcal{L} = \left(G^+(\xi)\nabla_\xi \mathcal{L}\right)^i \partial_i = \left(G^+(\theta)\nabla_\theta \mathcal{L}\right)^j \partial_j = d\psi_\theta \widetilde{\nabla}_\theta \mathcal{L}. \tag{40}$$

The natural gradient method is performed by updating the current parameter vector $\xi \in \mathbb{R}^d$ in the direction of the vector $G^+(\xi)\nabla_\xi \mathcal{L} \in \mathbb{R}^d$. Equation (40) implies that if we would update the parameters for both parametrisations an infinitesimal amount, this would give us the same result on the model. We would like to emphasise however that updating the parameters by a finite amount will in general result in different locations on the model. Therefore the natural gradient method trajectory is dependent on the choice of parametrisation. This is however not an issue specific to overparametrised models but with the natural gradient method in general. See Section 12 of [12] for exact bounds on the invariance.

Occurrence of non-proper points

We saw in the proof of Theorem 2 that when $\text{span}\{\partial_i(\xi)\}_i = \text{span}\{\partial_j(\theta)\}_j$ for two parametrisations ϕ and ψ with $\phi(\xi) = \psi(\theta)$, we have $d\phi_\xi \widetilde{\nabla}_\xi \mathcal{L} = d\psi_\theta \widetilde{\nabla}_\theta \mathcal{L}$. For $\phi(\xi) \in \text{Smooth}(\mathcal{M})$ note that this equality holds in particular when $\text{span}\{\partial_i(\xi)\}_i = \text{span}\{\partial_j(\theta)\}_j = T_p\mathcal{M}$, i.e. ϕ is proper in ξ. Therefore we will now study when this is the case. We start by recalling some basic facts from smooth manifold theory: Let M, N be smooth manifolds and $F : M \to N$ a smooth map. We call a point $p \in M$ a *regular point* if $dF_p : T_pM \to T_{F(p)}N$ is surjective and a *critical point* otherwise. A point $q \in N$ is called a *regular value* if all the elements in $F^{-1}(q)$ are regular points, and a *critical value* otherwise. If M is n-dimensional, we say that a subset $S \subset M$ has measure zero in M, if for every smooth chart (U, ψ) for M, the subset $\psi(S \cap U) \subset \mathbb{R}^n$ has n-dimensional measure zero. That is: $\forall \delta > 0$, there exists a countable cover of $\psi(S \cap U)$ consisting of open rectangles, the sum of whose volumes is less than δ. We have the following result based on Sard's theorem:

Proposition 1 *If* $\text{Smooth}(\mathcal{M})$ *is a manifold, then the image of the set of points for which ϕ is not proper has measure zero in* $\text{Smooth}(\mathcal{M})$.

Proof From the definition of $\text{Smooth}(\mathcal{M})$ we know that for every $p \in \text{Smooth}(\mathcal{M})$ there exists a U_p open in $\mathcal{Z}$ such that $U_p \cap \mathcal{M}$ is an embedded submanifold of $\mathcal{Z}$. Let $U := \bigcup_{p \in \text{Smooth}(\mathcal{M})} U_p$. Note that: $U \cap \mathcal{M} = \text{Smooth}(\mathcal{M})$ and therefore $\phi^{-1}(U) = \phi^{-1}(\text{Smooth}(\mathcal{M}))$. Since U is open in $\mathcal{Z}$, $\phi^{-1}(\text{Smooth}(\mathcal{M}))$ is an open subset of Ξ and thus an embedded submanifold. Therefore we can consider the map:

$$\phi|_{\phi^{-1}(\text{Smooth}(\mathcal{M}))} : \phi^{-1}(\text{Smooth}(\mathcal{M})) \to \text{Smooth}(\mathcal{M}) \tag{41}$$

and note that the image of the set of points for which ϕ is not proper is equal to the set of critical values of $\phi|_{\phi^{-1}(\text{Smooth}(\mathcal{M}))}$ in $\text{Smooth}(\mathcal{M})$. A simple application of Sard's theorem gives the result. □

This proposition implies that when $\text{Smooth}(\mathcal{M})$ is a manifold, the set of points for which the pushforward of the natural parameter gradient is unequal to the natural gradient has measure zero in $\text{Smooth}(\mathcal{M})$.

5 Conclusion

In this paper we have studied the natural parameter gradient, which was defined as the update direction of the natural gradient method, and its pushforward to the model in an overparametrised setting. We have seen that the latter is equal to the natural gradient under certain conditions. Furthermore we have proposed different notions of invariance and studied whether the natural parameter gradient satisfies these. From the perspective of the model, we have seen that the natural parameter gradient is reparametrisation invariant but that it is not parametrisation independent. Additionally, we saw that the natural parameter gradient is not reparametrisation invariant on the parameter space. We have argued, however, that this notion is less suitable in an overparametrised setting since multiple vectors on the parameter space can correspond to the same vector on the model. Finally we have given some practical considerations for the natural gradient method.

Appendix

A.1 On the limitation of Definition 4

Proposition 2 *No non-trivial representation of a vector can be parametrisation independent in the sense of Definition 4. More precisely, any representation satisfying Definition 4 has the property that for all parametrisation-coordinate pairs ϕ, ξ the following holds:*

$$d\phi_\xi \mathcal{O}(\phi, \xi) = 0. \tag{42}$$

Proof Let $\mathcal{M}$ be a model and assume that $\mathcal{O}$ is a representation satisfying Definition 4. Let ϕ be a parametrisation of $\mathcal{M}$ and $\xi^* \in \Xi$ be a fixed (arbitrary) element on the

domain of ϕ. Now consider the following function:

$$f: \mathbb{R}^d \to \mathbb{R}^d \tag{43}$$

$$\theta = \left(\theta^1, .., \theta^d\right) \mapsto \left(\left(\theta^1\right)^3, ..., \left(\theta^d\right)^3\right) + \xi^*. \tag{44}$$

We define $\Theta = f^{-1}(\Xi)$ and $\psi = \phi \circ f|_\Theta$. First note that, since f is continuous, Θ is an open set. Secondly, since f is surjective, we have $\psi(\Theta) = \mathcal{M}$. Therefore ψ is a parametrisation of $\mathcal{M}$. It is easy to see that the differential of f at $\theta = 0, df_0$, is equal to zero and therefore by the chain rule we have $d\psi_0 = d(\phi \circ f|_\Theta)_0 = d\phi_{f(0)} \circ df_0 = 0$. Furthermore we have: $\psi(0) = \phi(\xi^*)$. Therefore in order for $\mathcal{O}$ to satisfy equation (18) we need that:

$$d\phi_\xi^* \mathcal{O}(\phi, \xi^*) = d\psi_0 \mathcal{O}(\psi, 0) = 0. \tag{45}$$

Since ξ^* was chosen arbitrarily, this implies that $\mathcal{O}$ is a trivial representation. □

Remark 3 The function f used in the proof above is actually a homeomorphism since it is a continuous bijection and its inverse, given by $f^{-1}(\xi) = \left(\sqrt[3]{\xi^1 - (\xi^*)^1}, ..., \sqrt[3]{\xi^d - (\xi^*)^d}\right)$, is also continuous. The inverse is however not differentiable and therefore f is not a diffeomorphism, which is required for Definition 5.

A.2 Example calculation of parametrisation dependence on the model

We illustrate Example 2 in Sect. 3.1 with a specific calculation. Let us consider the following parametrisation:

$$\phi: (-1/8\pi, 5/8\pi) \to (\mathbb{R}^2, \bar{g}) \tag{46}$$

$$\xi \mapsto (x, y) = (\sin(2\xi), \sin(\xi)). \tag{47}$$

This gives $\xi_1 = 0, \xi_2 = \frac{1}{2}\pi$ in the above discussion. We get the following calculation for $d\phi_0 \tilde{\nabla}_0 \mathcal{L}$:

$$\partial(\xi) = d\phi_\xi \left(\frac{\partial}{\partial\xi}\bigg|_\xi\right) \tag{48}$$

$$= 2\cos(2\xi) \frac{\partial}{\partial x}\bigg|_{\phi(\xi)} + \cos(\xi) \frac{\partial}{\partial y}\bigg|_{\phi(\xi)} \tag{49}$$

$$\partial(\xi_1) = 2 \frac{\partial}{\partial x}\bigg|_{(0,0)} + \frac{\partial}{\partial y}\bigg|_{(0,0)} \tag{50}$$

$$G(\xi_1) = 2^2 + 1^1 = 5 \tag{51}$$

$$d\phi_{\xi_1} \tilde{\nabla}_{\xi_1} \mathcal{L} = \frac{1}{5} \frac{\partial}{\partial \xi}\bigg|_{\xi=0} \left(\mathcal{L}(\sin(2\xi), \sin(\xi)) \right) \left(2 \frac{\partial}{\partial x}\bigg|_{(0,0)} + \frac{\partial}{\partial y}\bigg|_{(0,0)} \right). \tag{52}$$

Now let:

$$f : (-1/8\pi, 5/8\pi) \rightarrow (-1/8\pi, 5/8\pi) \tag{53}$$

$$\theta \mapsto - \left(\theta - \frac{1}{4}\pi \right). \tag{54}$$

This implies that $\theta_1 = f^{-1}(\xi_2) = 0$. We define the alternative parametrisation $\psi = \phi \circ f$. Note that we have $\psi(\theta) = (-\sin(2\theta), \sin(\theta))$ and thus $\psi(\theta_1) = \phi(\xi_1) = (0, 0)$. A similar calculation as before gives:

$$\partial(\theta_1) = -2 \frac{\partial}{\partial x}\bigg|_{(0,0)} + \frac{\partial}{\partial y}\bigg|_{(0,0)} \tag{55}$$

$$d\psi_{\theta_1} \tilde{\nabla}_{\theta_1} \mathcal{L} = \frac{1}{5} \frac{\partial}{\partial \theta}\bigg|_{\theta=0} \left(\mathcal{L}(-\sin(2\theta), \sin(\theta)) \right) \left(-2 \frac{\partial}{\partial x}\bigg|_{(0,0)} + \frac{\partial}{\partial y}\bigg|_{(0,0)} \right). \tag{56}$$

Note that because $\partial(\xi_1) \neq \partial(\theta_1)$, (52) and (56) are not equal to each other. We can therefore conclude that the natural parameter gradient is not parametrisation independent.

A.3 Example calculation of reparametrisation (in)variance on the parameter space

We illustrate the discussion in Section A.2 with a specific calculation. Let us consider the following setting:

$$\mathcal{Z} = \Xi = \Theta = \mathbb{R}^2 \tag{57}$$

$$\phi(\xi_1, \xi_2) = (\xi_1 + \xi_2, 0) \tag{58}$$

$$f(\theta_1, \theta_2) = (2\theta_1, \theta_2) \tag{59}$$

$$\mathcal{L}(x, y) = x^2 \tag{60}$$

Plugging this into the expressions derived above gives:

$$\psi(\theta_1, \theta_2) = (2\theta_1 + \theta_2, 0) \tag{61}$$

$$\partial_1(\xi) = \partial_2(\xi) = \frac{\partial}{\partial x}\bigg|_{\phi(\xi)} \qquad F(\theta) = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \tag{62}$$

$$\partial_1(\theta) = 2 \frac{\partial}{\partial x}\bigg|_{\psi(\theta)} \qquad G(\xi) = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \tag{63}$$

$$\partial_2(\theta) = \frac{\partial}{\partial x}\bigg|_{\psi(\theta)} \qquad G(\theta) = \begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix} \tag{64}$$

$$\nabla_\xi \mathcal{L} = (2(\xi_1 + \xi_2), 2(\xi_1 + \xi_2)) \tag{65}$$

$$\nabla_\theta \mathcal{L} = (4(2\theta_1 + \theta_2), 2(2\theta_1 + \theta_2)). \tag{66}$$

Now we fix $\theta = (1, 1)$ and $\xi = f(\theta) = (2, 1)$. We start by computing y_Ξ. From the above we know that:

$$y_\Xi = \arg\min_y \{||y|| : G(\xi)y = \nabla_\xi \mathcal{L}\}. \tag{67}$$

It can be easily verified that this gives $y_\Xi = (3, 3)$. For y_Θ we get:

$$y_\Theta = \arg\min_y \{|| \left(F^T(\theta) \right)^{-1} y|| : G(\xi)y = \nabla_\xi \mathcal{L}\}, \tag{68}$$

which gives: $y_\Theta = (4\frac{4}{5}, 1\frac{1}{5})$. Evidently we have $y_\Xi \neq y_\Theta$. Note however that when we map the difference of the two gradient vectors from $T_{(2,1)}\Xi$ to $T_{(3,0)}\mathcal{M}$ through $d\phi_{(2,1)}$ we get:

$$d\phi_\xi \left(df_\theta \widetilde{\nabla}_\theta \mathcal{L} - \widetilde{\nabla}_\xi \mathcal{L} \right) = (y_\Theta - y_\Xi)^i \, \partial_i(\xi) \tag{69}$$

$$= (4\frac{4}{5} - 3)\frac{\partial}{\partial x}|_{(3,0)} + (1\frac{1}{5} - 3)\frac{\partial}{\partial x}|_{(3,0)} \tag{70}$$

$$= 0. \tag{71}$$

This shows that the natural parameter gradient is in general not reparametrisation invariant on the parameter space, but that the dependency on the parametrisation disappears when the gradient is mapped to the model.

Declarations

Conflict of interest Nihat Ay is the Editor-in-Chief of the journal. He was not involved in the peer review or handling of the manuscript. Furthermore, he is the corresponding author's PhD thesis advisor and an external PhD co-advisor of Johannes Müller. On behalf of all authors, the corresponding author states that there is no other potential conflict of interest to declare.

References

1. Amari, S.-I.: Natural gradient works efficiently in learning. Neural Comput. **10**(2), 251–276 (1998)
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International Conference on Machine Learning, pp. 214–223. PMLR (2017)
3. Ay, N.: On the locality of the natural gradient for learning in deep Bayesian networks. Info. Geo. (2020). https://doi.org/10.1007/s41884-020-00038-y
4. Ay, N., Montúfar, G., Rauh, J.: Selection criteria for neuromanifolds of stochastic dynamics. In: Advances in Cognitive Neurodynamics, pp. 147–154. Springer (2013)
5. Bernacchia, A., Lengyel, M., Hennequin, G.: Exact natural gradient in deep linear networks and its application to the nonlinear case. In: Bengio, S., et al. (eds.) Advances in Neural Information Processing Systems, vol. 31. Curran Associates, Inc. (2018). https://proceedings.neurips.cc/paper/2018/file/7f018eb7b301a66658931cb8a93fd6e8Paper.pdf
6. Grosse, R., Martens, J.: A Kronecker-factored approximate Fisher matrix for convolution layers. In: International Conference on Machine Learning, pp. 573–582. PMLR (2016)
7. Gulrajani, I., et al.: Improved training of Wasserstein GANs. In: Guyon, I. et al. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017). https://proceedings.neurips.cc/paper/2017/file/892c3b1c6dccd52936e27cbd0ff683d6-Paper.pdf
8. Kushner, H.J., Clark, D.S.: Stochastic Approximation Methods for Constrained and Unconstrained Systems. Springer, New York (1978). https://doi.org/10.1007/978-1-4684-9352-8
9. Li, W., Montúfar, G.: Natural gradient via optimal transport. Inf. Geom. **1**(2), 181–214 (2018)
10. Lin, W., et al.: Tractable structured natural gradient descent using local parameterizations. In: International Conference on Machine Learning, pp. 6680–6691. PMLR (2021)
11. Malagò, L., Montrucchio, L., Pistone, G.: Wasserstein Riemannian geometry of Gaussian densities. Inf. Geom. **1**(2), 137–179 (2018)
12. Martens, J.: New insights and perspectives on the natural gradient method. J. Mach. Learn. Res. **21**, 1–76 (2020)
13. Ollivier, Y.: Riemannian metrics for neural networks I: feedforward networks. Inf. Inference J. IMA **4**(2), 108–153 (2015)
14. van Oostrum, J., Ay, N.: Parametrisation independence of the natural gradient in overparametrised systems. In: International Conference on Geometric Science of Information, pp. 726– 735. Springer (2021)
15. Pal Singh, S., Alistarh, D.: WoodFisher: efficient second-order approximation for neural network compression. Eprint: arXiv:2004.14340 (2020)
16. Van Hasselt, H.: Reinforcement learning in continuous state and action spaces. In: Reinforcement learning, pp. 207–251. Springer (2012)
17. Várady, C., et al.: Natural wake-sleep algorithm. In: NeurIPS. Deep Learning through Information Geometry Worksop (2020)
18. Zhang, G., Martens, J., Grosse, R.: Fast convergence of natural gradient descent for overparameterized neural networks. arXiv preprint arXiv:1905.10961 (2019)