Parametrisation Independence of the Natural Gradient in Overparametrised Systems

Jesse van Oostrum* Nihat Ay*

Hamburg University of Technology, 21073 Hamburg, Germany

Introduction

Within the field of deep learning, gradient methods have become ubiquitous tools for parameter optimisation. The natural gradient method, first proposed by Amari [1], is an efficient method for performing gradient descent. It is an active field of study within information geometry [2] and has been shown to be extremely effective in many applications [3, 9, 10]. 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 [12]. The natural gradient is calculated by multiplying the ordinary gradient by the inverse of the Fisher Information Matrix (FIM) associated with the statistical manifold. For high-dimensional parameter spaces, this inversion is computationally expensive but different solutions have been proposed [7, 5, 2]. In many practical applications of machine learning, and in particular deep learning, one deals with overparametrised models, in which different configurations of the parameters correspond to the same output distribution. This causes the FIM to be degenerate. In this case, one uses a generalised inverse of the FIM to calculate the natural gradient [4]. The Moore-Penrose (MP) inverse is the canonical choice for this. 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 independence of the natural gradient [7], and thus potentially the performance of the natural gradient method.

In the following we outline that for overparametrised models parametrisation independence is not affected when using a generalised, and in particular MP, inverse. It turns out that for non-singular points on the manifold, a generalised inverse does not introduce parametrisation dependence. Furthermore, we will discuss that for singular points parametrisation independence is not even guaranteed for non-overparametrised models. The proofs of the results are omitted here but can be found in [8].

*The authors acknowledge the support of the Deutsche Forschungsgemeinschaft Priority Programme "The Active Self" (SPP 2134).

Parametrisation (in)dependence of the natural gradient

Let $(\mathcal{Z}, g)$ be a Riemannian manifold, $\Xi \subset \mathbb{R}^d$ a smooth manifold of parameters, $\phi : \Xi \to \mathcal{Z}$ a smooth map (taking the role of a parametrisation), $\mathcal{M} \equiv \phi(\Xi) \subset \mathcal{Z}$ a model, and $\mathcal{L} : \mathcal{M} \to \mathbb{R}$ a smooth (loss) function. 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 $\text{Smooth}(\mathcal{M})$. A point p is called *singular* if it is not non-singular. The gradient of $\mathcal{L}$ at $p \in \text{Smooth}(\mathcal{M})$ is defined implicitly as follows:

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

By the Riesz representation theorem, this defines the gradient uniquely. We define the pushforward of the tangent vector on the parameter space through the parametrisation as $\partial_i(\xi) \equiv d\phi_\xi \left(\frac{\partial}{\partial \xi^i}|_\xi\right)$, and the matrix[1] $G(\xi)$ be such that $G_{ij}(\xi) = g_{\phi(\xi)}\left(\partial_i(\xi), \partial_j(\xi)\right)$. We denote the vector of coordinate derivatives with $\nabla_{\partial(\xi)}\mathcal{L} \equiv (\partial_1(\xi)\mathcal{L}, ..., \partial_d(\xi)\mathcal{L}) \in \mathbb{R}^d$. Furthermore, following the Einstein summation convention, we write $a^i b_i$ for the sum $\sum_i a^i b_i$.

Gradient in non-singular points, $p \in \text{Smooth}(\mathcal{M})$

We assume that ϕ is a *proper parametrisation* such that for all ξ for which $\phi(\xi) \in \text{Smooth}(\mathcal{M})$ the following holds: $\text{span}(\{\partial_1(\xi), ..., \partial_d(\xi)\}) = T_{\phi(\xi)}\mathcal{M}$.

Definition 1. *A generalised inverse of a matrix A, denoted A^+, is a matrix satisfying the following property:*

$$AA^+A = A. \tag{2}$$

Note that this definition implies that if w is in the image of A, i.e. $w = Av$, then:

$$AA^+w = AA^+Av = Av = w. \tag{3}$$

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

Theorem 1. *For $\xi \in \Xi$ such that $\phi(\xi) = p \in \text{Smooth}(\mathcal{M})$, the gradient of $\mathcal{L}$ at p can be calculated as follows:*

$$\text{grad}_p \mathcal{L} = \left(G^+(\xi)\nabla_{\partial(\xi)}\mathcal{L}\right)^i \partial_i(\xi). \tag{4}$$

From this theorem we can conclude that the natural gradient of a loss function $\mathcal{L}$, which is itself defined independently of any parametrisation in Equation (1), can be obtained by applying a generalised inverse of the FIM to the vector of coordinate derivatives $\nabla_{\partial(\xi)}\mathcal{L}$. In particular this proves that the MP inverse preserves the parameter independence in overparametrised models.

[1] Note that this becomes the FIM when g is the Fisher metric.

Note that although the calculation of the gradient vector from the perspective of the manifold is independent of the parametrisation, the final result of the gradient step of the natural gradient method might still depend on this. This is because the direction of the gradient step is only at the point p parallel to the gradient vector, but will in general not follow the gradient flow exactly after leaving point p. This is however not an issue specific to overparametrised models but with the natural gradient in general. See Section 12 of [6] for exact bounds on the invariance.

Now we will look at the behaviour of the natural gradient on the parameter space. Let us denote the gradient of $\mathcal{L}$ on the parameter space Ξ at a point ξ as follows:

$$\operatorname{grad}_\xi^\Xi \mathcal{L} = \left(G^+(\xi)\nabla_{\partial(\xi)}\mathcal{L}\right)^i \frac{\partial}{\partial \xi^i}|_\xi \in T_\xi \Xi. \tag{5}$$

It turns out that $\operatorname{grad}_\xi^\Xi \mathcal{L}$ *does* depend on the choice of parametrisation, when G^+ is the MP inverse of G. However, this dependence disappears when $\operatorname{grad}_\xi^\Xi \mathcal{L}$ is mapped to $T_{\phi(\xi)}\mathcal{M}$ through $d\phi$.

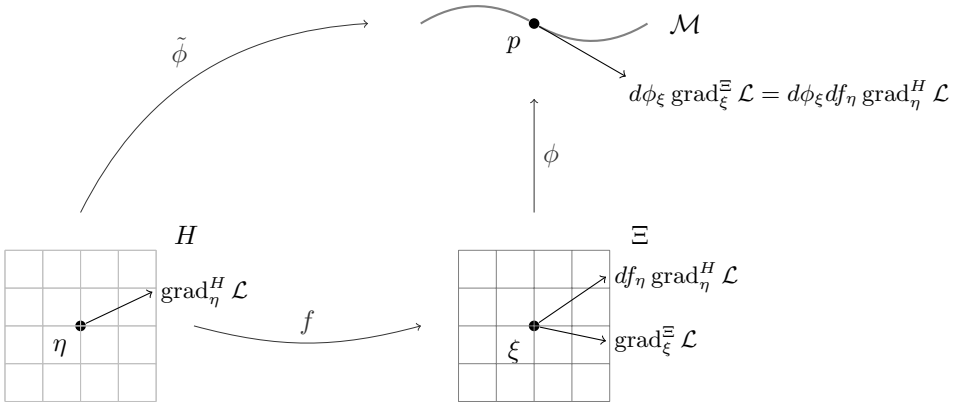

Figure 1: Two parametrisations of $\mathcal{M}$ with different gradient vectors on the parameter space

Let us consider an alternative parametrisation $\tilde{\phi} : H \ni \eta \mapsto \tilde{\phi}(\eta) \in \mathcal{M}$ such that $\tilde{\phi} = \phi \circ f$ for a diffeomorphism $f : H \to \Xi$ (see Figure 1). In order to compare the gradients on both parameter spaces, we use f to map $\operatorname{grad}_\eta^H \mathcal{L}$ to $T_\xi \Xi$. We can define the difference vector:

$$\Delta \operatorname{grad} \equiv \operatorname{grad}_\xi^\Xi \mathcal{L} - df_\eta \operatorname{grad}_\eta^H \mathcal{L} \tag{6}$$

It can be shown that $d\phi_\xi \Delta \operatorname{grad} = 0$. We conclude that although there is dependence on the parameter space itself, this dependence disappears when the gradient vectors are mapped to the manifold.

Gradient in singular points, $p \notin \text{Smooth}(\mathcal{M})$

It can be shown that for $p \notin \text{Smooth}(\mathcal{M})$, even for a non-degenerate FIM the procedure for calculating the gradient described in Theorem 1 gives a result that is parametrisation-dependent.

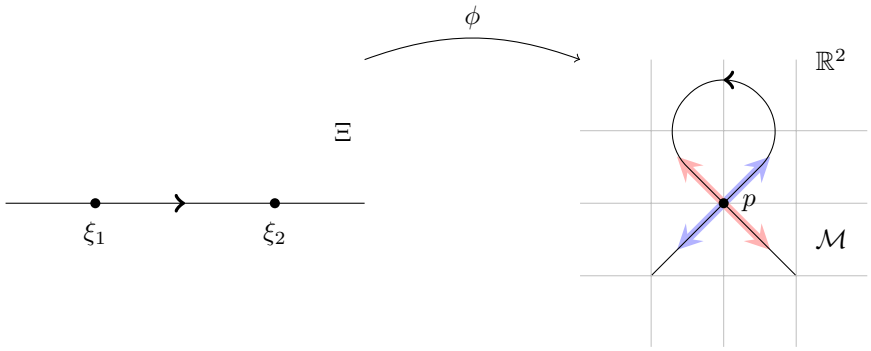

Figure 2: Example parametrisation that contains a singular point

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 Figure 2. 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. For $p \notin \text{Smooth}(\mathcal{M})$, we will denote the RHS of (4) with $\overline{\text{grad}}_p^{\partial(\xi)}\mathcal{L}$, to distinguish it from $\text{grad}_p\mathcal{L}$ which is only defined for non-singular points. It is straightforward to show that $\overline{\text{grad}}_p^{\partial(\xi_1)}\mathcal{L}$ is an element of the vector space spanned by the blue arrows. Now choosing a different parametrisation $\tilde{\phi} = \phi \circ f$, such that $f(\xi_1) = \xi_2$, will give us a gradient vector in the space spanned by the red arrows. This shows that for two different paramterisations, we get different gradient vectors and therefore the gradient is not parametrisation independent.

Conclusion

We have outlined that for non-singular points, the natural gradient of a loss function on an overparametrised model is parametrisation-independent. Subsequently, it turned out that in singular points, there is no guarantee for the natural gradient to be parametrisation-independent, even in non-overparametrised models. So far, we have only looked at the case for which the FIM is known. In practice, one often has to use the empirical FIM based on the available data. When inverting this matrix it can be beneficial to apply Tikhonov regularisation [6]. Note that by letting the regularisation parameter go to zero, the MP inverse is obtained. Also only one type of singularity has been discussed. A further direction of investigation is the behaviour of the natural gradient on different types of singular points on the manifold, see also [11].

References

[1] Shun-Ichi Amari. "Natural gradient works efficiently in learning". In: *Neural computation* 10.2 (1998), pp. 251–276.

[2] Nihat Ay. "On the locality of the natural gradient for learning in deep Bayesian networks". In: *Information Geometry* (2020), pp. 1–49.

[3] Nihat Ay, Guido Montúfar, and Johannes Rauh. "Selection criteria for neuromanifolds of stochastic dynamics". In: *Advances in Cognitive Neurodynamics*. Springer, 2013, pp. 147–154.

[4] Alberto Bernacchia, Mate Lengyel, and Guillaume Hennequin. "Exact natural gradient in deep linear networks and application to the nonlinear case". In: NIPS. 2019.

[5] Roger Grosse and James Martens. "A kronecker-factored approximate fisher matrix for convolution layers". In: *International Conference on Machine Learning*. PMLR. 2016, pp. 573–582.

[6] James Martens. "New insights and perspectives on the natural gradient method". In: *arXiv preprint arXiv:1412.1193* (2014).

[7] Yann Ollivier. "Riemannian metrics for neural networks I: feedforward networks". In: *Information and Inference: A Journal of the IMA* 4.2 (2015), pp. 108–153.

[8] Jesse van Oostrum and Nihat Ay. "Parametrisation Independence of the Natural Gradient in Overparametrised Systems". (in press).

[9] Hado Van Hasselt. "Reinforcement learning in continuous state and action spaces". In: *Reinforcement learning*. Springer, 2012, pp. 207–251.

[10] Csongor Várady et al. "Natural wake-sleep algorithm". In: *arXiv preprint arXiv:2008.06687* (2020).

[11] Sumio Watanabe. *Algebraic geometry and statistical learning theory*. Cambridge university press, 2009.

[12] Guodong Zhang, James Martens, and Roger Grosse. "Fast convergence of natural gradient descent for overparameterized neural networks". In: *arXiv preprint arXiv:1905.10961* (2019).