Invariance Properties of the Natural Gradient in Overparametrised Systems[1]

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

[1]Hamburg University of Technology, Institute for Data Science Foundations, 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

[1]Jesse van Oostrum, Johannes Müller, and Nihat Ay. "Invariance properties of the natural gradient in overparametrised systems". In: *Information Geometry* (2022), pp. 1–17

Introduction

The natural gradient and its invariance properties are usually studied in the setting where the parametrisation is a diffeomorphism.

The systems studied in deep learning are often overparametrised. This forces us to loosen the assumptions on the parametrisation. This is thought to affect parametrisation invariance of the natural gradient.[2]

In this presentation:

- Slowly relax the assumptions on the parametrisation, from diffeomorphism $\rightarrow$ smooth function, see the consequences.
- Define and study parametrisation invariance in this more general setting.

[2] 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

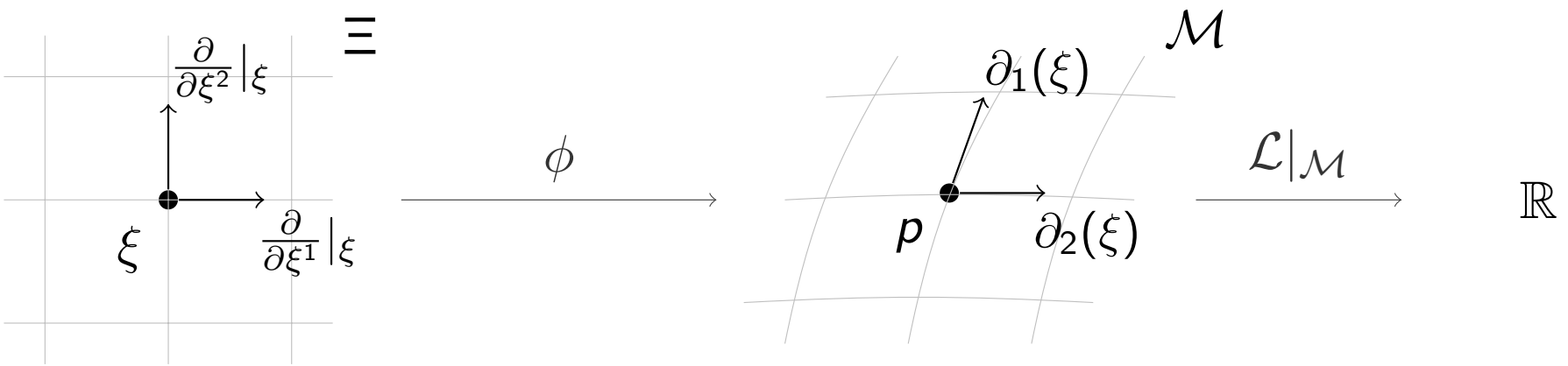
Overview

1. Notation

2. Assumptions on parametrisation
   - Diffeomorphism
   - Diffeomorphism onto its image
   - Local diffeomorphism onto its image
   - Smooth map

3. Definitions of invariance for general parametrisations

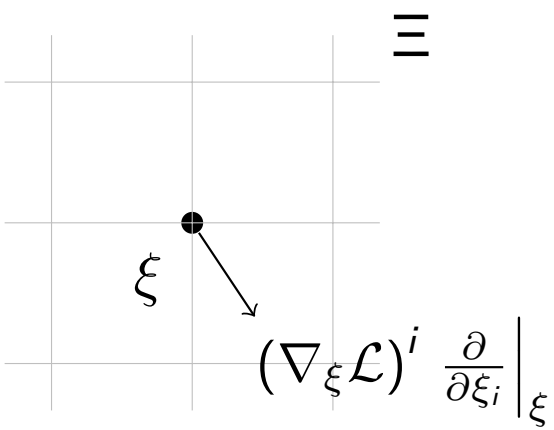
Statistical model

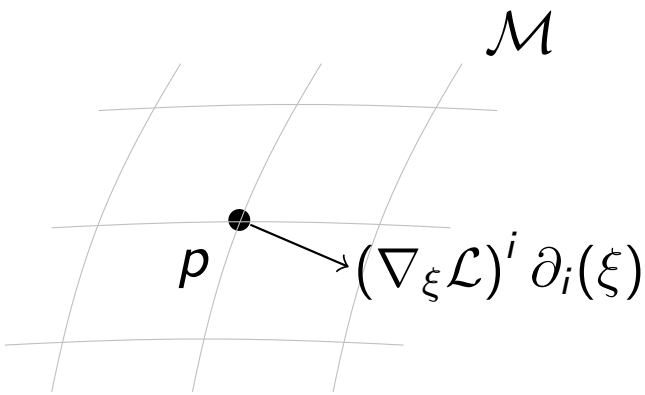


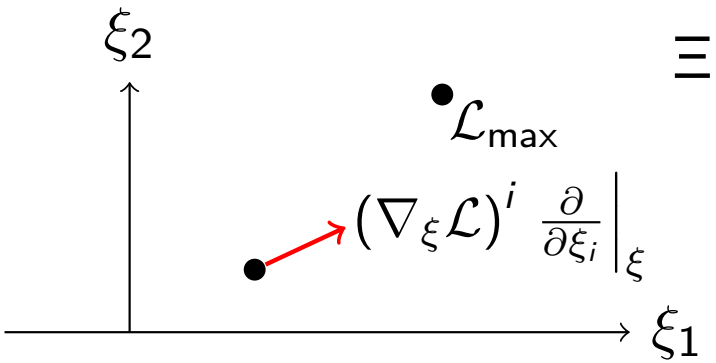
$p = \phi(\xi)$	
$\Xi \subset \mathbb{R}^d$ (open)	parameter space
$(\mathcal{Z}, g)$	Riemannian manifold of probability distributions
$\phi : \Xi \to \mathcal{Z}$	parametrisation (local diffeomorphism onto its image)
$\mathcal{M} = \phi(\Xi) \subset \mathcal{Z}$	statistical model
$\mathcal{L} : \mathcal{Z} \to \mathbb{R}$	objective function
$\frac{\partial}{\partial \xi^i}\|_\xi$	tangent vector on the parameter space Ξ
$\partial_i(\xi) = d\phi_\xi \left(\frac{\partial}{\partial \xi^i}\|_\xi \right)$	tangent vector on the manifold
$\nabla_\xi \mathcal{L} = \left(\frac{\partial \mathcal{L} \circ \phi}{\partial \xi^1}(\xi), ..., \frac{\partial \mathcal{L} \circ \phi}{\partial \xi^d}(\xi) \right) \in \mathbb{R}^d$	vector of coordinate derivatives

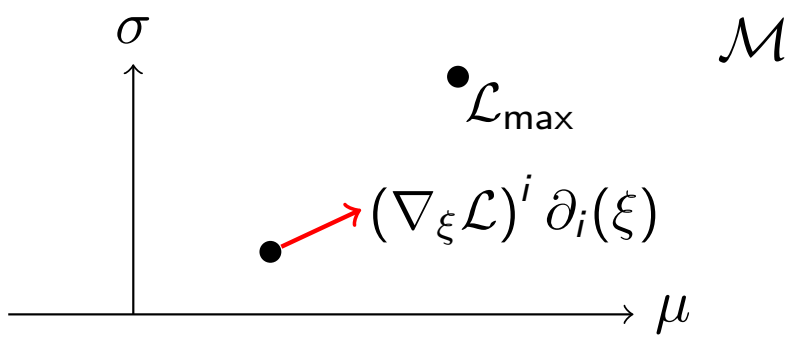
Standard gradient optimisation

For standard gradient optimisation[3], we distinguish two different vectors:









[3]The result is dependent on the choice of parametrisation of the manifold $\mathcal{M}$.

Natural gradient

Definition (Riemannian gradient)

The *Riemannian gradient* of a function $\mathcal{L} : \mathcal{Z} \to \mathbb{R}$ at a point $p \in \mathcal{Z}$ is defined implicitly through Riesz representation theorem as follows:

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

Note that this definition makes no reference to a parametrisation.

Definition (Natural gradient)

The *natural gradient*, denoted $\text{grad}_p^{\mathcal{M}}\,\mathcal{L}$, is a gradient field on $\mathcal{M}$ and is equal to the Riemannian gradient of $\mathcal{L}$.

Natural gradient

Given a basis $\{\partial_1(\xi), ..., \partial_d(\xi)\}$ of the vector space $T_p\mathcal{M}$, the natural gradient can be calculated as follows:

$$\text{grad}_p^{\mathcal{M}} \mathcal{L} = \left(G^{-1}(\xi)\nabla_\xi \mathcal{L}\right)^i \partial_i(\xi). \tag{2}$$

Where G is the Gram matrix, given by:

$$G_{ij}(\xi) = g_{\phi(\xi)}\left(\partial_i(\xi), \partial_j(\xi)\right) \tag{3}$$

The matrix G^{-1} rescales the vector used in standard gradient optimisation.

Natural parameter gradient

The corresponding vector on the parameter space is called the *natural parameter gradient*:

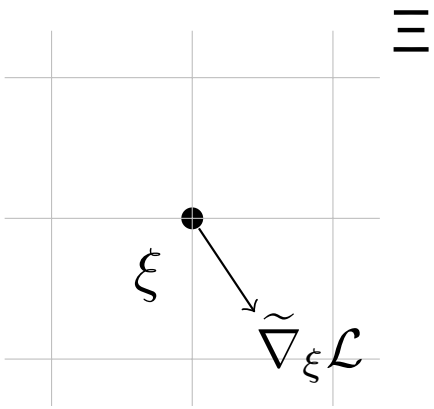
$$\widetilde{\nabla}_\xi \mathcal{L} = \left(G^{-1}(\xi)\nabla_\xi \mathcal{L}\right)^i \left.\frac{\partial}{\partial \xi^i}\right|_\xi \tag{4}$$

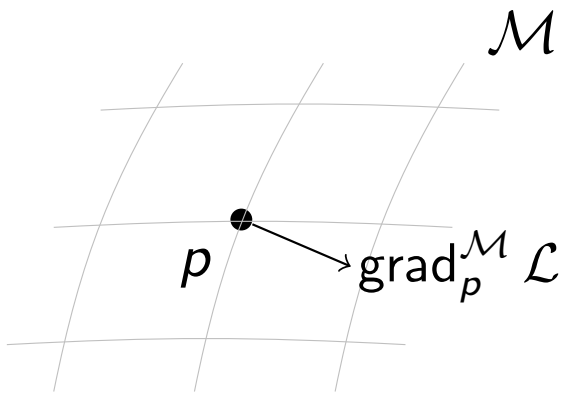
We have

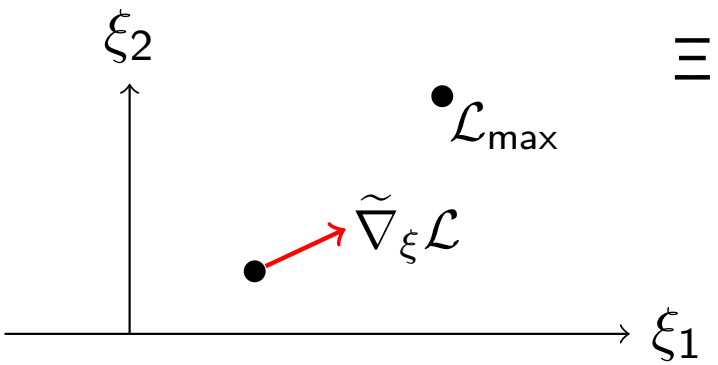
$$d\phi_\xi \widetilde{\nabla}_\xi \mathcal{L} = \operatorname{grad}_p^{\mathcal{M}} \mathcal{L}. \tag{5}$$

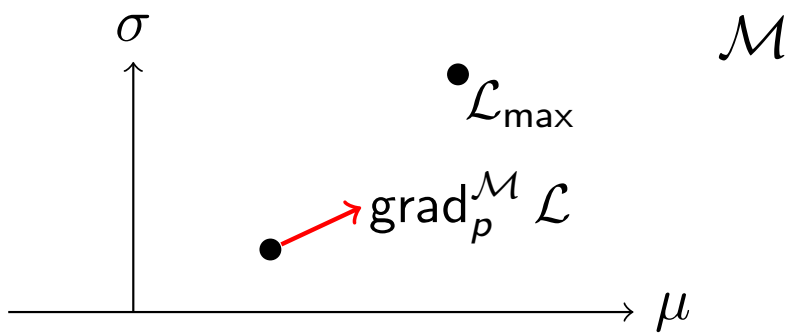
Does (5) remain true when loosening the assumptions on the parametrisation?

Natural gradient vs natural parameter gradient



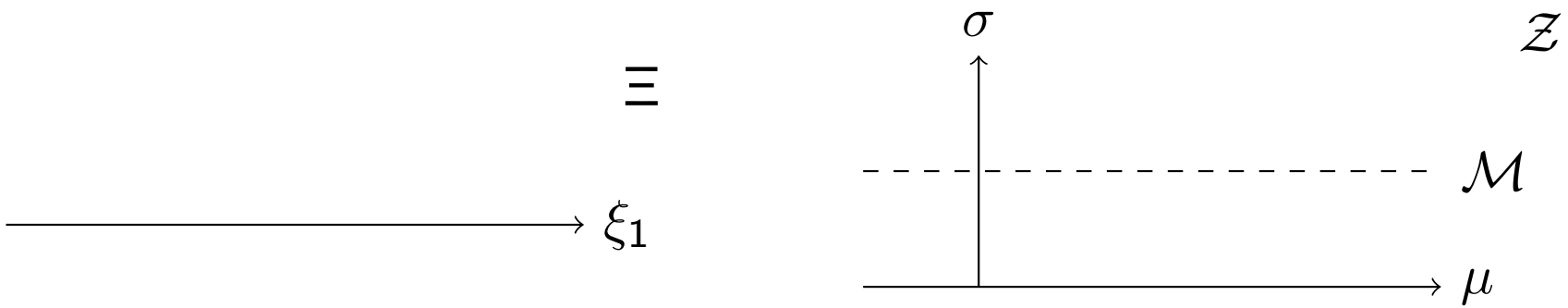






ϕ is a diffeomorphism onto its image

Consequence: $\mathcal{Z} \neq \mathcal{M}$



Definition (Natural gradient)

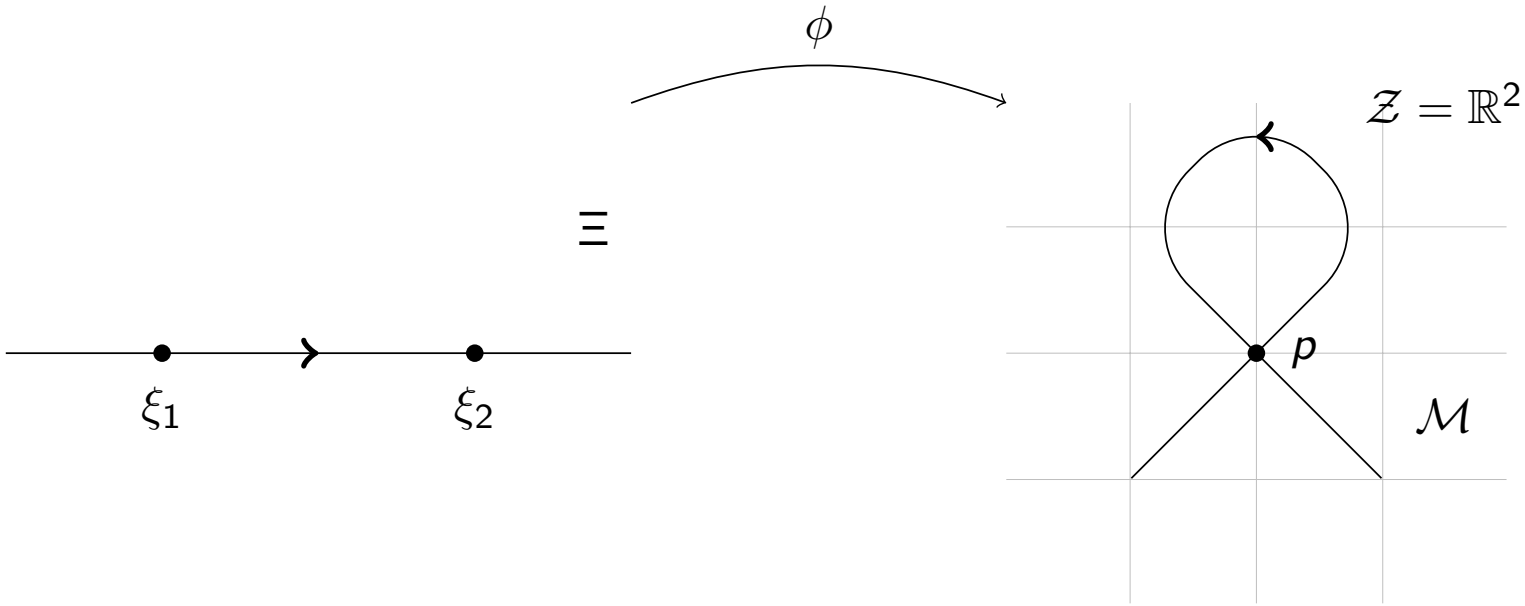
The *natural gradient*, denoted $\operatorname{grad}_p^{\mathcal{M}} \mathcal{L}$, is a gradient field on $\mathcal{M}$ and is equal to the Riemannian gradient of $\mathcal{L}|_{\mathcal{M}}$.

$$\operatorname{grad}_p^{\mathcal{M}} \mathcal{L} = \Pi_p(\operatorname{grad}_p^{\mathcal{Z}} \mathcal{L}), \tag{6}$$

where Π_p is the projection onto $T_p\mathcal{M}$.

ϕ is a local diffeomorphism onto its image

Consequence: $\mathcal{M}$ is no longer a smooth manifold.



Definition

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.

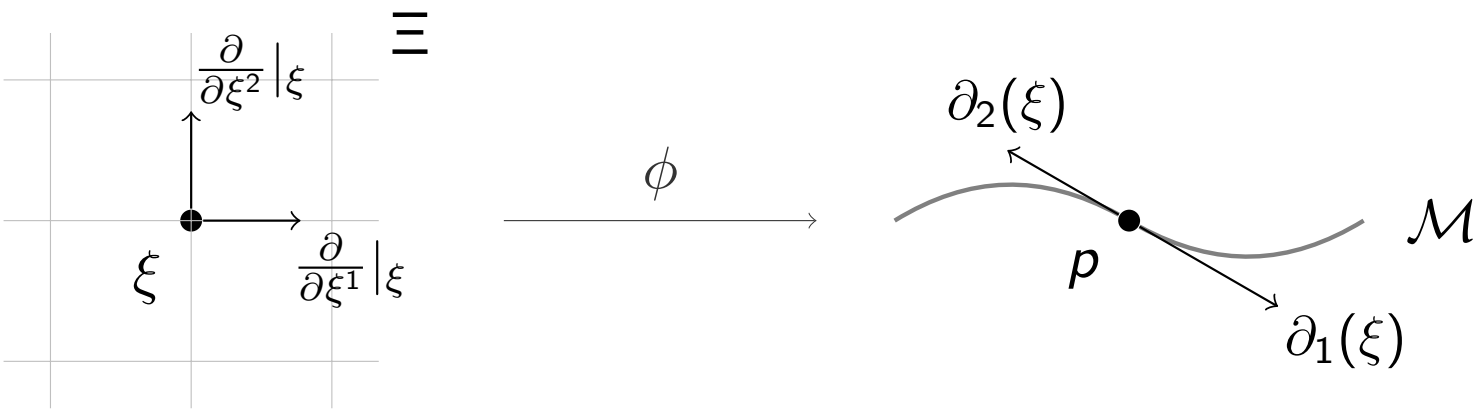
Definition (Natural gradient)

For $p \in \text{Smooth}(\mathcal{M})$, the *natural gradient*, denoted $\text{grad}_p^{\mathcal{M}} \mathcal{L}$, is a gradient field on $\mathcal{M}$ and is equal to the Riemannian gradient of $\mathcal{L}|_{\mathcal{M}}$.

ϕ is a smooth map

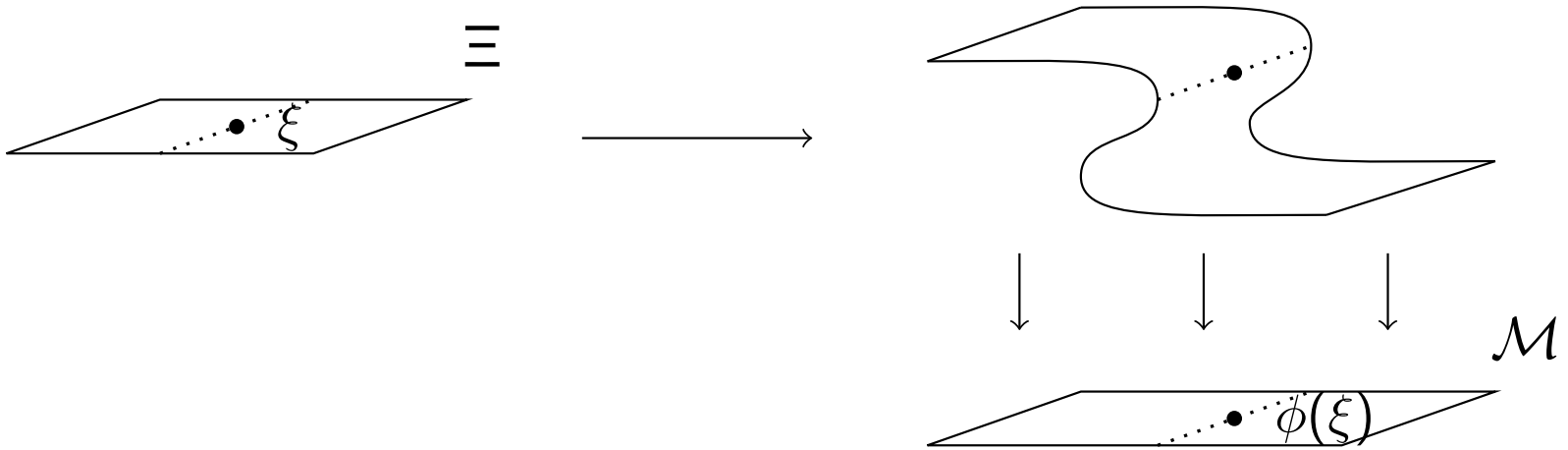
Consequences:

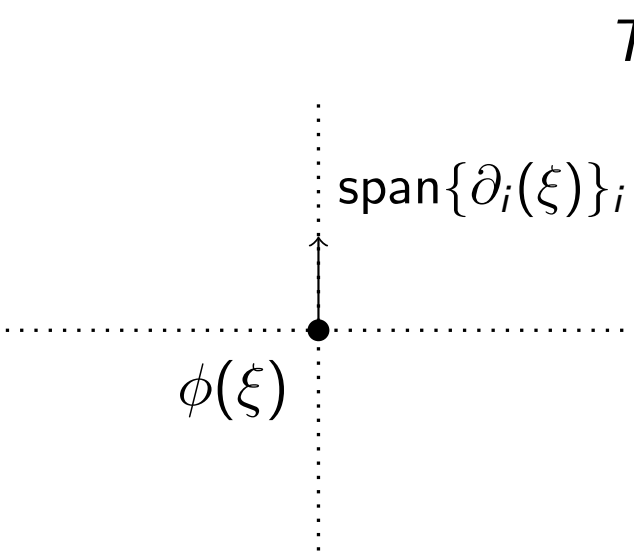
- $\mathcal{M}$ can have lower dimension than Ξ.
- Tangent vectors $\partial_i(\xi)$ are no longer linearly independent. The Gram matrix $G(\xi)$ is degenerate.



consequences (continued)

- The tangent vectors $\partial_i(\xi)$ do no longer span the tangent space $T_p\mathcal{M}$.





Definition (Generalised inverse)

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

$$AA^+A = A. \tag{7}$$

Note that the generalised inverse is in not uniquely determined for degenerate matrices.

Definition (Natural parameter gradient)

For a fixed choice of generalised inverse, the *natural parameter gradient* is defined to be the following vector on the parameter space:

$$\widetilde{\nabla}_\xi \mathcal{L} = \left(G^+(\xi)\nabla_\xi \mathcal{L}\right)^i \left.\frac{\partial}{\partial \xi^i}\right|_\xi . \tag{8}$$

When is the pushforward of the natural parameter gradient equal to the natural gradient?

Theorem

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

$$d\phi_\xi \widetilde{\nabla}_\xi \mathcal{L} = \Pi_\xi \left(\operatorname{grad}_p^{\mathcal{Z}} \mathcal{L} \right), \tag{9}$$

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

$$d\phi_\xi \widetilde{\nabla}_\xi \mathcal{L} = \operatorname{grad}_p^{\mathcal{M}} \mathcal{L}. \tag{10}$$

Invariance properties

The natural gradient is defined independently of a parametrisation. It therefore makes no sense to talk about parametrisation invariance of the natural gradient. One can however talk about parametrisation invariance of the natural parameter gradient, which *is* defined in terms of a parametrisation.

A parametrisation can be used to represent tangent vectors on the model space by elements of $\mathbb{R}^d$. A *representation* (of a vector 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 $\widetilde{\nabla}_\xi \mathcal{L} = \left(G^+(\xi) \nabla_\xi \mathcal{L}\right)^i \left.\frac{\partial}{\partial \xi^i}\right|_\xi$ is an example of a representation.

Definitions of invariance

Definition (Parametrisation independence)

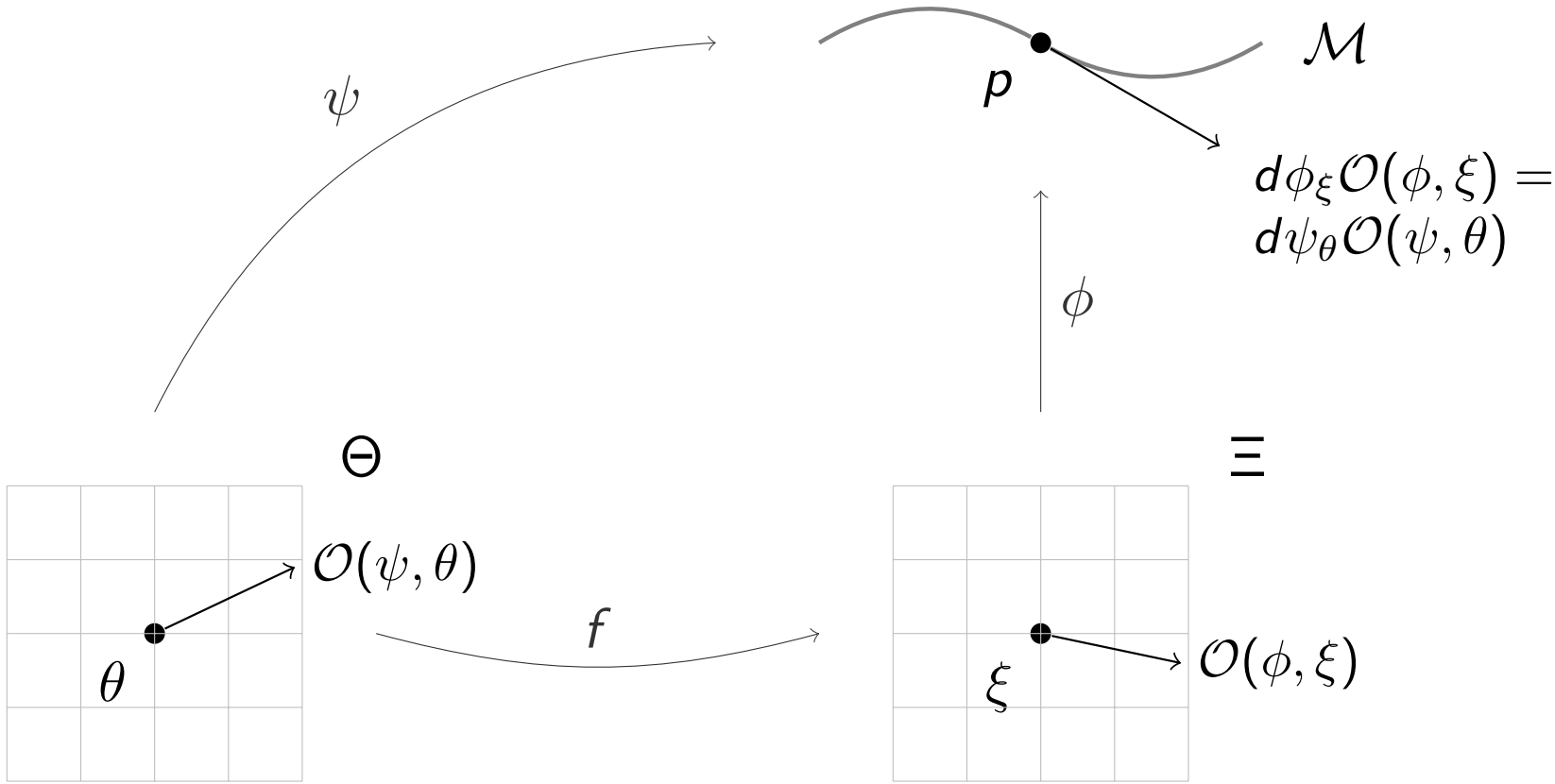
Let $\mathcal{M}$ be a model. A representation $\mathcal{O}(\cdot,\cdot)$ is called *parametrisation independent* if for any pair ϕ, ψ of parametrisations of $\mathcal{M}$, and coordinates ξ, θ such that $\psi(\theta) = \phi(\xi)$, the following holds:

$$d\psi_\theta \mathcal{O}(\psi, \theta) = d\phi_\xi \mathcal{O}(\phi, \xi). \tag{11}$$

No non-trivial representation can be parametrisation independent in the sense of this definition.

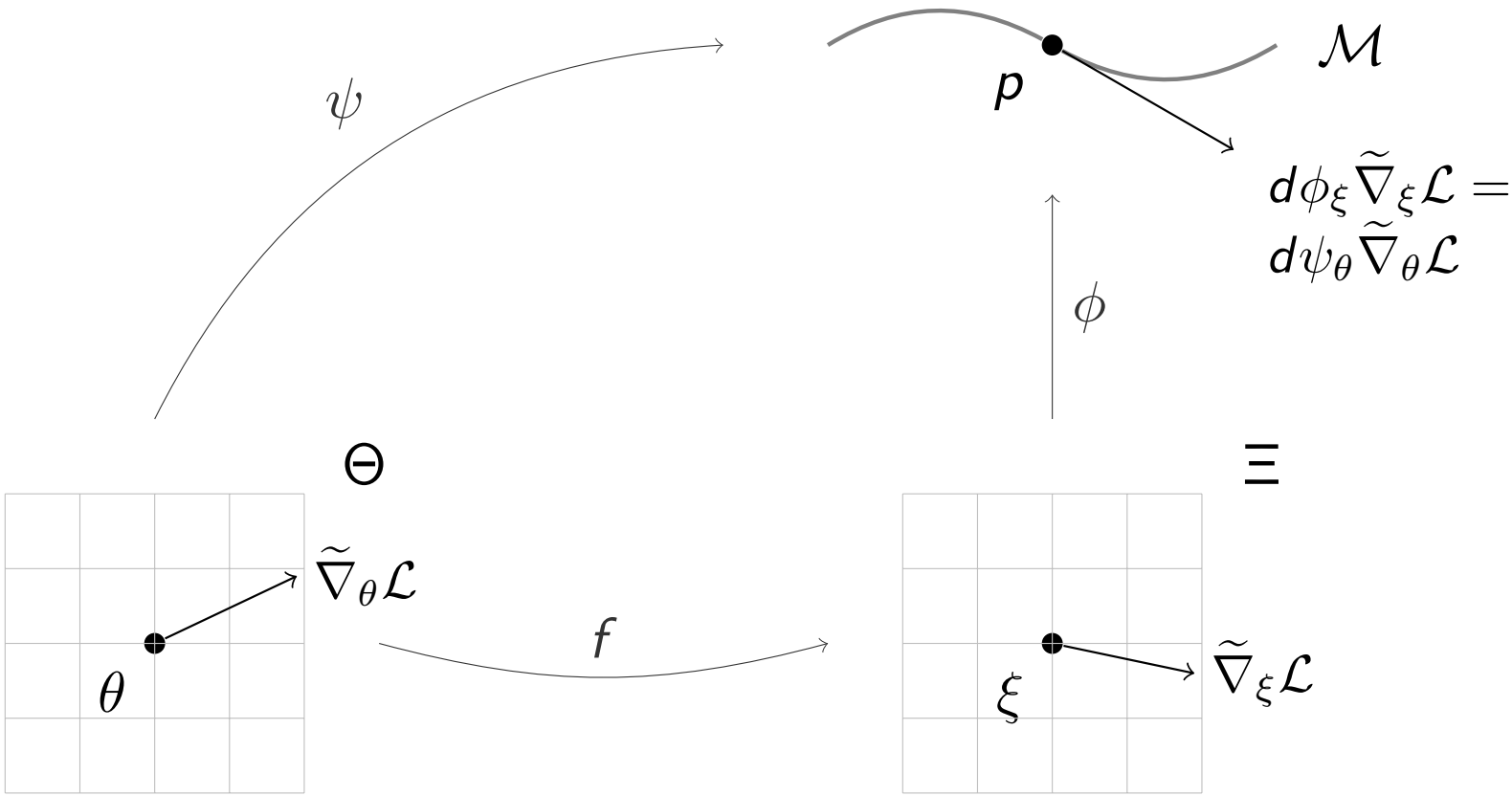
Definition (Reparametrisation invariance)

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 (11) holds.


$\mathcal{M}$
p
ψ
$d\phi_\xi \mathcal{O}(\phi, \xi) =$
$d\psi_\theta \mathcal{O}(\psi, \theta)$
ϕ
Θ
Ξ
$\mathcal{O}(\psi, \theta)$
θ
f
ξ
$\mathcal{O}(\phi, \xi)$

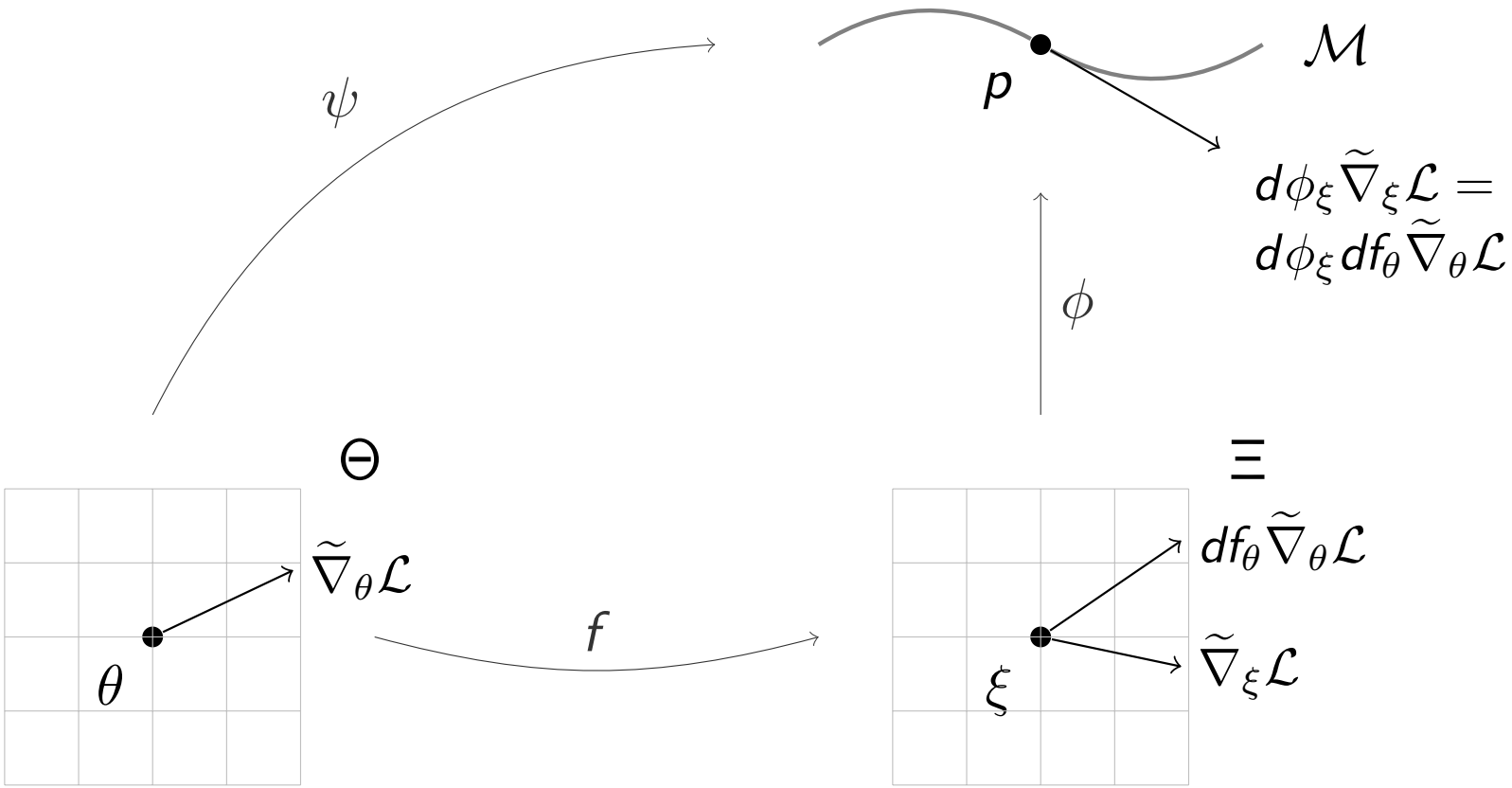
Theorem

The natural parameter gradient is reparametrisation invariant.



Parametrisation invariance on the parameter space

Parametrisation invariance can also be defined on the parameter space itself, see [4]. We argue that this is a less suitable definition since the difference between the vectors disappears when mapping to the model.



[4] James Martens. "New Insights and Perspectives on the Natural Gradient Method". In: *Journal of Machine Learning Research* 21 (2020), pp. 1–76

The trajectory of the natural gradient *method* is parametrisation dependent

The natural gradient method is performed in discrete time steps. This causes the gradient vector to be parallel to the gradient flow only at the start of each step. The exact trajectory of the procedure will therefore depend on the choice of parametrisation. This is however also a problem in non-overparametrised systems.

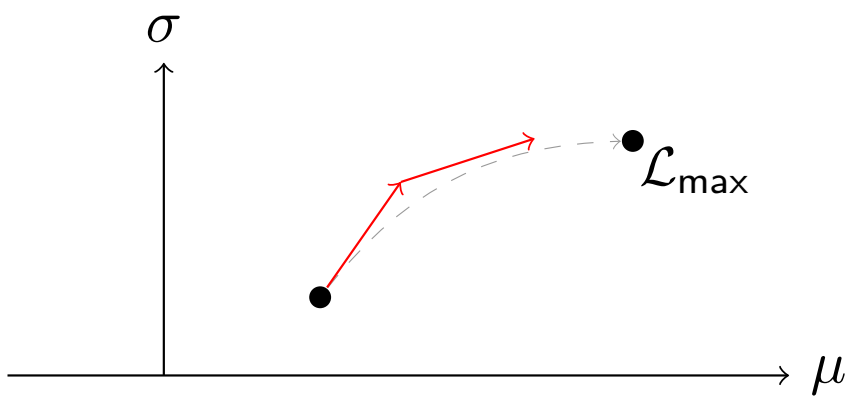


Figure: Natural gradient descent trajectory on a manifold of normal distributions

The End