

# A Geometric Modeling of Occam’s Razor in Deep Learning

Ke Sun 

CSIRO’s Data61, Australia  
Ke.Sun@data61.csiro.au, sunk@ieee.org

Frank Nielsen 

Sony Computer Science Laboratories Inc. (Sony CSL)  
Tokyo, Japan  
Frank.Nielsen@acm.org

Version: March 2025

## Abstract

Why do deep neural networks (DNNs) benefit from very high dimensional parameter spaces? Their huge parameter complexities *vs.* stunning performances in practice is all the more intriguing and not explainable using the standard theory of model selection for regular models. In this work, we propose a geometrically flavored information-theoretic approach to study this phenomenon. With the belief that simplicity is linked to better generalization, as grounded in the theory of minimum description length, the objective of our analysis is to examine and bound the complexity of DNNs. We introduce the locally varying dimensionality of the parameter space of neural network models by considering the number of significant dimensions of the Fisher information matrix, and model the parameter space as a manifold using the framework of singular semi-Riemannian geometry. We derive model complexity measures which yield short description lengths for deep neural network models based on their singularity analysis thus explaining the good performance of DNNs despite their large number of parameters.

## 1 Introduction

Deep neural networks (DNNs) are usually large models in terms of storage costs. In the classical model selection theory, such models are not favored as compared to simple models with the same training performance. For example, if one applies the Bayesian information criterion (BIC) [70] to DNN, a shallow neural network (NN) will be preferred over a deep NN due to the penalty term with respect to (w.r.t.) the complexity. A basic principle in science is the Occam<sup>1</sup>’s Razor, which favors simple models over complex ones that accomplish the same task. This raises the fundamental question of *how to measure the simplicity or the complexity of a model*.

Formally, the preference of simple models has been studied in the area of minimum description length (MDL) [25, 66, 67], also known in another thread of research as the minimum message length (MML) [77]. By the theory of MDL [25], statistical models that can most concisely

---

\*This work first appeared under the former title “Lightlike Neuromanifolds, Occam’s Razor and Deep Learning” in 2019.

<sup>1</sup>William of Ockham (ca. 1287 — ca. 1347), a monk (friar) and philosopher.

communicate the observed data are favored and expected to generalize better [9, 11, 24, 26, 83]. This is intuitive, as complex models often lead to overfit.

Consider a parametric family of distributions  $\mathcal{M} = \{p(\mathbf{x} \mid \boldsymbol{\theta})\}$  with  $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^D$ . The distributions are mutually absolutely continuous, which guarantees all densities to have the same support. Otherwise, many problems of non-regularity will arise as described by [28, 61]. The Fisher information matrix (FIM)  $\mathcal{I}(\boldsymbol{\theta})$  is a  $D \times D$  positive semi-definite (psd) matrix:  $\mathcal{I}(\boldsymbol{\theta}) \succeq 0$ . The model is called *regular* if it is (i) identifiable [13] with (ii) a non-degenerate and finite Fisher information matrix (i.e.,  $\mathcal{I}(\boldsymbol{\theta}) \succ 0$ ).

In a Bayesian setting, the description length of a set of  $N$  i.i.d. observations  $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N \subset \mathcal{X}$  w.r.t.  $\mathcal{M}$  can be defined as the number of *nats* with the coding scheme of a parametric model  $p(\mathbf{x} \mid \boldsymbol{\theta})$  and a prior  $p(\boldsymbol{\theta})$ . The code length of any  $\mathbf{x}_i$  is given by the cross entropy between the empirical distribution  $\delta_i(\mathbf{x}) = \delta(\mathbf{x} - \mathbf{x}_i)$ , where  $\delta(\cdot)$  denotes the Dirac’s delta function, and  $p(\mathbf{x}) = \int p(\mathbf{x} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}) d\boldsymbol{\theta}$ . Therefore, the description length of  $\mathbf{X}$  is

$$-\log p(\mathbf{X}) = \sum_{i=1}^N h^\times(\delta_i : p) = -\sum_{i=1}^N \log \int p(\mathbf{x}_i \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}, \quad (1)$$

where  $h^\times(p : q) := -\int p(\mathbf{x}) \log q(\mathbf{x}) d\mathbf{x}$  denotes the cross entropy between  $p(\mathbf{x})$  and  $q(\mathbf{x})$ , and  $\log$  denotes natural logarithm throughout the paper. The code length means the cumulative loss of the Bayesian mixture model  $p(\mathbf{x})$  w.r.t. the observations  $\mathbf{X}$ . Equation (1) corresponds to the Bayesian universal code. In MDL, the optimal code in terms of the minimax strategy [71] is given by the normalized maximum likelihood (NML) code. With a suitable choice of the prior, the Bayesian universal code and the NML code asymptotically coincide [25] with  $O(1)$  difference.

By using Jeffreys<sup>2</sup>, non-informative prior [3] as  $p(\boldsymbol{\theta})$ , the MDL in eq. (1) can be approximated (see [7, 66, 67]) as

$$\chi = \underbrace{-\log p(\mathbf{X} \mid \hat{\boldsymbol{\theta}})}_{\text{fitness}} + \underbrace{\frac{D}{2} \log \frac{N}{2\pi}}_{\text{penalize high dof}} + \overbrace{\log \left( \int \sqrt{|\mathcal{I}(\boldsymbol{\theta})|} d\boldsymbol{\theta} \right)}^{\text{geometric complexity}} \underbrace{\hspace{1cm}}_{\text{model capacity}}, \quad (2)$$

where  $\hat{\boldsymbol{\theta}} \in \Theta$  is the maximum likelihood estimation (MLE), or the projection [3] of  $\mathbf{X}$  onto the model,  $D = \dim(\Theta)$  is the model size,  $N$  is the number of observations, and  $|\cdot|$  denotes the matrix determinant. In this paper, the symbols  $\chi$  and  $\mathcal{O}$  and the term “razor” all refer to the same concept, that is the description length of the data  $\mathbf{X}$  by the model  $\mathcal{M}$ . The smaller those quantities, the better.

The first term in eq. (2) is the fitness of the model to the observed data. The second and the third terms measures the *geometric complexity* [50] and make  $\chi$  favor simple models. The second  $O(\log N)$  term only depends on the number of parameters  $D$  and the number of observations  $N$ . It penalizes large models with a high degree of freedom (dof). The third  $O(1)$  term is independent to the observed data and measures the model capacity, or the total “number” of distinguishable distributions [50] in the model.

Unfortunately, this razor  $\chi$  in eq. (2) does not fit straightforwardly into DNNs, which are high-dimensional *singular* models. The FIM  $\mathcal{I}(\boldsymbol{\theta})$  are large singular matrices (not full rank) and the last term may be difficult to evaluate. Based on the second term on the right-hand-side (RHS), a DNN can have very high complexity and therefore is less favored against a shallow

<sup>2</sup>Sir Harold Jeffreys (1891–1989), a British statistician.

network. This contradicts the good generalization of DNNs as compared to shallow NNs. These issues call for a new analysis of the MDL in the DNN setting.

Towards this direction, we made the following contributions in this paper:

- New concepts and methodologies from singular semi-Riemannian geometry [41] to analyze the space of neural networks;
- A definition of the local dimensionality in this space, that is the amount of non-singularity, with bounding analysis;
- A connection between  $f$ -mean and DNN model complexity with related bounds;
- A new MDL formulation, which explains how the singularity contribute to the “negative complexity” of DNNs: That is, the model turns simpler as the number of parameters grows.

The rest of this paper is organized as follows. Section 2 reviews singularities in information geometry. In the setting of a DNN, section 3 introduces its singular parameter manifold. Section 4 bounds the number of singular dimensions of the parameter manifold of the DNN. Sections 5 to 8 derive our MDL criterion based on two different priors, and discusses how model complexity is affected by the singular geometry. We discuss related work in section 9 and conclude in section 10. Proofs and related derivations of our main results are provided in the appendix.

## 2 Lightlike Statistical Manifold

In this paper, bold capital letters like  $\mathbf{A}$  denote matrices, bold small letters like  $\mathbf{a}$  denote vectors, normal capital/small letters like  $A/a$  and Greek letters like  $\alpha$  denote scalars, and calligraphy letters like  $\mathcal{M}$  denote manifolds (with exceptions). We use  $\mathbf{X}$ ,  $\mathbf{Y}$  and  $\mathbf{Z}$  to denote a collection of  $N$  random observations and use  $\mathbf{x}$ ,  $\mathbf{y}$  and  $\mathbf{z}$  to denote one single observation.

The term “statistical manifold” refers to  $\mathcal{M} = \{p(\mathbf{x}|\boldsymbol{\theta})\}$ , where each point of  $\mathcal{M}$  corresponds to a probability distribution  $p(\mathbf{x}|\boldsymbol{\theta})$ <sup>3</sup>. The discipline of information geometry [3] studies such a space in the Riemannian and more generally differential geometry framework. Hotelling [31] and independently Rao [63, 64] proposed to endow a parametric space of statistical models with the Fisher information matrix as a Riemannian metric:

$$\mathcal{I}(\boldsymbol{\theta}) := \mathbb{E}_p \left( \frac{\partial \log p(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log p(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \right), \quad (3)$$

where  $\mathbb{E}_p$  denotes the expectation w.r.t.  $p(\mathbf{x}|\boldsymbol{\theta})$ . The corresponding infinitesimal squared length element  $ds^2 = \text{tr}(\mathcal{I}(\boldsymbol{\theta})d\boldsymbol{\theta}d\boldsymbol{\theta}^\top) = \langle d\boldsymbol{\theta}, d\boldsymbol{\theta} \rangle_{\mathcal{I}(\boldsymbol{\theta})} = d\boldsymbol{\theta}^\top \mathcal{I}(\boldsymbol{\theta}) d\boldsymbol{\theta}$ , where  $\text{tr}(\cdot)$  means the matrix trace<sup>4</sup>, is independent of the underlying parameterization of the population space.

Amari further developed this approach by revealing the dualistic structure of statistical manifolds which extends the Riemannian framework [3, 54]. The MDL criterion arising from the geometry of Bayesian inference with Jeffreys’ prior for regular models is detailed in [7]. In information geometry, the regular assumption is (1) an open connected parameter space in some Euclidean space; and (2) the FIM exists and is non-singular. However, in general, the FIM is only positive semi-definite and thus for non-regular models like neuromanifolds [3] or Gaussian mixture models [78], the manifold is not Riemannian but *singular semi-Riemannian* [17, 41]. In the machine learning community, singularities have often been dealt with as *a minor issue*:

<sup>3</sup>To be more precise, a statistical manifold [42] is a structure  $(\nabla, g, C)$  on a smooth manifold  $\mathcal{M}$ , where  $g$  is a metric tensor,  $\nabla$  a torsion-free affine connection, and  $C$  is a symmetric covariant tensor of order 3.

<sup>4</sup>Using the cyclic property of the matrix trace, we have  $ds^2 = \text{tr}(\mathcal{I}(\boldsymbol{\theta})d\boldsymbol{\theta}d\boldsymbol{\theta}^\top) = d\boldsymbol{\theta}^\top \mathcal{I}(\boldsymbol{\theta}) d\boldsymbol{\theta}$ .

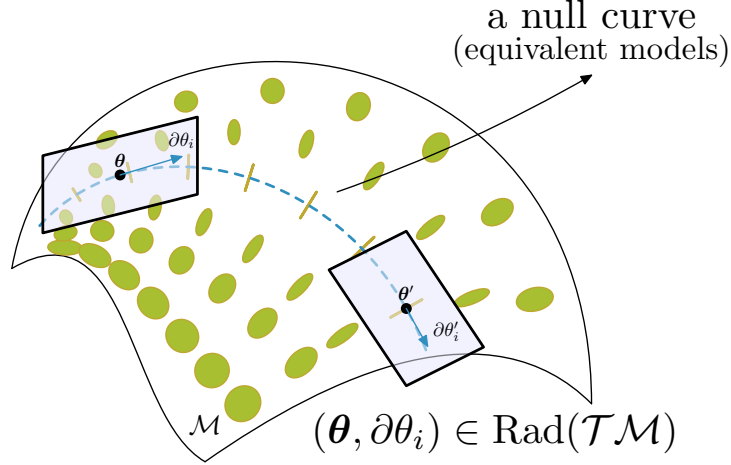


Figure 1: A toy lightlike manifold  $\mathcal{M}$  with a null curve. The ellipses are Tissot’s indicatrices, showing how circles of infinitesimal radius are distorted by the lightlike geometry on  $\mathcal{M}$ . On the null curve, the FIM is degenerate so that  $\langle \partial\theta_i, \partial\theta_i \rangle_{\mathcal{I}} = 0$ . Therefore the local dynamic  $\partial\theta_i$  (tangent vector of the null curve) has zero length, meaning that it does not change the model. The radical distribution  $\text{Rad}(\mathcal{TM})$  is formed by the null curve and its tangent vectors. In the context of DNNs, such dynamics refer to small changes of neural networks weights/biases that do not alter the global model.

For example, the natural gradient has been generalized based on the Moore-Penrose inverse of  $\mathcal{I}(\theta)$  [75] to avoid potential non-invertible FIMs. Watanabe [78] addressed the fact that most usual learning machines are singular in his singular learning theory which relies on algebraic geometry. Nakajima and Ohmoto [52] discussed dually flat structures for singular models.

Recently, preliminary efforts [6, 33] tackle singularity at the core, mostly from a mathematical standpoint. For example, Jain et al. [33] studied the Ricci curvature tensor of such manifolds. These mathematical notions are used in the community of differential geometry or general relativity but have not yet been ported to the machine learning community.

Following these efforts, we first introduce informally some basic concepts from a machine learning perspective to define the differential geometry of non-regular statistical manifolds. The *tangent space*  $\mathcal{T}_{\theta}(\mathcal{M})$  is a  $D$ -dimensional ( $D = \dim(\mathcal{M})$ ) real vector space, that is the local linear approximation of the manifold  $\mathcal{M}$  at the point  $\theta \in \mathcal{M}$ , equipped with the inner product induced by  $\mathcal{I}(\theta)$ . The *tangent bundle*  $\mathcal{TM} := \{(\theta, v), \theta \in \mathcal{M}, v \in \mathcal{T}_{\theta}\}$  is the  $2D$ -dimensional manifold obtained by combining all tangent spaces for all  $\theta \in \mathcal{M}$ . A *vector field* is a smooth mapping from  $\mathcal{M}$  to  $\mathcal{TM}$  such that each point  $\theta \in \mathcal{M}$  is attached a tangent vector originating from itself. Vector fields are cross-sections of the tangent bundle. In a local coordinate chart  $\theta$ , the vector fields along the frame are denoted as  $\partial\theta_i$ . A *distribution* (not to be confused with probability distributions which are points on  $\mathcal{M}$ ) means a vector subspace of the tangent bundle spanned by several independent vector fields, such that each point  $\theta \in \mathcal{M}$  is associated with a subspace of  $\mathcal{T}_{\theta}(\mathcal{M})$  and those subspaces vary smoothly with  $\theta$ . Its dimensionality is defined by the dimensionality of the subspace, *i.e.*, the number of vector fields that span the distribution.

In a *lightlike* manifold [17, 41]  $\mathcal{M}$ ,  $\mathcal{I}(\theta)$  can be degenerate. The tangent space  $\mathcal{T}_{\theta}(\mathcal{M})$  is a vector space with a kernel subspace, *i.e.*, a nullspace. A null vector field is formed by null

vectors, whose lengths measured according to the Fisher metric tensor are all zero. The *radical*<sup>5</sup> distribution  $\text{Rad}(\mathcal{TM})$  is the distribution spanned by the null vector fields. Locally at  $\theta \in \mathcal{M}$ , the tangent vectors in  $\mathcal{T}_\theta(\mathcal{M})$  which span the kernel of  $\mathcal{I}(\theta)$  are denoted as  $\text{Rad}_\theta(\mathcal{TM})$ . In a local coordinate chart,  $\text{Rad}(\mathcal{TM})$  is well defined if these  $\text{Rad}_\theta(\mathcal{TM})$  form a valid distribution. We write  $\mathcal{TM} = \text{Rad}(\mathcal{TM}) \oplus \mathcal{S}(\mathcal{TM})$ , where ‘ $\oplus$ ’ is the direct sum, and the *screen distribution*  $\mathcal{S}(\mathcal{TM})$  is complementary to the radical distribution  $\text{Rad}(\mathcal{TM})$  and has a non-degenerate induced metric. See fig. 1 for an illustration of the concept of radical distribution.

We can find a local coordinate frame (a frame is an ordered basis)  $(\theta_1, \dots, \theta_d, \theta_{d+1}, \dots, \theta_D)$ , where the first  $d$  dimensions  $\theta^s = (\theta_1, \dots, \theta_d)$  correspond to the screen distribution, and the remaining  $\bar{d} := D - d$  dimensions  $\theta^r = (\theta_{d+1}, \dots, \theta_D)$  correspond to the radical distribution. The local inner product  $\langle \cdot, \cdot \rangle_{\mathcal{I}}$  satisfies

$$\begin{aligned} \langle \partial\theta_i, \partial\theta_j \rangle_{\mathcal{I}} &= \delta_{ij}, \quad (\forall 1 \leq i, j \leq d) \\ \langle \partial\theta_i, \partial\theta_k \rangle_{\mathcal{I}} &= 0, \quad (\forall d+1 \leq i \leq D, 1 \leq k \leq D) \end{aligned}$$

where  $\delta_{ij} = 1$  if and only if (iff)  $i = j$  and  $\delta_{ij} = 0$ , otherwise. Unfortunately, this frame is not unique [16]. We will abuse  $\mathcal{I}$  to denote both the FIM of  $\theta$  and the FIM of  $\theta^s$ . One has to remember that  $\mathcal{I}(\theta) \succeq 0$ , while  $\mathcal{I}(\theta^s) \succ 0$  is a proper Riemannian metric. Hence, both  $\mathcal{I}^{-1}(\theta^s)$  and  $\log|\mathcal{I}(\theta^s)|$  are well-defined.

**Remark 1.** Notice that the Fisher information matrix is covariant under reparameterization. That is, let  $\theta(\lambda)$  be an invertible smooth reparameterization of  $\lambda$ . Then the FIM rewrites in the  $\theta$ -parameterization as:

$$\mathcal{I}(\theta) = \mathbf{J}_{\theta \rightarrow \lambda}^T \mathcal{I}(\lambda(\theta)) \mathbf{J}_{\theta \rightarrow \lambda}, \quad (4)$$

where  $\mathbf{J}_{\theta \rightarrow \lambda}$  is the full rank Jacobian matrix.

The natural gradient flows (vector fields on  $\mathcal{M}$ ) with respect to  $\lambda$  and  $\theta$  coincide but not the natural gradient descent methods (learning paths that consist of sequences of points on  $\mathcal{M}$ ) because of the non-zero learning step sizes.

Furthermore, the ranks of  $\mathcal{I}(\theta)$  and  $\mathcal{I}(\lambda)$  as well as the dimensions of the screen and radical distributions coincide. Hence, the notion of singularities is intrinsic and independent of the smooth reparameterization.

### 3 Lightlike Neuromanifold

This section instantiates the concepts in the previous section 2 in terms of a simple DNN predictive model. The random variable  $\mathbf{x} = (\mathbf{z}, y)$  of interest consists of two components:  $\mathbf{z}$ , referred to as the “input”, and  $y$ , referred to as the “target”. By assumption, their joint probability distribution is specified by

$$\log p(\mathbf{x} \mid \psi, \theta) = \log p(\mathbf{z} \mid \psi) + \log p(y \mid \mathbf{z}, \theta),$$

where  $p(\mathbf{z} \mid \psi)$  is a generative model of  $\mathbf{z}$  which is parameterized by  $\psi$ ,  $p(y \mid \mathbf{z}, \theta)$  is a predictive DNN, and  $\theta$  consists of all neural network parameters.

Our main subject is the latter predictive model  $p(y \mid \mathbf{z}, \theta)$  and its parameter manifold  $\mathcal{M}_\theta$ . Here, we need the generative model  $p(\mathbf{z} \mid \psi)$  for the purpose of discussing how the geometry of  $\mathcal{M}_\theta$  is affected by the choice of  $p(\mathbf{z} \mid \psi)$  and can be studied independent of the parameter space of  $p(\mathbf{z} \mid \psi)$ , which we denote as  $\mathcal{M}_\psi$ . In the end, our results do not depend on the specific form of  $p(\mathbf{z})$  or whether it is parametric.

---

<sup>5</sup>Radical stems from Latin and means root.

For  $p(y | \mathbf{z}, \boldsymbol{\theta})$ , we consider a deep feed-forward network with  $L$  layers, uniform width  $M$  except the last layer which has  $m$  output units ( $m < M$ ), input  $\mathbf{z} \in \mathcal{Z}$  with  $\dim(\mathcal{Z}) = M$ , pre-activations  $\mathbf{h}^l$  of size  $M$  (except that in the last layer,  $\mathbf{h}^L$  has  $m$  elements), post-activations  $\mathbf{z}^l$  of size  $M$ , weight matrices  $\mathbf{W}^l$  and bias vectors  $\mathbf{b}^l$  ( $1 \leq l \leq L$ ). The layers are given by

$$\begin{aligned} \mathbf{z}^l &= \phi(\mathbf{h}^l), \\ \mathbf{h}^l &= \mathbf{W}^l \mathbf{z}^{l-1} + \mathbf{b}^l, \\ \mathbf{z}^0 &= \mathbf{z}, \end{aligned} \tag{5}$$

where  $\phi$  is an element-wise nonlinear activation function such as ReLU [22].

Without loss of generality, we assume multinomial<sup>6</sup> output units and the DNN output [23]

$$y \sim \text{Multinomial}(\text{SoftMax}(\mathbf{h}^L))$$

is a random label in the set  $\{1, \dots, m\}$ , where

$$\text{SoftMax}(\mathbf{t}) := \frac{1}{\sum_{i=1}^m \exp(t_i)} (\exp(t_1), \exp(t_2), \dots, \exp(t_m))$$

denotes the softmax function.  $\text{SoftMax}(\mathbf{h}^L)$  is a random point in  $\Delta^m$ , the  $(m-1)$  dimensional statistical simplex. Therefore,  $p(y = k) = \exp(h_k^L) / \sum_{j=1}^m \exp(h_j^L)$ ,  $k = 1, \dots, m$ . The neural network parameters  $\boldsymbol{\theta}$  consists of  $\mathbf{W}^l$  and  $\mathbf{b}^l$ ,  $l = 1, \dots, L$ . In this supervised setting, the code length in eq. (1) means the predictive loss of the Bayesian mixture model  $p(\mathbf{x}) = p(\mathbf{z}) \int p(y | \mathbf{z}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$  w.r.t. to the observed pairs  $(\mathbf{z}_i, y_i)$ . The smaller the code length, the more accurate the prediction.

All such neural networks  $\text{NN}_{\boldsymbol{\theta}}$  when  $\boldsymbol{\theta}$  varies in a parameter space are referred to as the *neuromanifold*:  $\mathcal{M}_{\boldsymbol{\theta}} = \{\text{NN}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ . Similarly, the parameter space of the distribution family  $p(\mathbf{z} | \boldsymbol{\psi})$  is denoted as  $\mathcal{M}_{\boldsymbol{\psi}}$ . In machine learning, we are often interested in the FIM w.r.t.  $\boldsymbol{\theta}$  as it reveals the geometry of the parameter space. However, the FIM can also be computed relatively w.r.t. a subset of  $\boldsymbol{\theta}$  in a sub-system [73].

By the definition in eq. (3), the FIM on the product manifold  $\mathcal{M}_{\boldsymbol{\psi}} \times \mathcal{M}_{\boldsymbol{\theta}}$  is in a block-diagonal form

$$\mathcal{I}(\boldsymbol{\psi}, \boldsymbol{\theta}) = \begin{bmatrix} \mathcal{I}(\boldsymbol{\psi}) & \mathbf{0} \\ \mathbf{0} & \mathcal{I}(\boldsymbol{\theta}) \end{bmatrix}. \tag{6}$$

The off-diagonal blocks are zero-matrices (denoted as  $\mathbf{0}$ ) because the generative and predictive models do not share parameters. Indeed, we have

$$\begin{aligned} \mathbb{E}_p \left( \frac{\partial \log p(\mathbf{x} | \boldsymbol{\psi}, \boldsymbol{\theta})}{\partial \boldsymbol{\psi}} \frac{\partial \log p(\mathbf{x} | \boldsymbol{\psi}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \right) &= \mathbb{E}_p \left( \frac{\partial \log p(\mathbf{z} | \boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \frac{\partial \log p(y | \mathbf{z}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \right) \\ &= \mathbb{E}_{p(\mathbf{z} | \boldsymbol{\psi})} \left( \frac{\partial \log p(\mathbf{z} | \boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \mathbb{E}_{p(y | \mathbf{z}, \boldsymbol{\theta})} \left( \frac{\partial \log p(y | \mathbf{z}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^\top \right) = \mathbf{0}, \end{aligned}$$

where  $\mathbb{E}_{p(y | \mathbf{z}, \boldsymbol{\theta})} \left( \frac{\partial \log p(y | \mathbf{z}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)$  is the expectation of the score function and is always zero. The metric  $\mathcal{I}(\boldsymbol{\psi}, \boldsymbol{\theta})$  is a *product metric*, meaning that the geometry of  $\mathcal{M}_{\boldsymbol{\theta}}$  defined by  $\mathcal{I}(\boldsymbol{\theta})$  can be studied separately to the geometry of  $\mathcal{M}_{\boldsymbol{\psi}}$ .

<sup>6</sup>In fact, a generalization of the Bernoulli distribution with integer  $k \geq 2$  mutually exclusive events, called informally a multinoulli distribution since it is a multinomial distribution with a single trial.

As we are interested in the predictive model corresponding to the diagonal block  $\mathcal{I}(\boldsymbol{\theta})$ , we further have (see *e.g.* [57][72] for derivations)

$$\mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{z})} \left[ \left( \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}} \right)^\top \mathbf{C}(\mathbf{z}) \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}} \right], \quad (7)$$

where the expectation is taken w.r.t.  $p(\mathbf{z}) := p(\mathbf{z} | \boldsymbol{\psi})$ , an underlying true distribution in the input space depending on the parameter  $\boldsymbol{\psi}$ .  $\frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}}$  is the  $m \times D$  parameter-output Jacobian matrix, based on a given input  $\mathbf{z}$ ,  $\mathbf{C}(\mathbf{z}) := \text{diag}(\mathbf{o}(\mathbf{z})) - \mathbf{o}(\mathbf{z})\mathbf{o}(\mathbf{z})^\top \succeq 0$ ,  $\text{diag}(\cdot)$  means the diagonal matrix with the given diagonal entries, and  $\mathbf{o}(\mathbf{z}) := \text{SoftMax}(\mathbf{h}^L(\mathbf{z}))$  is the predicted class probabilities of  $\mathbf{z}$ . By the definition of SoftMax, each dimension of  $\mathbf{o}(\mathbf{z})$  represents a positive probability, although  $\mathbf{o}(\mathbf{z})$  can be arbitrarily close to a one-hot vector. As a result, the kernel of the psd matrix  $\mathbf{C}(\mathbf{z})$  is given by  $\{\lambda \mathbf{1} : \lambda \in \mathbb{R}\}$ , where  $\mathbf{1}$  is the vector of all 1's.

In eq. (7),  $\mathcal{I}(\boldsymbol{\theta})$  is the single-observation FIM. It is obvious that the FIM w.r.t. the joint distribution  $p(\mathbf{X} | \boldsymbol{\theta})$  of multiple observations is  $N\mathcal{I}(\boldsymbol{\theta})$  (Fisher information is additive), so that  $\mathcal{I}(\boldsymbol{\theta})$  does not scale with  $N$ . In theory, computing  $\mathcal{I}(\boldsymbol{\theta})$  requires assuming  $p(\mathbf{z})$ , which depends on the parameter  $\boldsymbol{\psi}$ . This makes sense as  $(\boldsymbol{\psi}_1, \boldsymbol{\theta})$  and  $(\boldsymbol{\psi}_2, \boldsymbol{\theta})$  with  $\boldsymbol{\psi}_1 \neq \boldsymbol{\psi}_2$  are different points on the product manifold  $\mathcal{M}_{\boldsymbol{\psi}} \times \mathcal{M}_{\boldsymbol{\theta}}$  and thus their  $\mathcal{I}(\boldsymbol{\theta})$  should be different. In practice, one only gets access to a set of  $N$  i.i.d. samples drawn from an unknown  $p(\mathbf{z} | \boldsymbol{\psi})$ . In this case, it is reasonable to take  $p(\mathbf{z})$  in eq. (7) to be the empirical distribution  $\hat{p}(\mathbf{z})$  so that  $p(\mathbf{z}) = \hat{p}(\mathbf{z}) := \frac{1}{N} \sum_{i=1}^N \delta(\mathbf{z} - \mathbf{z}_i)$ , then

$$\mathcal{I}(\boldsymbol{\theta}) = \hat{\mathcal{I}}(\boldsymbol{\theta}) := \frac{1}{N} \sum_{i=1}^N \left[ \left( \frac{\partial \mathbf{h}^L(\mathbf{z}_i)}{\partial \boldsymbol{\theta}} \right)^\top \mathbf{C}(\mathbf{z}_i) \frac{\partial \mathbf{h}^L(\mathbf{z}_i)}{\partial \boldsymbol{\theta}} \right]. \quad (8)$$

The FIM computed in this way does not rely on the assumption of a parametric generative model  $p(\mathbf{z} | \boldsymbol{\psi})$  and the choice of a  $\boldsymbol{\psi}$ .  $\hat{\mathcal{I}}(\boldsymbol{\theta})$  can be directly computed from the observed  $\mathbf{z}_i$ 's and does not depend on the observed  $y_i$ 's. Although denoted differently than  $\mathcal{I}(\boldsymbol{\theta})$  in the current paper, this  $\hat{\mathcal{I}}(\boldsymbol{\theta})$  is a standard version of the definition of the FIM for neural networks [40, 47, 72, 73].

By considering the neural network weights and biases as random variables satisfying a prescribed prior distribution [35, 60], this  $\mathcal{I}(\boldsymbol{\theta})$  can be regarded as a random matrix [48] depending on the structure of the DNN and the prior. The empirical density of  $\mathcal{I}(\boldsymbol{\theta})$  is the empirical distribution of its eigenvalues  $\{\lambda_i\}_{i=1}^D$ , that is,  $\rho_D(\lambda) = \frac{1}{D} \sum_{i=1}^D \delta(\lambda_i)$ . If at the limit  $D \rightarrow \infty$ , the empirical density converges to a probability density function (pdf), then

$$\rho_{\mathcal{I}}(\lambda) := \lim_{D \rightarrow \infty} \rho_D(\lambda) \quad (9)$$

is called the *spectral density* of the Fisher information matrix.

For DNN, we assume that

- (A1) At the MLE  $\hat{\boldsymbol{\theta}}$ , the prediction  $\text{SoftMax}(\mathbf{h}^L(\mathbf{z}_i))$  perfectly recovers (tending to be one-hot vectors) the training target  $y_i$ , for all the training samples  $(\mathbf{z}_i, y_i)$ .

In this case, the negative Hessian of the average log-likelihood

$$\mathfrak{J}(\boldsymbol{\theta}) := -\frac{1}{N} \frac{\partial^2 \log p(\mathbf{X} | \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = -\frac{1}{N} \sum_{i=1}^N \frac{\partial^2 \log p(y_i | \mathbf{z}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}$$

is called the observed FIM (sample-based FIM), which is also known as the ‘‘empirical Fisher’’ in machine learning literature [40, 47]. In our notations explained in table 1, the FIM  $\mathcal{I}$  depends



on the true distribution  $p(\mathbf{z})$  and does not depend on the observed samples. In the expression of the FIM in eq. (7), if  $p(\mathbf{z}) = \hat{p}(\mathbf{z})$ , then  $\mathcal{I}$  become  $\hat{\mathcal{I}}$ , which depends on the observed input  $\mathbf{z}_i$ 's. The observed FIM  $\mathfrak{J}$  depends on both the observed input  $\mathbf{z}_i$ 's and the observed target  $y_i$ 's. If  $p(\mathbf{z}) = \hat{p}(\mathbf{z})$ , the observed FIM coincides with the FIM at the MLE  $\hat{\boldsymbol{\theta}}$  and  $\mathfrak{J}(\hat{\boldsymbol{\theta}}) = \hat{\mathcal{I}}(\hat{\boldsymbol{\theta}})$ . For general statistical models, there is a residual term in between these two matrices which scales with the training error (see *e.g.* Eq. 6.19 in section 6 of [4], or eq. (24) in the appendix). How these different metric tensors are called is just a matter of terminology. One should distinguish them by examining whether/how they depend (partially) on the observed information.

Table 1: The FIM and the observed FIM. The last three columns explains whether the tensor depends on the observed  $\mathbf{z}_i$ 's, whether it depends on the observed  $y_i$ 's, and whether they can be computed in practice based on empirical observations.

Notation	Name	Depend on $\mathbf{z}_i$	Depend on $y_i$	Computable
$\mathcal{I}(\boldsymbol{\theta})$	FIM (w.r.t. true $p(\mathbf{z})$ )	No	No	No
$\hat{\mathcal{I}}(\boldsymbol{\theta})$	FIM (w.r.t. empirical $\hat{p}(\mathbf{z})$ )	Yes	No	Yes
$\mathfrak{J}(\boldsymbol{\theta})$	observed FIM	Yes	Yes	Yes

## 4 Local Dimensionality

This section quantitatively measures the singularity of the neuromanifold. Our main definitions and results do *not* depend on the settings introduced in the previous section and can be generalized to similar models including stochastic neural networks [12]. For example, if the output units or the network structure is changed, the expression of the FIM and related results can be adapted straightforwardly. Our derivations depend on that (1) DNNs have a large amount of singularity corresponding zero eigenvalues of the FIM; and (2) the spectrum of the (observed) FIM has many eigenvalues close to zero [36]. That being said, our results also apply to singular models [78] with similar properties.

**Definition 1** (Local dimensionality). *The local dimensionality  $d(\boldsymbol{\theta}) := \text{rank}(\mathcal{I}(\boldsymbol{\theta}))$  of the neuromanifold  $\mathcal{M}$  at  $\boldsymbol{\theta} \in \mathcal{M}$  refers to the rank of the FIM  $\mathcal{I}(\boldsymbol{\theta})$ . If  $p(\mathbf{z}) = \hat{p}(\mathbf{z})$ , then  $d(\boldsymbol{\theta}) = \hat{d}(\boldsymbol{\theta}) := \text{rank}(\hat{\mathcal{I}}(\boldsymbol{\theta}))$ .*

The local dimensionality  $d(\boldsymbol{\theta})$  is the number of degrees of freedom at  $\boldsymbol{\theta} \in \mathcal{M}$  which can change the probabilistic model  $p(y|\mathbf{z}, \boldsymbol{\theta})$  in terms of information theory. One can find a reparameterized DNN with  $d(\boldsymbol{\theta})$  parameters, which is locally equivalent to the original DNN with  $D$  parameters. Recall the dimensionality of the tangent bundle is two times the dimensionality of the manifold.

**Remark 2.** *The dimensionality of the screen distribution  $\mathcal{S}(\mathcal{T}\mathcal{M})$  at  $\boldsymbol{\theta}$  is  $2d(\boldsymbol{\theta})$ .*

By definition, the FIM as the singular semi-Riemannian metric of  $\mathcal{M}$  must be psd. Therefore it only has positive and zero eigenvalues, and the number of positive eigenvalues  $d(\boldsymbol{\theta})$  is not constant as  $\boldsymbol{\theta}$  varies in general.

**Remark 3.** *The local metric signature (number of positive, negative, zero eigenvalues of the FIM) of the neuromanifold  $\mathcal{M}$  is  $(d(\boldsymbol{\theta}), 0, D - d(\boldsymbol{\theta}))$ , where  $d(\boldsymbol{\theta})$  is the local dimensionality.*

The local dimensionality  $d(\boldsymbol{\theta})$  depends on the specific choice of  $p(\mathbf{z})$ . If  $p(\mathbf{z}) = \hat{p}(\mathbf{z})$ , then  $d(\boldsymbol{\theta}) = \hat{d}(\boldsymbol{\theta}) = \text{rank}(\hat{\mathcal{I}}(\boldsymbol{\theta}))$ . On the other hand, one can use the rank of the negative Hessian



$\mathfrak{J}(\boldsymbol{\theta})$  (i.e., observed rank) to get an approximation of the local dimensionality  $d(\boldsymbol{\theta}) \approx \text{rank}(\mathfrak{J}(\boldsymbol{\theta}))$ . In the MLE  $\hat{\boldsymbol{\theta}}$ , this approximation becomes accurate. We simply denote  $d$  and  $\hat{d}$ , instead of  $d(\boldsymbol{\theta})$  and  $\hat{d}(\boldsymbol{\theta})$ , if  $\boldsymbol{\theta}$  is clear from the context.

We first show that the lightlike dimensions of  $\mathcal{M}$  do not affect the neural network model in eq. (5).

**Lemma 1.** *If  $(\boldsymbol{\theta}, \sum_j \alpha_j \partial \theta_j) \in \text{Rad}(\mathcal{T}\mathcal{M})$ , i.e.  $\langle \sum_j \alpha_j \partial \theta_j, \sum_j \alpha_j \partial \theta_j \rangle_{\mathcal{I}(\boldsymbol{\theta})} = 0$ , then almost surely we have  $\frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}} \boldsymbol{\alpha} = \lambda(\mathbf{z}) \mathbf{1}$ , where  $\lambda(\mathbf{z}) \in \mathbb{R}$ .*

By lemma 1, the Jacobian  $\frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}}$  is the local linear approximation of the map  $\boldsymbol{\theta} \rightarrow \mathbf{h}^L$ . The dynamic  $\boldsymbol{\alpha}$  (coordinates of a tangent vector) on  $\mathcal{M}$  causes a uniform increment on the output  $\mathbf{h}^L$ , which, after the SoftMax function, does not change the neural network map  $\mathbf{z} \rightarrow y$ .

Then, we can upper-bound the local dimensionality using the rank of the parameter-output Jacobian  $\partial \mathbf{h}^L(\mathbf{z}) / \partial \boldsymbol{\theta}$ .

**Proposition 2.**  $\forall \boldsymbol{\theta} \in \mathcal{M}, \hat{d}(\boldsymbol{\theta}) \leq \sum_{i=1}^N \min \left\{ \text{rank} \left( \frac{\partial \mathbf{h}^L(\mathbf{z}_i)}{\partial \boldsymbol{\theta}} \right), m-1 \right\}.$

**Remark 4.** *While the total number  $D$  of free parameters is unbounded in DNNs, the local dimensionality estimated by  $\hat{d}(\boldsymbol{\theta})$  grows at most linearly w.r.t. the sample size  $N$ , given fixed  $m$  (size of the last layer). If both  $N$  and  $m$  are fixed, then  $\hat{d}(\boldsymbol{\theta})$  is bounded even when the network width  $M \rightarrow \infty$  and/or depth  $L \rightarrow \infty$ .*

The above bound is based on the inequality  $\text{rank}(\sum_i \mathbf{A}_i) \leq \sum_i \text{rank}(\mathbf{A}_i)$  for any matrices  $\mathbf{A}_i$ , which could lead to loose bounds. Alternatively, an upper bound can be established directly based on the definition of the matrix rank.

**Proposition 3.** *For all  $\boldsymbol{\theta} \in \mathcal{M}$ , we have  $d(\boldsymbol{\theta}) \leq \dim \text{span} \left( \bigcup_{\mathbf{z} \in \text{supp}(p)} \text{Row} \left( \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}} \right) \right)$ , where  $\text{supp}(p)$  is the support of  $p(\mathbf{z})$ , and  $\text{Row}(\mathbf{A})$  denotes the row vectors of the matrix  $\mathbf{A}$ . Similarly,  $\hat{d}(\boldsymbol{\theta})$  has an upper bound obtained by replacing  $\text{supp}(p)$  with  $\text{supp}(\hat{p})$  on the RHS, i.e. the union is over the observed  $\mathbf{z}_i$ 's.*

In summary, the less the rank of  $\frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}}$ , the more potential singularities in the neuromanifold. Note the Jacobian  $\frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}}$  can be further written as

$$\frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}} = \left( \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \mathbf{w}^1}, \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \mathbf{w}^2}, \dots, \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \mathbf{w}^L} \right), \quad (10)$$

where  $\mathbf{w}^l$  contains all parameters in the  $l$ 'th layer ( $l = 1, \dots, L$ ) and is obtained by stacking the columns of  $\mathbf{W}^l$  and  $\mathbf{b}^l$  into a long vector. We can bound  $\frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \mathbf{w}^l}$  individually as below.

**Proposition 4.** *It holds that*

$$\text{rank} \left( \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \mathbf{w}^l} \right) \leq \text{rank} (\mathbf{W}^L \boldsymbol{\Phi}^{L-1} \mathbf{W}^{L-1} \dots \mathbf{W}^{l+1} \boldsymbol{\Phi}^l) \leq \min_{s=l}^{L-1} \text{rank} (\boldsymbol{\Phi}^s), \quad (11)$$

where  $\boldsymbol{\Phi}^l = \text{diag}(\phi'(\mathbf{h}_1^l), \dots, \phi'(\mathbf{h}_M^l))$  is the Jacobian of the  $l$ 'th activation layer.

Observe that the upper bounds in proposition 4 are monotonically decreasing with respect to  $L$ . For example, we have

$$\text{rank} (\mathbf{W}^L \boldsymbol{\Phi}^{L-1} \dots \mathbf{W}^{l+1} \boldsymbol{\Phi}^l) \leq \text{rank} (\mathbf{W}^L \boldsymbol{\Phi}^{L-1} \dots \mathbf{W}^{l+2} \boldsymbol{\Phi}^{l+1}). \quad (12)$$

Based on these upper bounds,  $\frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \mathbf{w}^l}$  is potentially more singular for layers that are close to the input.

**Remark 5.** If  $\phi$  is ReLU, then the diagonal entries of  $\Phi^s$  is a binary vector, and  $\text{rank}(\Phi^s)$  is the number of activated neurons in the  $s$ 'th layer. In this case, the upper bound  $\min_{s=l}^{L-1} \text{rank}(\Phi^s)$  means the smallest number of activated neurons across all layers.

To understand  $d(\theta)$ , one can parameterize the DNN, locally, with only  $d(\theta)$  free parameters while maintaining the same predictive model. The log-likelihood is a function of these  $d(\theta)$  parameters, and therefore its Hessian has at most rank  $d(\theta)$ . In theory, one can only reparameterize  $\mathcal{M}$  so that at one single point  $\hat{\theta}$ , the screen and radical distributions are separated based on the coordinate chart. Such a chart may neither exist locally (in a neighborhood around  $\hat{\theta}$ ) nor globally.

The local dimensionality is not constant and may vary with  $\theta$ . The global topology of the neuromanifold is therefore like a *stratifold* [5, 18]. As  $\theta$  has a large dimensionality in DNNs, singularities are more likely to occur in  $\mathcal{M}$ . Compared to the notion of *intrinsic dimensionality* [43], our  $d(\theta)$  is well-defined mathematically rather than based on empirical evaluations. One can regard our local dimensionality as an upper bound of the intrinsic dimensionality, because a very small singular value of  $\mathcal{I}$  still counts towards the local dimensionality. Notice that random matrices have full rank with probability 1 [19].

We can regard small singular values (below a prescribed threshold  $\varepsilon > 0$ ) as  $\varepsilon$ -singular dimensions, and use  $\varepsilon$ -rank defined below to estimate the local dimensionality.

**Definition 2.** The  $\varepsilon$ -rank of the FIM  $\mathcal{I}(\theta)$  is the number of eigenvalues of  $\mathcal{I}(\theta)$  which is not less than some given  $\varepsilon > 0$ .

By definition, the  $\varepsilon$ -rank is a lower bound of the rank of the FIM, which depends on the  $\theta$ -parameterization — different parameterizations of the DNN may yield different  $\varepsilon$ -ranks of the corresponding FIM. If  $\varepsilon \rightarrow 0$ , the  $\varepsilon$ -rank of  $\mathcal{I}(\theta)$  becomes the true rank of  $\mathcal{I}(\theta)$  given by  $d(\theta)$ . The spectral density  $\rho_{\mathcal{I}}$  (probability distribution of the eigenvalues of  $\mathcal{I}(\theta)$ ) affects the  $\varepsilon$ -rank of  $\mathcal{I}(\theta)$  and the expected local dimensionality of  $\mathcal{M}$ . On the support of  $\rho_{\mathcal{I}}$ , the higher the probability of the region  $[0, \varepsilon)$ , the more likely  $\mathcal{M}$  is singular. By the Cramér-Rao lower bound, the variance of an unbiased 1D estimator  $\hat{\theta}$  must satisfy

$$\text{var}(\hat{\theta}) \geq \mathcal{I}(\theta)^{-1} \geq \frac{1}{\varepsilon}.$$

Therefore the  $\varepsilon$ -singular dimensions lead to a large variance of the estimator  $\hat{\theta}$ : a single observation  $\mathbf{x}_i$  carries little or no information regarding  $\theta$ , and it requires a large number of observations to achieve the same precision. The notion of thresholding eigenvalues close to zero may depend on the parameterization but the intrinsic ranks given by the local dimensionality are invariant.

In a DNN, there are several typical *sources of singularities*:

- First, if the neuron is saturated and gives constant output regardless of the input sample  $\mathbf{z}_i$ , then all dynamics of its input and output connections are in  $\text{Rad}(\mathcal{TM})$ .
- Second, two neurons in the same layer can have linearly dependent output, *e.g.* when they share the same weight vector and bias. They can be merged into one single neuron, as there exists redundancy in the original reparameterization.
- Third, if the activation function  $\phi(\cdot)$  is homogeneous, *e.g.* ReLU, then any neuron in the DNN induces a reparameterization by multiplying the input links by  $\alpha$  and output links by  $1/\alpha^k$  ( $k$  is the degree of homogeneity). This reparameterization corresponds to a null curve in the neuromanifold parameterized by  $\alpha$ .

- Fourth, certain structures such as recurrent neural networks (RNNs) suffer from vanishing gradient [23]. As the FIM is the variance of the gradient of the log-likelihood (known as variance of the score in statistics), its scale goes to zero along the dimensions associated with such structures.

It is meaningful to formally define the notion of “lightlike neuromanifold”. Using the geometric tools, related studies can be invariant w.r.t. neural network reparametrization. Moreover, the connection between neuromanifold and singular semi-Riemannian geometry, which is used in general relativity, is not yet widely adopted in machine learning. For example, the textbook [78] in singular statistics mainly used tools from algebraic geometry which is a different field.

Notice that the Fisher-Rao distance along a null curve is undefined because there the FIM is degenerate and there is no arc-length reparameterization along null curves [37].

## 5 General Formulation of Our Razor

In this section, we derive a new formula of MDL for DNNs, aiming to explain *how does the high dimensional DNN structure can have a short code length* of the given data? Notice that, this work focuses on the concept of model complexity but not the generalization bounds. We aim to show that the DNN model is intrinsically simple because it can be described shortly. The theoretical connection between generalization power and MDL is studied in PAC-Bayesian theory and PAC-MDL (see [9, 11, 24, 26, 83] and references therein). This is beyond the scope of this paper.

We derive a simple asymptotic formula for the case of large sample size and large network size. Therefore crude approximations are taken and the low-order terms are ignored, which are common practices in deriving information criteria [1, 70].

In the following, we will abuse  $p(\mathbf{x} \mid \boldsymbol{\theta})$  to denote the DNN model  $p(y \mid \mathbf{z}, \boldsymbol{\theta})$  for shorter equations and to be consistent with the introduction. Assume

(A2) The absolute values of the third-order derivatives of  $\log p(\mathbf{x} \mid \boldsymbol{\theta})$  w.r.t.  $\boldsymbol{\theta}$  are bounded by some constant.

(A3)  $\forall i, |\theta_i - \hat{\theta}_i| = O(1/\sqrt{M})$ , where  $O(\cdot)$  is the Bachmann–Landau’s big-O notation.

Recall that  $M$  is the width of the neural network. We consider that the neural networks weights have a order of  $O(1/\sqrt{M})$ . For example, if the input to a neuron follows the standard Gaussian distribution, then its weights with order  $O(1/\sqrt{M})$  guarantee the output is  $O(1)$ . In practice, this constraint can be enforced by clipping the weight vector to a prescribed range. This scaling is commonly used by random initialization techniques [21, 29] for training DNNs.

We rewrite the code length in eq. (1) based on the Taylor expansion of  $\log p(\mathbf{X} \mid \boldsymbol{\theta})$  at  $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$  up to the second order:

$$\begin{aligned} -\log p(\mathbf{X}) = & -\log \int_{\mathcal{M}} p(\boldsymbol{\theta}) \exp \left( \log p(\mathbf{X} \mid \hat{\boldsymbol{\theta}}) - \frac{N}{2} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathfrak{J}(\hat{\boldsymbol{\theta}}) (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right. \\ & \left. + O \left( N \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|^3 \right) \right) d\boldsymbol{\theta}. \end{aligned} \quad (13)$$

Notice that the first order term vanishes because  $\hat{\boldsymbol{\theta}}$  is a local optimum of  $\log p(\mathbf{X} \mid \boldsymbol{\theta})$ , and in the second order term,  $-N\mathfrak{J}(\hat{\boldsymbol{\theta}})$  is the Hessian matrix of the likelihood function  $\log p(\mathbf{X} \mid \boldsymbol{\theta})$  evaluated at  $\hat{\boldsymbol{\theta}}$ . At the MLE,  $\mathfrak{J}(\hat{\boldsymbol{\theta}}) \succeq 0$ , while in general the Hessian of the loss of a DNN evaluated at  $\boldsymbol{\theta} \neq \hat{\boldsymbol{\theta}}$  can have a negative spectrum [2, 68].

Through a change of variable  $\phi := \sqrt{N}(\theta - \hat{\theta})$ , the density of  $\phi$  is  $p(\phi) = \frac{1}{\sqrt{N}}p(\frac{\phi}{\sqrt{N}} + \hat{\theta})$  so that  $\int_{\mathcal{M}} p(\phi)d\phi = 1$ . In the integration in eq. (13), the term  $-\frac{N}{2}(\theta - \hat{\theta})^\top \mathfrak{J}(\hat{\theta})(\theta - \hat{\theta})$  has an order of  $O(\|\phi\|^2)$ . The cubic remainder term has an order of  $O(\frac{1}{\sqrt{N}}\|\phi\|^3)$ . If  $N$  is sufficiently large, this remainder can be ignored. Therefore we can write

$$-\log p(\mathbf{X}) \approx -\log p(\mathbf{X} | \hat{\theta}) - \log \mathbb{E}_p \exp \left( -\frac{N}{2}(\theta - \hat{\theta})^\top \mathfrak{J}(\hat{\theta})(\theta - \hat{\theta}) \right). \quad (14)$$

On the RHS, the first term measures the error of the model w.r.t. the observed data  $\mathbf{X}$ . The second term measures the model complexity. We have the following bound.

**Proposition 5.** *We have  $\forall \theta \in \mathcal{M}$ ,*

$$\begin{aligned} 0 &\leq -\log \mathbb{E}_p \exp \left( -\frac{N}{2}(\theta - \hat{\theta})^\top \mathfrak{J}(\hat{\theta})(\theta - \hat{\theta}) \right) \\ &\leq \frac{N}{2} \text{tr} \left( \mathfrak{J}(\hat{\theta}) \left( (\mu(\theta) - \hat{\theta})(\mu(\theta) - \hat{\theta})^\top + \text{cov}(\theta) \right) \right), \end{aligned}$$

where  $\mu(\theta)$  and  $\text{cov}(\theta)$  denote the mean and covariance matrix of the prior  $p(\theta)$ , respectively.

Therefore the complexity is always non-negative and its scale is bounded by the prior  $p(\theta)$ . The model has low complexity when  $\hat{\theta}$  is close to the mean of  $p(\theta)$  and/or when the variance of  $p(\theta)$  is small.

Consider the prior  $p(\theta) = \kappa(\theta) / \int_{\mathcal{M}} \kappa(\theta)d\theta$ , where  $\kappa(\theta) > 0$  is a positive measure on  $\mathcal{M}$  so that  $0 < \int_{\mathcal{M}} \kappa(\theta)d\theta < \infty$ . Based on the above approximation of  $-\log p(\mathbf{X})$ , we arrive at a general formula

$$\begin{aligned} \mathcal{O} &:= -\log p(\mathbf{X} | \hat{\theta}) + \log \int_{\mathcal{M}} \kappa(\theta)d\theta \\ &\quad - \log \int_{\mathcal{M}} \kappa(\theta) \exp \left( -\frac{N}{2}(\theta - \hat{\theta})^\top \mathfrak{J}(\hat{\theta})(\theta - \hat{\theta}) \right) d\theta, \end{aligned} \quad (15)$$

where “ $\mathcal{O}$ ” stands for Occam’s razor. Compared with previous formulations of MDL [7, 66, 67], eq. (15) relies on a quadratic approximation of the log-likelihood function and can be instantiated based on different assumptions of  $\kappa(\theta)$ . The non-normalized  $\kappa(\theta)$  in Bayesian coding serves a similar role to the luckiness function in NML coding [25], as they both incorporate prior knowledge to favor certain parameters in the parameter space  $\mathcal{M}$ .

Informally, the term  $\int_{\mathcal{M}} \kappa(\theta)d\theta$  gives the total capacity of models in  $\mathcal{M}$  specified by the *improper prior*  $\kappa(\theta)$ , up to constant scaling. For example, if  $\kappa(\theta)$  is uniform on a subregion in  $\mathcal{M}$ , then  $\int_{\mathcal{M}} \kappa(\theta)d\theta$  corresponds to the size of this region w.r.t. the base measure  $d\theta$ . The term  $\int_{\mathcal{M}} \kappa(\theta) \exp \left( -\frac{N}{2}(\theta - \hat{\theta})^\top \mathfrak{J}(\hat{\theta})(\theta - \hat{\theta}) \right) d\theta$  gives the model capacity specified by the posterior  $p(\theta | \mathbf{X}) \propto p(\theta)p(\mathbf{X} | \theta) \propto \kappa(\theta) \exp \left( -\frac{N}{2}(\theta - \hat{\theta})^\top \mathfrak{J}(\hat{\theta})(\theta - \hat{\theta}) \right)$ . It shrinks to zero when the number  $N$  of observations increases. The last two terms in eq. (15) is the log-ratio between the model capacity w.r.t. the prior and the capacity w.r.t. the posterior. A large log-ratio means there are many distributions on  $\mathcal{M}$  which have a relatively large value of  $\kappa(\theta)$  but a small value of  $\kappa(\theta) \exp \left( -\frac{N}{2}(\theta - \hat{\theta})^\top \mathfrak{J}(\hat{\theta})(\theta - \hat{\theta}) \right)$ . The associated model is considered to have a high complexity, meaning that only a small “percentage” of the models are helpful to describe the given data.

DNNs have a large amount of *symmetry*: the parameter space consists many pieces that looks exactly the same. This can be caused *e.g.* by permutate the neurons in the same layer. This

is a different *non-local* property than singularity that is a local differential property. Our  $\mathcal{O}$  is not affected by the model size caused by symmetry, because these symmetric models are both counted in the prior and the posterior, and the log-ratio in eq. (15) cancels out symmetric models. Formally,  $\mathcal{M}$  has  $\zeta$  symmetric pieces denoted by  $\mathcal{M}_1, \dots, \mathcal{M}_\zeta$ . Note any MLE on  $\mathcal{M}_i$  is mirrored on those  $\zeta$  pieces. Then both integrations on the RHS of eq. (15) are multiplied by a factor of  $\zeta$ . Therefore  $\mathcal{O}$  is invariant to symmetry.

## 6 Connection with $f$ -mean

In this section, we show that the model complexity terms in eqs. (14) and (15) can be studied based on the notion of  $f$ -mean, defined below.

**Definition 3** ( $f$ -mean). *Given a set  $\mathfrak{T} = \{t_i\}_{i=1}^n \subset \mathbb{R}$  and a continuous and strictly monotonous function  $f : \mathbb{R} \rightarrow \mathbb{R}$ , the  $f$ -mean of  $\mathfrak{T}$  is*

$$M_f(\mathfrak{T}) := f^{-1} \left( \frac{1}{n} \sum_{i=1}^n f(t_i) \right).$$

The  $f$ -mean, also known as the quasi-arithmetic mean was studied in [38, 51]: Thus they are also called Kolmogorov-Nagumo means [39]. By definition, the image of  $M_f(\mathfrak{T})$  under  $f$  is the arithmetic mean of the image of  $\mathfrak{T}$  under the same mapping. Therefore,  $M_f(\mathfrak{T})$  is in between the smallest and largest elements of  $\mathfrak{T}$ . If  $f(x) = x$ , then  $M_f$  becomes the arithmetic mean, which we denote as  $\bar{\mathfrak{T}}$ . We have the following bound.

**Lemma 6.** *Given a real matrix  $\mathbf{T} = (t_{ij})_{n \times m}$ , we use  $\mathbf{t}_i$  to denote  $i$ 'th row of  $\mathbf{T}$ , and  $\mathbf{t}_{:,j}$  to denote the  $j$ 'th column of  $\mathbf{T}$ . If  $f(t) = \exp(-t)$ , then*

$$M_f(\mathbf{T}) \leq \overline{\{M_f(\mathbf{t}_{:,1}), \dots, M_f(\mathbf{t}_{:,m})\}} \leq M_f(\{\bar{\mathbf{t}}_1, \dots, \bar{\mathbf{t}}_n\}) \leq \bar{\mathbf{T}},$$

where  $M_f(\mathbf{T})$  is the  $f$ -mean of all  $n \times m$  elements of  $\mathbf{T}$ , and  $\bar{\mathbf{T}}$  is their arithmetic mean.

Particular attention should be given to the second " $\leq$ ". If the arithmetic mean of each row is first evaluated, and then their  $f$ -mean is evaluated, we get an upper bound of the arithmetic mean of the  $f$ -mean of the columns. In simple terms, for  $f(t) = \exp(-t)$ , the  $f$ -mean of arithmetic mean is lower bounded by the arithmetic mean of the  $f$ -mean. The proof is straightforward from Jensen's inequality, and by noting that  $-\log \sum_i \exp(-t_i)$  is a concave function of  $\mathbf{t}$ . The last " $\leq$ " leads to a proof of the upper bound in proposition 5. Lemma 6 pertains to the mean of a matrix. It leads to bounds of the mean value of a bi-variable function w.r.t. a probability mass function or a probability density function, or a mix of both. This is straightforward and omitted.

**Remark 6.** All instances of " $\leq$ " in lemma 6 are derived from Jensen's inequality. Consequently, the gaps of these bounds shrink as the variance of the matrix elements decreases, where the specific way of measuring the variance depends on the particular " $\leq$ ". For instance, the gap associated with the second " $\leq$ " becomes smaller as the variance across each row  $\mathbf{t}_i$  reduces.

**Remark 7.** The second complexity term on the RHS of eq. (14) is the  $f$ -mean of the quadratic term  $\frac{N}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathfrak{J}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$  w.r.t. the prior  $p(\boldsymbol{\theta})$ , where  $f(t) = \exp(-t)$ .

Based on the spectrum decomposition  $\mathfrak{J}(\hat{\boldsymbol{\theta}}) = \sum_{j=1}^{\text{rank}(\mathfrak{J}(\hat{\boldsymbol{\theta}}))} \lambda_j^+ \mathbf{v}_j \mathbf{v}_j^\top$ , where the positive eigenvalues  $\lambda_j^+ := \lambda_j(\mathfrak{J}(\hat{\boldsymbol{\theta}}))$  and the eigenvectors  $\mathbf{v}_j := \mathbf{v}_j(\hat{\boldsymbol{\theta}})$  depend on the MLE  $\hat{\boldsymbol{\theta}}$ , we further write

this term as

$$\frac{N}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathfrak{J}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) = \sum_{j=1}^{\text{rank}(\mathfrak{J}(\hat{\boldsymbol{\theta}}))} \frac{\lambda_j^+}{\text{tr}(\mathfrak{J}(\hat{\boldsymbol{\theta}}))} \cdot \frac{N}{2} \text{tr}(\mathfrak{J}(\hat{\boldsymbol{\theta}})) \langle \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}, \mathbf{v}_j \rangle^2.$$

By lemma 6, we have

$$\begin{aligned} & -\log \mathbb{E}_p \exp \left( -\frac{N}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathfrak{J}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right) \\ & \geq - \sum_{j=1}^{\text{rank}(\mathfrak{J}(\hat{\boldsymbol{\theta}}))} \frac{\lambda_j^+}{\text{tr}(\mathfrak{J}(\hat{\boldsymbol{\theta}}))} \log \mathbb{E}_p \exp \left( -\frac{N}{2} \text{tr}(\mathfrak{J}(\hat{\boldsymbol{\theta}})) \langle \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}, \mathbf{v}_j \rangle^2 \right), \end{aligned}$$

where the  $f$ -mean and the mean w.r.t.  $\frac{\lambda_j^+}{\text{tr}(\mathfrak{J}(\hat{\boldsymbol{\theta}}))}$  is swapped on the RHS. Denote  $\varphi_j = \langle \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}, \mathbf{v}_j \rangle$ , which in matrix form is written as  $\boldsymbol{\varphi} = \mathbf{V}^\top(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$ .  $\mathbf{V}$  has orthonormal columns and the  $j$ 'th column of  $\mathbf{V}$  is  $\mathbf{v}_j$ . The prior of  $\boldsymbol{\varphi}$  is given by  $p(\mathbf{V}\boldsymbol{\varphi} + \hat{\boldsymbol{\theta}})$ . Then

$$-\log \mathbb{E}_p \exp \left( -\frac{N}{2} \text{tr}(\mathfrak{J}(\hat{\boldsymbol{\theta}})) \langle \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}, \mathbf{v}_j \rangle^2 \right) = -\log \mathbb{E}_{p(\varphi_j)} \exp \left( -\frac{N}{2} \text{tr}(\mathfrak{J}(\hat{\boldsymbol{\theta}})) \varphi_j^2 \right),$$

where the RHS is a variance-like measure of  $p(\boldsymbol{\theta})$  up to scaling, as it is the  $f$ -mean of  $\frac{N}{2} \text{tr}(\mathfrak{J}(\hat{\boldsymbol{\theta}})) \varphi_j^2$  evaluated at point  $\hat{\boldsymbol{\theta}}$  along the direction  $\mathbf{v}_j$ . In summary, we get a lower bound of the model complexity, which is tighter than the lower bound in proposition 5, given by

$$\begin{aligned} & -\log \mathbb{E}_p \exp \left( -\frac{N}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathfrak{J}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right) \\ & \geq - \sum_{j=1}^{\text{rank}(\mathfrak{J}(\hat{\boldsymbol{\theta}}))} \frac{\lambda_j^+}{\text{tr}(\mathfrak{J}(\hat{\boldsymbol{\theta}}))} \log \mathbb{E}_{p(\varphi_j)} \exp \left( -\frac{N}{2} \text{tr}(\mathfrak{J}(\hat{\boldsymbol{\theta}})) \varphi_j^2 \right). \end{aligned} \quad (16)$$

The RHS is determined by the quantity  $\frac{N}{2} \text{tr}(\mathfrak{J}(\hat{\boldsymbol{\theta}})) \varphi_j^2$  after evaluating the  $f$ -mean and some weighted mean, where  $\varphi_j$  is an orthogonal transformation of the local coordinates  $\theta_i$  based on the spectrum of  $\mathfrak{J}(\hat{\boldsymbol{\theta}})$ . Recall that the trace of the observed FIM  $\mathfrak{J}(\hat{\boldsymbol{\theta}})$  means the overall amount of information a random observation contains w.r.t. the underlying model. Given the same sample size  $N$ , a larger  $\text{tr}(\mathfrak{J}(\hat{\boldsymbol{\theta}}))$  indicates that the samples are more informative and the likelihood is more sensitive to the choice of the parameters on  $\mathcal{M}$ . Consequently, it is reasonable to regard the model as more complex, because small changes of model parameters more easily lead to different representations.

The bound in eq. (16) is tight when the variance of  $\frac{N}{2} \text{tr}(\mathfrak{J}(\hat{\boldsymbol{\theta}})) \varphi_j^2$  w.r.t. the discrete distribution  $\frac{\lambda_j^+}{\text{tr}(\mathfrak{J}(\hat{\boldsymbol{\theta}}))}$  is small. In the case when  $\mathfrak{J}(\hat{\boldsymbol{\theta}})$  is rank-one, “ $\geq$ ” becomes “ $=$ ”. In practice, the FIM of DNNs exhibits a pathological spectrum [36], where most eigenvalues of  $\mathfrak{J}(\hat{\boldsymbol{\theta}})$  are near zero, with a small fraction taking large values. This means that the  $\epsilon$ -rank of  $\text{tr}(\mathfrak{J}(\hat{\boldsymbol{\theta}}))$  is limited, and the distribution  $\frac{\lambda_j^+}{\text{tr}(\mathfrak{J}(\hat{\boldsymbol{\theta}}))}$  has lower variance as compared to a uniform spectrum. This distinctive property of DNNs offers some basis for considering the lower bound in eq. (16) as a proxy of the model complexity.

As  $\hat{\boldsymbol{\theta}}$  is the MLE, we have  $\mathfrak{J}(\hat{\boldsymbol{\theta}}) = \hat{\mathcal{I}}(\hat{\boldsymbol{\theta}})$ . Recall from eq. (7) that the FIM  $\hat{\mathcal{I}}(\hat{\boldsymbol{\theta}})$  is a numerical average over all observed samples. We can have alternative lower bounds of the model complexity based on lemma 6:

$$\begin{aligned}
& -\log \mathbb{E}_p \exp \left( -\frac{N}{2} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathfrak{J}(\hat{\boldsymbol{\theta}}) (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right) \\
& \geq -\frac{1}{N} \sum_{i=1}^N \log \mathbb{E}_p \exp \left( -\frac{N}{2} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \left( \frac{\partial \mathbf{h}^L(\mathbf{z}_i)}{\partial \boldsymbol{\theta}} \right)^\top \mathbf{C}_i \frac{\partial \mathbf{h}^L(\mathbf{z}_i)}{\partial \boldsymbol{\theta}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right) \\
& \geq -\frac{1}{N} \sum_{i=1}^N \mathbb{E}_{p(y|\mathbf{z}_i)} \log \mathbb{E}_p \exp \left( -\frac{N}{2} \left[ \frac{\partial \log p(y|\mathbf{z}_i)}{\partial \boldsymbol{\theta}^\top} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right]^2 \right). \tag{17}
\end{aligned}$$

The bounds are obtained by swapping the  $f$ -mean with the numerical average of the samples, and by swapping  $f$ -mean with the expectation w.r.t.  $p(y|\mathbf{z}_i)$ . Therefore the model complexity can be bounded by the average scale of the vector  $\frac{\partial \mathbf{h}^L(\mathbf{z}_i)}{\partial \boldsymbol{\theta}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$ , where  $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$ . Note that  $\frac{\partial \mathbf{h}^L(\mathbf{z}_i)}{\partial \boldsymbol{\theta}}$  is the parameter-output Jacobian matrix, or a linear approximation of the neural network mapping  $\boldsymbol{\theta} \rightarrow \mathbf{h}^L$ . The complexity lower bounds in eq. (17) mean how the local parameter change  $(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$  w.r.t. the prior  $p(\boldsymbol{\theta})$  affect the output. If the output is sensitive to these parameter variations, then the model is considered to have high complexity. In summary, the  $f$ -mean offers a powerful tool to analyze our model complexity and obtain its approximations.

## 7 The Razor based on Gaussian Prior

The simplest and most widely-used choice of the prior  $p(\boldsymbol{\theta})$  is the Gaussian prior (see *e.g.* [35, 46] among many others). In eq. (15), we set

$$\kappa(\boldsymbol{\theta}) = \exp \left( -\boldsymbol{\theta}^\top \text{diag} \left( \frac{1}{\boldsymbol{\sigma}} \right) \boldsymbol{\theta} \right),$$

where  $\text{diag}(\cdot)$  means a diagonal matrix constructed with given entries, and  $\boldsymbol{\sigma} > 0$  (elementwisely). Equivalently, the associated prior is  $p_G(\boldsymbol{\theta}) = \mathcal{G}(\boldsymbol{\theta} | \mathbf{0}, \text{diag}(\boldsymbol{\sigma}))$ , meaning a Gaussian distribution with mean  $\mathbf{0}$  and covariance matrix  $\text{diag}(\boldsymbol{\sigma})$ . We further assume

(A4)  $\mathcal{M}$  has a global coordinate chart and  $\mathcal{M}$  is homeomorphic to  $\mathbb{R}^D$ .

(A5) Regardless of  $D$ ,  $\hat{\boldsymbol{\theta}}^\top \text{diag}(\frac{1}{\boldsymbol{\sigma}}) \hat{\boldsymbol{\theta}} < \infty$ .

Assumption (A4) enables us to define a Gaussian distribution in a global coordinate system, which typically represents the neural network weights and biases. By assumption (A5), the MLE  $\hat{\boldsymbol{\theta}}$  has a non-zero probability under the Gaussian prior.

From eq. (15), we get a closed form expression of the razor

$$\begin{aligned}
\mathcal{O}_G & := -\log p(\mathbf{X} | \hat{\boldsymbol{\theta}}) + \frac{\text{rank}(\mathfrak{J}(\hat{\boldsymbol{\theta}}))}{2} \log N \\
& + \frac{1}{2} \sum_{i=1}^{\text{rank}(\mathfrak{J}(\hat{\boldsymbol{\theta}}))} \log \left( \lambda_i^+ \left( \mathfrak{J}(\hat{\boldsymbol{\theta}}) \text{diag}(\boldsymbol{\sigma}) \right) + \frac{1}{N} \right) + O(1), \tag{18}
\end{aligned}$$



where  $\lambda_i^+(\mathfrak{J}(\hat{\theta})\text{diag}(\sigma))$  denotes the  $i$ 'th positive eigenvalue of  $\mathfrak{J}(\hat{\theta})\text{diag}(\sigma)$ . Notice that  $\mathfrak{J}(\hat{\theta})\text{diag}(\sigma)$  and  $\text{diag}(\sqrt{\sigma})\mathfrak{J}(\hat{\theta})\text{diag}(\sqrt{\sigma})$  share the same set of non-zero eigenvalues, and the latter is psd with  $\text{rank}(\mathfrak{J}(\hat{\theta}))$  positive eigenvalues.

In our razor expressions, all terms that do not scale with the sample size  $N$  or the number of parameters  $D$  are discarded. The first two terms on the RHS are similar to BIC [70] up to scaling. The complexity terms (second and third terms on the RHS of eq. (18)) do not scale with  $D$  but are bounded by the rank of the Hessian, or the observed FIM. In other words, the radical distribution associated with zero-eigenvalues of  $\mathfrak{J}(\hat{\theta})$  does not affect the model complexity. This is different from previous formulations of MDL [7, 66, 67] and BIC [70]. For example, the 2nd term on the RHS of eq. (2) increases linearly with  $D$ , while the 2nd term on the RHS of eq. (18) increases linearly with  $\text{rank}(\mathfrak{J}(\hat{\theta})) \leq D$ .

Interestingly, if  $\lambda_i^+(\mathfrak{J}(\hat{\theta})) < \frac{1}{\sigma_{\max}}(1 - \frac{1}{N})$ , the third term on the RHS of eq. (18) becomes negative. In the extreme case when  $\lambda_i^+(\mathfrak{J}(\hat{\theta}))$  tends to zero,  $\frac{1}{2} \log(\sigma_{\max} \lambda_i^+(\mathfrak{J}(\hat{\theta})) + \frac{1}{N}) \rightarrow -\frac{1}{2} \log N$ , which cancels out the model complexity penalty in the term  $\frac{\text{rank}(\mathfrak{J}(\hat{\theta}))}{2} \log N$ . In other words, the corresponding parameter is added free (without increasing the model complexity). Informally, we call similar terms that are helpful in decreasing the complexity while contributing to model flexibility the *negative complexity*.

We have

$$\begin{aligned} \sum_{i=1}^{\text{rank}(\mathfrak{J}(\hat{\theta}))} \log\left(\sigma_{\min} \lambda_i^+(\mathfrak{J}(\hat{\theta})) + \frac{1}{N}\right) &\leq \sum_{i=1}^{\text{rank}(\mathfrak{J}(\hat{\theta}))} \log\left(\lambda_i^+(\mathfrak{J}(\hat{\theta})\text{diag}(\sigma)) + \frac{1}{N}\right) \\ &\leq \sum_{i=1}^{\text{rank}(\mathfrak{J}(\hat{\theta}))} \log\left(\sigma_{\max} \lambda_i^+(\mathfrak{J}(\hat{\theta})) + \frac{1}{N}\right), \end{aligned}$$

where  $\sigma_{\max}$  and  $\sigma_{\min}$  denote the largest and smallest elements of  $\sigma$ , respectively. Therefore the term can be bounded based on the spectrum of  $\mathfrak{J}(\hat{\theta})$ . If  $\sigma = \sigma \mathbf{1}$ , where  $\sigma > 0$ , then both of the above " $\leq$ "s become equalities. In this case, we let  $D \rightarrow \infty$  and rewrite the razor in terms of the spectrum density  $\rho_{\mathcal{I}}(\lambda)$  of  $\mathfrak{J}(\hat{\theta})$ :

$$\mathcal{O}_G = -\log p(\mathbf{X} | \hat{\theta}) + \frac{\text{rank}(\mathfrak{J}(\hat{\theta}))}{2} \mathbb{E}_{\rho_{\mathcal{I}}(\lambda)} \log(N\sigma\lambda + 1) + O(1). \quad (19)$$

Note  $\text{rank}(\mathfrak{J}(\hat{\theta})) = \hat{d}(\hat{\theta})$  is the local dimensionality at  $\hat{\theta}$ , which could have a smaller order than  $D$ , especially when  $N$  is finite. If  $\rho_{\mathcal{I}}(\lambda)$  is highly concentrated around 0 as shown in [36], then the expectation of  $\log(N\sigma\lambda + 1)$  can be roughly approximated as zero. This approximation is also linked to the low intrinsic complexity of DNNs.

The Gaussian prior  $p_G$  is helpful to give simple and intuitive expressions of  $\mathcal{O}_G$ . However, the problem in choosing  $p_G$  is two fold. First, it is not invariant. Under a reparametrization (*e.g.* normalization or centering techniques), the Gaussian prior in the new parameter system does not correspond to the original prior. Second, it double counts equivalent models. Because of the many singularities of the neuromanifold, a small dynamic in the parameter system may not change the prediction model. However, the Gaussian prior is defined in a real vector space and may not fit in this singular semi-Riemannian structure. Gaussian distributions are defined on Riemannian manifolds [69] which lead to potential extensions of the discussed prior  $p_G(\theta)$ .

## 8 The Razor based on Jeffreys' Non-informative Prior

Jeffreys' prior is specified by  $p_J(\boldsymbol{\theta}) \propto \sqrt{|\mathcal{I}(\boldsymbol{\theta})|}$ . It is *non-informative* in the sense that no neural network model  $\boldsymbol{\theta}_1$  is prioritized over any other model  $\boldsymbol{\theta}_2$ . It is invariant to the choice of the coordinate system. Under a reparameterization  $\boldsymbol{\theta} \rightarrow \boldsymbol{\eta}$ ,

$$\sqrt{|\mathcal{I}(\boldsymbol{\eta})|}d\boldsymbol{\eta} = \sqrt{\left| \left( \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}} \right)^\top \mathcal{I}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}} \right|} \cdot d\boldsymbol{\eta} = \sqrt{|\mathcal{I}(\boldsymbol{\theta})|} \cdot \left( \left| \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}} \right| d\boldsymbol{\eta} \right) = \sqrt{|\mathcal{I}(\boldsymbol{\theta})|}d\boldsymbol{\theta},$$

showing that the Riemannian volume element is the same in different coordinate systems. Unfortunately, the Jeffreys' prior is *not* well defined on the lightlike neuromanifold  $\mathcal{M}$ , where the metric  $\mathcal{I}(\boldsymbol{\theta})$  is degenerate and  $\sqrt{|\mathcal{I}(\boldsymbol{\theta})|}$  becomes zero. The stratifold structure of  $\mathcal{M}$ , where  $d(\boldsymbol{\theta})$  varying with  $\boldsymbol{\theta} \in \mathcal{M}$ , makes it difficult to properly define the base measure  $d\boldsymbol{\theta}$  and integrate functions as in eq. (15). From a mathematical standpoint, one has to integrate on the screen distribution  $\mathcal{S}(\mathcal{T}\mathcal{M})$ , which has a Riemannian structure. We refer the reader to [34, 74] for other extensions of Jeffreys' prior.

In this paper, we take a simple approach by examining a submanifold of  $\mathcal{M}$  denoted as  $\widetilde{\mathcal{M}}$  and parameterized by  $\boldsymbol{\xi}$ , which has a Riemannian metric  $\mathcal{I}(\boldsymbol{\xi}) \succ 0$  that is induced by the FIM  $\mathcal{I}(\boldsymbol{\theta}) \succeq 0$  and the mapping  $\boldsymbol{\xi} \rightarrow \boldsymbol{\theta}$ . The dimensionality of  $\widetilde{\mathcal{M}}$  is upper-bounded by the local dimensionality  $d(\boldsymbol{\theta})$ . Intuitively, any infinitesimal dynamic on  $\widetilde{\mathcal{M}}$  means such a change of neural network parameters that leads to a non-zero change of the global predictive model  $\mathbf{z} \rightarrow y$ . For example,  $\widetilde{\mathcal{M}}$  can be defined based on a subset of sensitive parameters. In theory, we would like to construct  $\widetilde{\mathcal{M}}$  so that it is representative of  $\mathcal{M}$ , meaning that  $\dim(\widetilde{\mathcal{M}})$  is close to the local dimensionality  $d(\boldsymbol{\theta})$ , and at the same time  $\widetilde{\mathcal{M}}$  remains Riemannian. The following results are constrained to the choice of the submanifold  $\widetilde{\mathcal{M}}$ .

In eq. (15), let  $\kappa(\boldsymbol{\xi}) = \sqrt{|\mathcal{I}(\boldsymbol{\xi})|}$ . We further assume

$$(A6) \quad 0 < \int_{\widetilde{\mathcal{M}}} \sqrt{|\mathcal{I}(\boldsymbol{\xi})|} d\boldsymbol{\xi} < \infty;$$

meaning that the Riemannian volume of  $\widetilde{\mathcal{M}}$  is bounded. After straightforward derivations, we arrive at

$$\begin{aligned} \mathcal{O}_J(\boldsymbol{\xi}) &= -\log p(\mathbf{X} | \hat{\boldsymbol{\xi}}) + \log \int_{\widetilde{\mathcal{M}}} \sqrt{|\mathcal{I}(\boldsymbol{\xi})|} d\boldsymbol{\xi} \\ &\quad - \log \int_{\widetilde{\mathcal{M}}} \exp \left( -\frac{N}{2} (\boldsymbol{\xi} - \hat{\boldsymbol{\xi}})^\top \mathfrak{J}(\hat{\boldsymbol{\xi}}) (\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}) \right) \sqrt{|\mathcal{I}(\boldsymbol{\xi})|} d\boldsymbol{\xi} \\ &= -\log p(\mathbf{X} | \hat{\boldsymbol{\xi}}) + \log \int_{\widetilde{\mathcal{M}}} \sqrt{|\mathcal{I}(\boldsymbol{\xi})|} d\boldsymbol{\xi} - \log \int_{\widetilde{\mathcal{M}}} \omega(\boldsymbol{\xi}) \sqrt{|\mathcal{I}(\boldsymbol{\xi})|} d\boldsymbol{\xi}, \end{aligned} \quad (20)$$

where  $\omega(\boldsymbol{\xi}) := \exp \left( -\frac{N}{2} (\boldsymbol{\xi} - \hat{\boldsymbol{\xi}})^\top \mathfrak{J}(\hat{\boldsymbol{\xi}}) (\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}) \right)$  is a shorthand. Let us examine the meaning of  $\mathcal{O}_J(\boldsymbol{\xi})$ . As  $\mathcal{I}(\boldsymbol{\xi})$  is the Riemannian metric of  $\widetilde{\mathcal{M}}$  based on information geometry,  $\sqrt{|\mathcal{I}(\boldsymbol{\xi})|}d\boldsymbol{\xi}$  is a Riemannian volume element (volume form). In the second term on the RHS of eq. (20), the integral  $\int_{\widetilde{\mathcal{M}}} \sqrt{|\mathcal{I}(\boldsymbol{\xi})|} d\boldsymbol{\xi}$  is the information volume, or the total “number” of different DNN models [50] on  $\widetilde{\mathcal{M}}$ . In the last (third) term, because  $0 < \omega(\boldsymbol{\xi}) \leq 1$ , the integral on the LHS of

$$\int_{\widetilde{\mathcal{M}}} \omega(\boldsymbol{\xi}) \sqrt{|\mathcal{I}(\boldsymbol{\xi})|} d\boldsymbol{\xi} \leq \int_{\widetilde{\mathcal{M}}} \sqrt{|\mathcal{I}(\boldsymbol{\xi})|} d\boldsymbol{\xi}$$

means a “weighted volume” of  $\widetilde{\mathcal{M}}$ , where the positive weights  $\omega(\boldsymbol{\xi})$  are determined by the observed FIM  $\mathfrak{J}(\hat{\boldsymbol{\xi}})$ . Combining these two terms, the model complexity is the log-ratio between the unweighted volume and the weighted volume and is lower bounded by 0.

Assume the spectrum decomposition  $\mathfrak{J}(\hat{\xi}) = \mathbf{Q} \text{diag} \left( \lambda_i^+(\mathfrak{J}(\hat{\xi})) \right) \mathbf{Q}^\top$ , where  $\mathbf{Q}$  has orthonormal columns, and  $\lambda_i^+(\mathfrak{J}(\hat{\xi}))$  are the positive eigenvalues of  $\mathfrak{J}(\hat{\xi})$ . Equation (20) becomes

$$\begin{aligned} \mathcal{O}_J(\zeta) = & -\log p(\mathbf{X} | \hat{\zeta}) + \log \int_{\widetilde{\mathcal{M}}} \sqrt{|\mathcal{I}(\zeta)|} d\boldsymbol{\xi} \\ & - \log \int_{\widetilde{\mathcal{M}}} \exp \left( -\frac{N}{2} \sum_{i=1}^{\text{rank}(\mathfrak{J}(\hat{\xi}))} \lambda_i^+(\mathfrak{J}(\hat{\xi})) (\zeta_i - \hat{\zeta}_i)^2 \right) \sqrt{|\mathcal{I}(\zeta)|} d\zeta, \end{aligned} \quad (21)$$

where  $\zeta = \mathbf{Q}^\top \boldsymbol{\xi}$  is an orthogonal transformation of  $\boldsymbol{\xi}$ , and  $\mathcal{O}_J$  is invariant to such transformations. If an eigenvalue of  $\mathfrak{J}(\hat{\xi})$  has an order of  $o(\frac{1}{N})$ , the last two terms in eq. (21) cancel out in the corresponding direction, meaning no complexity is added. This is similar to how the positive and negative complexity terms cancel out in eq. (18) – small eigenvalues of  $\mathfrak{J}(\hat{\xi})$  are helpful to enhance the representation power of DNNs without increasing the model complexity. Only eigenvalues that are large enough contribute significantly to the model complexity.

In the rest of this section, we connect our  $\mathcal{O}_J$  with previous formulations of MDL [7, 50]. Observe that  $\omega(\boldsymbol{\xi})$  in eq. (20) resembles a Gaussian density up to a scaling factor. If  $\mathfrak{J}(\hat{\xi})$  has full rank, we can further write

$$\begin{aligned} \mathcal{O}_J(\widetilde{\mathcal{M}}) = & -\log p(\mathbf{X} | \hat{\xi}) + \frac{\dim(\widetilde{\mathcal{M}})}{2} \log \frac{N}{2\pi} + \log \int_{\widetilde{\mathcal{M}}} \sqrt{|\mathcal{I}(\boldsymbol{\xi})|} d\boldsymbol{\xi} \\ & - \log \int_{\widetilde{\mathcal{M}}} G \left( \boldsymbol{\xi} | \hat{\xi}, \frac{1}{N} \mathfrak{J}^{-1}(\hat{\xi}) \right) \frac{|\mathcal{I}(\boldsymbol{\xi})|^{1/2}}{|\mathfrak{J}(\hat{\xi})|^{1/2}} d\boldsymbol{\xi}. \end{aligned} \quad (22)$$

By assumption (A6), the RHS of eq. (20) is well defined, while the RHS of eq. (22) is only meaningful for a full rank  $\mathfrak{J}(\hat{\xi})$ . If  $\mathfrak{J}(\hat{\xi})$  is *not* invertible, one can consider the limit case when the zero eigenvalues of  $\mathfrak{J}(\hat{\xi})$  are replaced by a small  $\epsilon > 0$  and still enjoy the expression in eq. (22). One has to note that

$$\int_{\widetilde{\mathcal{M}}} G \left( \boldsymbol{\xi} | \hat{\xi}, \frac{1}{N} \mathfrak{J}^{-1}(\hat{\xi}) \right) d\boldsymbol{\xi} \leq 1,$$

as the integral is over  $\widetilde{\mathcal{M}}$  which is a subset of  $\mathbb{R}^{\dim(\widetilde{\mathcal{M}})}$ . The last term on the RHS of eq. (22) resembles an expectation w.r.t. a Gaussian distribution centered at  $\hat{\xi}$  on  $\widetilde{\mathcal{M}}$ , except that the Gaussian density may be truncated by  $\widetilde{\mathcal{M}}$ . One can therefore take the rough approximation based on the mean of the Gaussian:

$$-\log \int_{\widetilde{\mathcal{M}}} G \left( \boldsymbol{\xi} | \hat{\xi}, \frac{1}{N} \mathfrak{J}^{-1}(\hat{\xi}) \right) \frac{|\mathcal{I}(\boldsymbol{\xi})|^{1/2}}{|\mathfrak{J}(\hat{\xi})|^{1/2}} d\boldsymbol{\xi} \approx \frac{1}{2} \log \frac{|\mathfrak{J}(\hat{\xi})|}{|\mathcal{I}(\hat{\xi})|}. \quad (23)$$

Under this approximation, eq. (22) gives the MDL criterion discussed in [7, 50], where the term on the RHS of eq. (23) is interpreted as a penalty to models that lack robustness and are sensitive to the choice of parameters. We therefore consider the spectrum of both matrices  $\mathcal{I}(\boldsymbol{\xi})$  and  $\mathfrak{J}(\boldsymbol{\xi})$ , noting that in the large sample limit  $N \rightarrow \infty$ , they become identical. Because of the finite  $N$ , the observed FIM  $\mathfrak{J}(\boldsymbol{\xi})$  is singular in potentially many directions. The true FIM  $\mathcal{I}(\hat{\xi})$  can be regarded as the sum of the observed FIM  $\mathfrak{J}(\hat{\xi})$  and the FIM w.r.t. unobserved samples, up to a scaling factor. Based on how  $\widetilde{\mathcal{M}}$  is constructed,  $\mathcal{I}(\hat{\xi}) \succ 0$  is positive definite and suffers less from singularities. In the directions where  $\mathfrak{J}(\hat{\xi})$  is nearly singular, the log-ratio  $\log |\mathfrak{J}(\hat{\xi})|/|\mathcal{I}(\hat{\xi})|$  contributes significantly and negatively to the model complexity. As a result, eq. (23) serves as a

negative complexity term and explains how singularities of  $\mathfrak{J}(\hat{\xi})$  correspond to the simplicity of DNNs.

Compared with  $\mathcal{O}_G$ ,  $\mathcal{O}_J$  is based on a more accurate geometric modeling. However, it is hard to be computed numerically, as it depends on how  $\hat{\mathcal{M}}$  is constructed, and  $\mathcal{I}(\xi)$  and  $p(z)$  which are unknown due to limited observations. Despite that  $\mathcal{O}_G$  and  $\mathcal{O}_J$  have different expressions, their preference to model dimensions with small Fisher information (as in DNNs) is similar.

Hence, we can conclude that *the intrinsic complexity of a DNN is affected by the singularity and spectral properties of the Fisher information matrix.*

## 9 Related Work

The dynamics of supervised learning of a DNN describes a trajectory on the parameter space of the DNN geometrically modeled as a manifold when endowed with the FIM (e.g., ordinary/natural gradient descent learning the parameters of a MLP). Singular regions of the neuromanifold [79] correspond to non-identifiable parameters with rank-deficient FIM, and the learning trajectory typically exhibit chaotic patterns [4] with the singularities which translate into slowdown plateau phenomena when plotting the loss function value against time. By building an elementary singular DNN, [4] (and references therein) show that stochastic gradient descent learning dynamics yields a Milnor-type attractor with both attractor/repulser subregions where the learning trajectory is attracted in the attractor region, then stay a long time there before escaping through the repulser region. The natural gradient is shown to be free of critical slowdowns. Furthermore, although DNNs have potentially many singular regions, it is shown that the interaction of elementary units cancels out the Milnor-type attractors. It was shown [55] that skip connections are helpful to reduce the effect of singularities. However, a full understanding of the learning dynamics [81] for generic DNN architectures with multiple output values or recurrent DNNs is yet to be investigated.

The MDL criterion has undergone several fundamental revisions, such as the original crude MDL [66] and refined MDL with the introduction of stochastic complexity [8, 67], and the NML [65, 71] as a modern refinement. We refer the reader to the book [25] for a comprehensive introduction to this area and [24] for a recent review. We should also mention that the relationship between MDL and generalization has been explored in the PAC-MDL framework [9, 11, 25, 26, 83]. See [24] (section 6.4) for related remarks.

The relationship between MDL and information geometry is well established [3, 7, 49, 50, 67]. For example, they both rely on fundamental concepts such as the Fisher information. The geometric complexity of statistical models is commonly formulated using tools from information geometry [1, 7, 49, 67]. The stochastic complexity in singular mixture models can be bounded [80] and therefore is smaller than that of regular models. On this line of research, our derivations based on a Taylor expansion of the log-likelihood are similar to [7]. This technique is also used for deriving natural gradient optimization for deep learning [4, 45, 57].

Recently MDL has been ported to deep learning [10] focusing on variational methods. Practical techniques such as weight sharing [20], binarization [32], model compression [14], etc., follows similar principles of MDL. In the same community, many efforts have been made to develop a theory of deep learning, for example, based on PAC-Bayes theory [53], statistical learning theory [82], algorithmic information theory [76], information geometry [44], geometry of the DNN mapping [62], or through defining an intrinsic dimensionality [43] that is much smaller than the network size. Our analysis depends on  $\mathfrak{J}(\hat{\theta})$  and therefore is related to the flatness/sharpness of the local minima [15, 30], which is known to affect generalization. Using advanced mathematical tools such as random matrix theory, investigations are conducted on the spectrum of the input-output Jacobian matrix [59], the Hessian matrix w.r.t. the neural network weights [58], and the

FIM [27, 35, 36, 56, 60].

## 10 Conclusion

We considered mathematical tools from singular semi-Riemannian geometry to study the locally varying intrinsic dimensionality of a deep learning model. These models fall in the category of non-identifiable parameterizations. We take a meaningful step to quantify *geometric singularity* through the notion of local dimensionality  $d(\boldsymbol{\theta})$  yielding a singular semi-Riemannian neuromanifold with varying metric signature. We show that  $d(\boldsymbol{\theta})$  grows at most linearly with the sample size  $N$ . Recent findings show that the spectrum of the Fisher information matrix shifts towards  $0^+$  with a large number of small eigenvalues. We show that these singular dimensions help to reduce the model complexity. As a result, we contribute a simple and general MDL for deep learning. It provides theoretical insights on the description length of DNNs. DNNs benefit from a high-dimensional parameter space in that the singular dimensions impose a negative complexity to describe the data, which can be seen in our derivations based on Gaussian and Jeffreys' priors. How the short description length is connected to the empirical performance of DNNs and related generalization bounds require further examinations. This is not addressed in the current work. A more careful analysis of the FIM's spectrum, e.g. through considering higher-order terms, could give more practical formulations of the proposed criterion. We leave empirical studies as potential future work.

## Appendix A Proof of $\mathfrak{J}(\hat{\boldsymbol{\theta}}) = \hat{\mathcal{I}}(\hat{\boldsymbol{\theta}})$

*Proof.*

$$p(y_i | \mathbf{z}_i, \boldsymbol{\theta}) = \exp \left( \text{OneHot}(y_i)^\top \mathbf{h}^L(\mathbf{z}_i) - \log \sum_j \exp(h_j^L(\mathbf{z}_i)) \right),$$

where  $\text{OneHot}(y)$  is the binary vector with the same dimensionality as  $\mathbf{h}^L(\mathbf{z}_i)$ , with the  $y$ 'th bit set to 1 and the rest bits set to 0. Therefore,

$$\frac{\partial \log p(y_i | \mathbf{z}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left[ \frac{\partial \mathbf{h}^L}{\partial \boldsymbol{\theta}} \right]^\top [\text{OneHot}(y_i) - \text{SoftMax}(\mathbf{h}^L(\mathbf{z}_i))].$$

Therefore,

$$\frac{\partial^2 \log p(y_i | \mathbf{z}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = \sum_j [\text{OneHot}(y_i) - \text{SoftMax}(\mathbf{h}^L(\mathbf{z}_i))]_j \frac{\partial^2 \mathbf{h}_j^L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} - \left[ \frac{\partial \mathbf{h}^L}{\partial \boldsymbol{\theta}} \right]^\top \cdot \mathbf{C}_i \cdot \frac{\partial \mathbf{h}^L}{\partial \boldsymbol{\theta}}. \quad (24)$$

where

$$\mathbf{C}_i = \frac{\partial \text{SoftMax}(\mathbf{h}^L(\mathbf{z}_i))}{\partial \mathbf{h}^L(\mathbf{z}_i)} = \text{diag}(\mathbf{o}_i) - \mathbf{o}_i \mathbf{o}_i^\top, \quad \mathbf{o}_i = \text{SoftMax}(\mathbf{h}^L(\mathbf{z}_i)).$$

By (A1), at the MLE  $\hat{\boldsymbol{\theta}}$ ,

$$\forall i, \quad \text{SoftMax}(\mathbf{h}^L(\mathbf{z}_i)) = \text{OneHot}(y_i).$$

Therefore

$$\forall i, \quad -\frac{\partial^2 \log p(y_i | \mathbf{z}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = \left[ \frac{\partial \mathbf{h}^L}{\partial \boldsymbol{\theta}} \right]^\top \cdot \mathbf{C}_i \cdot \frac{\partial \mathbf{h}^L}{\partial \boldsymbol{\theta}}.$$

Taking the sample average on both sides, we get

$$\mathfrak{J}(\hat{\boldsymbol{\theta}}) = \hat{\mathcal{I}}(\hat{\boldsymbol{\theta}}).$$

□

## Appendix B Proof of Lemma 1

*Proof.* If  $(\boldsymbol{\theta}, \sum_j \alpha_j \partial \theta_j) \in \text{Rad}(\mathcal{TM})$ , Then

$$\left\langle \sum_j \alpha_j \partial \theta_j, \sum_j \alpha_j \partial \theta_j \right\rangle_{\mathcal{I}(\boldsymbol{\theta})} = 0.$$

In matrix form, it is simply  $\boldsymbol{\alpha}^\top \mathcal{I}(\boldsymbol{\theta}) \boldsymbol{\alpha} = 0$ . We have the analytical expression

$$\mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}_p \left[ \left( \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}} \right)^\top \mathbf{C}(\mathbf{z}) \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}} \right].$$

Therefore

$$E_p \left[ \left( \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}} \boldsymbol{\alpha} \right)^\top \mathbf{C}(\mathbf{z}) \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}} \boldsymbol{\alpha} \right] = 0.$$

By noting that  $\mathbf{C}(\mathbf{z}) \succeq 0$  is psd, we have almost surely that

$$\left( \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}} \boldsymbol{\alpha} \right)^\top \mathbf{C}(\mathbf{z}) \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}} \boldsymbol{\alpha} = 0.$$

Any eigenvector of  $\mathbf{C}(\mathbf{z})$  associated with the zero eigenvalues must be a multiple of  $\mathbf{1}$ . Indeed,

$$\mathbf{v}^\top \mathbf{C}(\mathbf{z}) \mathbf{v} = \mathbf{v}^\top (\text{diag}(\mathbf{o}(\mathbf{z})) - \mathbf{o}(\mathbf{z}) \mathbf{o}(\mathbf{z})^\top) \mathbf{v} = \sum_j o_j(\mathbf{z}) (v_j - \sum_j o_j(\mathbf{z}) v_j)^2 = 0 \Leftrightarrow \mathbf{v} \propto \mathbf{1},$$

where  $o_j(\mathbf{z}) > 0$  is the  $j$ 'th element of  $\mathbf{o}(\mathbf{z})$ . Hence, almost surely

$$\frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}} \boldsymbol{\alpha} = \lambda(\mathbf{z}) \mathbf{1}.$$

□

*Remark.*  $\boldsymbol{\alpha}$  is associated with a tangent vector in  $\text{Rad}(\mathcal{TM})$ , meaning a dynamic along the lightlike dimensions. The Jacobian  $\frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}}$  is the local linear approximation of the mapping  $\boldsymbol{\theta} \rightarrow \mathbf{h}^L(\mathbf{z})$ . By lemma 1, with probability 1 such a dynamic leads to uniform increments in the output units, meaning  $\mathbf{h}^L(\mathbf{z}) \rightarrow \mathbf{h}^L(\mathbf{z}) + \lambda(\mathbf{z}) \mathbf{1}$ ,  $\forall i$ , and therefore the output distribution  $\text{SoftMax}(\mathbf{h}^L(\mathbf{z}))$  is not affected. In summary, we have verified that the radical distribution does not affect the neural network mapping.

## Appendix C Proof of Proposition 2

*Proof.*

$$\begin{aligned}\hat{d}(\boldsymbol{\theta}) &= \text{rank} \left( \hat{\mathcal{I}}(\boldsymbol{\theta}) \right) = \text{rank} \left( \sum_{i=1}^N \left( \frac{\partial \mathbf{h}^L(\mathbf{z}_i)}{\partial \boldsymbol{\theta}} \right)^\top \mathbf{C}_i \frac{\partial \mathbf{h}^L(\mathbf{z}_i)}{\partial \boldsymbol{\theta}} \right) \\ &\leq \sum_{i=1}^N \text{rank} \left( \left( \frac{\partial \mathbf{h}^L(\mathbf{z}_i)}{\partial \boldsymbol{\theta}} \right)^\top \mathbf{C}_i \frac{\partial \mathbf{h}^L(\mathbf{z}_i)}{\partial \boldsymbol{\theta}} \right) \leq \sum_{i=1}^N \min \left\{ \text{rank} \left( \frac{\partial \mathbf{h}^L(\mathbf{z}_i)}{\partial \boldsymbol{\theta}} \right), \text{rank}(\mathbf{C}_i) \right\}.\end{aligned}$$

Note the matrix  $\frac{\partial \mathbf{h}^L(\mathbf{z}_i)}{\partial \boldsymbol{\theta}}$  has size  $m \times D$ , and  $\mathbf{C}_i$  has size  $m \times m$  and rank  $(m-1)$ . We also have  $\hat{d}(\boldsymbol{\theta}) = \text{rank} \left( \hat{\mathcal{I}}(\boldsymbol{\theta}) \right) \leq D = \dim(\boldsymbol{\theta})$ . Therefore

$$\hat{d}(\boldsymbol{\theta}) \leq \sum_{i=1}^N \min \left\{ \text{rank} \left( \frac{\partial \mathbf{h}^L(\mathbf{z}_i)}{\partial \boldsymbol{\theta}} \right), m-1 \right\}.$$

□

## Appendix D Proof of Proposition 3

*Proof.* We only prove the upper bound of  $d(\boldsymbol{\theta})$ . The upper bound of  $\hat{d}(\boldsymbol{\theta})$  can be proved similarly.

$$d(\boldsymbol{\theta}) = \text{rank}(\mathcal{I}(\boldsymbol{\theta})) = \text{rank} \left( \mathbb{E}_{p(\mathbf{z})} \left[ \left( \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}} \right)^\top \mathbf{C}(\mathbf{z}) \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}} \right] \right).$$

That means  $d(\boldsymbol{\theta})$  is the dimensionality of the image of  $\mathbb{E}_{p(\mathbf{z})} \left[ \left( \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}} \right)^\top \mathbf{C}(\mathbf{z}) \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}} \right]$ .  $\forall \boldsymbol{\theta}$ , we have

$$\begin{aligned}\mathbb{E}_{p(\mathbf{z})} \left[ \left( \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}} \right)^\top \mathbf{C}(\mathbf{z}) \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}} \right] \boldsymbol{\theta} &= \mathbb{E}_{p(\mathbf{z})} \left[ \left( \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}} \right)^\top \mathbf{C}(\mathbf{z}) \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}} \boldsymbol{\theta} \right] \\ &= \mathbb{E}_{p(\mathbf{z})} \left[ \left( \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}} \right)^\top \boldsymbol{\beta}(\mathbf{z}) \right],\end{aligned}$$

where  $\boldsymbol{\beta}(\mathbf{z}) = \mathbf{C}(\mathbf{z}) \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}} \boldsymbol{\theta}$  is an  $m$ -dimensional vector. Therefore

$$\mathbb{E}_{p(\mathbf{z})} \left[ \left( \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}} \right)^\top \mathbf{C}(\mathbf{z}) \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}} \right] \boldsymbol{\theta} \in \text{span} \bigcup_{\mathbf{z} \in \text{supp}(p)} \text{Row} \left( \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \boldsymbol{\theta}} \right).$$

Letting  $\boldsymbol{\theta}$  vary in  $\mathbb{R}^D$ , and applying  $\dim(\cdot)$  on both sides, the statement follows immediately. □

## Appendix E Proof of Proposition 4

*Proof.* We have the total derivative

$$d\mathbf{h}^L(\mathbf{z}) = \sum_{l=1}^L \mathbf{W}^L \boldsymbol{\Phi}^{L-1} \mathbf{W}^{L-1} \dots \boldsymbol{\Phi}^l (d\mathbf{W}^l \mathbf{z}^{l-1} + d\mathbf{b}^l).$$



Therefore,

$$\forall \mathbf{w} \in \mathbb{R}^{\dim(\mathbf{w}^l)}, \quad \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \mathbf{w}^l} \mathbf{w} = \mathbf{W}^L \mathbf{\Phi}^{L-1} \mathbf{W}^{L-1} \dots \mathbf{\Phi}^l \text{mat}(\mathbf{w}) \begin{pmatrix} \mathbf{z}^{l-1} \\ 1 \end{pmatrix},$$

where  $\text{mat}(\cdot)$  means to rearrange the vector into a matrix. Therefore,

$$\text{rank} \left( \frac{\partial \mathbf{h}^L(\mathbf{z})}{\partial \mathbf{w}^l} \right) \leq \text{rank} (\mathbf{W}^L \mathbf{\Phi}^{L-1} \mathbf{W}^{L-1} \dots \mathbf{\Phi}^l).$$

The second “ $\leq$ ” in the statement is because

$$\begin{aligned} \text{rank} (\mathbf{W}^L \mathbf{\Phi}^{L-1} \mathbf{W}^{L-1} \dots \mathbf{\Phi}^l) &\leq \min \{ \text{rank} (\mathbf{W}^L), \text{rank} (\mathbf{\Phi}^{L-1}), \dots, \text{rank} (\mathbf{\Phi}^l) \} \\ &\leq \min \{ \text{rank} (\mathbf{\Phi}^{L-1}), \dots, \text{rank} (\mathbf{\Phi}^l) \} \\ &= \min_{s=l}^{L-1} \text{rank} (\mathbf{\Phi}^s). \end{aligned}$$

□

## Appendix F Metric Signature of the Neuromanifold

The metric signature of  $\mathcal{M}$

$$(d(\boldsymbol{\theta}), 0, D - d(\boldsymbol{\theta}))$$

is straightforward from the fact that  $\mathcal{I}(\boldsymbol{\theta})$  is positive semi-definite (there is no negative eigenvalues), and the local dimensionality  $d(\boldsymbol{\theta})$ , by definition, is  $\text{rank}(\mathcal{I}(\boldsymbol{\theta}))$  (the number of non-zero eigenvalues).

We also show that  $\text{rank}(\mathfrak{J}(\boldsymbol{\theta})) \neq \hat{d}(\boldsymbol{\theta})$ . Recall that  $\hat{d}(\boldsymbol{\theta}) = \text{rank}(\hat{\mathcal{I}}(\boldsymbol{\theta}))$ , and

$$\text{rank}(\mathfrak{J}(\boldsymbol{\theta})) = \text{rank} \left( \frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right) = \text{rank} \left( \sum_i \frac{\partial^2 \ell_i}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right),$$

where  $\ell$  is the log-likelihood, and  $\ell_i = \log p(y_i | \mathbf{z}_i, \boldsymbol{\theta})$ . We write the analytical form of the elementwise Hessian

$$\frac{\partial^2 \ell_i}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = \sum_{j=1}^m \frac{\partial h_j^L(\mathbf{z}_i)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} (\text{OneHot}_j(y) - \text{SoftMax}_j(\mathbf{h}^L)) - \mathcal{I}(\boldsymbol{\theta}),$$

where  $\text{OneHot}(\cdot)$  denote the one-hot vector associated with the given target label  $y$ . Therefore

$$\boldsymbol{\alpha}^\top \frac{\partial^2 \ell_i}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \boldsymbol{\alpha} = \sum_{j=1}^m \boldsymbol{\alpha}^\top \left( \frac{\partial h_j^L(\mathbf{z}_i)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \boldsymbol{\alpha} \right) (\text{OneHot}_j(y) - \text{SoftMax}_j(\mathbf{h}^L)) - \boldsymbol{\alpha}^\top \mathcal{I}(\boldsymbol{\theta}) \boldsymbol{\alpha}.$$

Because of the first term on the RHS, the kernels of the two matrices  $\mathfrak{J}(\boldsymbol{\theta})$  and  $\hat{\mathcal{I}}(\boldsymbol{\theta})$  are different, and thus their ranks are also different.

## Appendix G Proof of Proposition 5

*Proof.* As  $\hat{\boldsymbol{\theta}}$  is the MLE, we have  $\mathfrak{J}(\hat{\boldsymbol{\theta}}) \succeq 0$ , and  $\forall \boldsymbol{\theta} \in \mathcal{M}$ ,

$$-\frac{N}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathfrak{J}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \leq 0.$$

Hence,

$$\mathbb{E}_p \exp \left( -\frac{N}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathfrak{J}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right) \leq 1.$$

Hence,

$$-\log \mathbb{E}_p \exp \left( -\frac{N}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathfrak{J}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right) \geq 0.$$

This proves the first “ $\leq$ ”.

As  $-\log(x)$  is convex, by Jensen’s inequality, we get

$$\begin{aligned} & -\log \mathbb{E}_p \exp \left( -\frac{N}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathfrak{J}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right) \\ & \leq \mathbb{E}_p \left( -\log \exp \left( -\frac{N}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathfrak{J}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right) \right) \\ & = \mathbb{E}_p \left( \frac{N}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathfrak{J}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right) \\ & = \frac{N}{2} \text{tr} \left( \mathbb{E}_p \left( \mathfrak{J}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \right) \right) \\ & = \frac{N}{2} \text{tr} \left( \mathfrak{J}(\hat{\boldsymbol{\theta}}) \left( (\mu(\boldsymbol{\theta}) - \hat{\boldsymbol{\theta}})(\mu(\boldsymbol{\theta}) - \hat{\boldsymbol{\theta}})^\top + \text{cov}(\boldsymbol{\theta}) \right) \right). \end{aligned}$$

This proves the second “ $\leq$ ”. □

## Appendix H Proof of Lemma 6

*Proof.* Due to the convexity of  $-\log t$ , we have

$$\begin{aligned} \overline{\{M_f(\mathbf{t}_{:,1}), \dots, M_f(\mathbf{t}_{:,m})\}} &= \frac{1}{m} \sum_{j=1}^m \left[ -\log \left( \frac{1}{n} \sum_{i=1}^n \exp(-t_{ij}) \right) \right] \\ &\geq -\log \left[ \frac{1}{m} \sum_{j=1}^m \frac{1}{n} \sum_{i=1}^n \exp(-t_{ij}) \right] = M_f(\mathbf{T}). \end{aligned}$$

This proves the first “ $\leq$ ”. To prove the second “ $\leq$ ”, we note that  $-\log \frac{1}{n} \sum_{i=1}^n \exp(-t_i)$  is a concave function. Therefore

$$\begin{aligned} \overline{\{M_f(\mathbf{t}_{:,1}), \dots, M_f(\mathbf{t}_{:,m})\}} &= \frac{1}{m} \sum_{j=1}^m \left[ -\log \left( \frac{1}{n} \sum_{i=1}^n \exp(-t_{ij}) \right) \right] \\ &\leq -\log \left( \frac{1}{n} \sum_{i=1}^n \exp \left( -\frac{1}{m} \sum_{j=1}^m t_{ij} \right) \right) = M_f(\{\bar{\mathbf{t}}_1, \dots, \bar{\mathbf{t}}_n\}). \end{aligned}$$

The last “ $\leq$ ” is based on the convexity of  $-\log t$ . Once again, by Jensen’s inequality, we have

$$M_f(\{\overline{t_1}, \dots, \overline{t_n}\}) \leq \frac{1}{n} \sum_{i=1}^n -\log \left( \exp \left( -\frac{1}{m} \sum_{j=1}^m t_{ij} \right) \right) = \overline{T}.$$

□

## Appendix I Derivations of $\mathcal{O}_G$

We recall the general formulation in eq. (15):

$$\begin{aligned} \mathcal{O} &:= -\log p(\mathbf{X} | \hat{\boldsymbol{\theta}}) + \log \int_{\mathcal{M}} \kappa(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &\quad - \log \int_{\mathcal{M}} \kappa(\boldsymbol{\theta}) \exp \left( -\frac{N}{2} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathfrak{J}(\hat{\boldsymbol{\theta}}) (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right) d\boldsymbol{\theta}. \end{aligned}$$

If  $\kappa(\boldsymbol{\theta}) = \exp \left( -\frac{1}{2} \boldsymbol{\theta}^\top \text{diag} \left( \frac{1}{\boldsymbol{\sigma}} \right) \boldsymbol{\theta} \right)$ , then the second term on the RHS is

$$\begin{aligned} \log \int_{\mathcal{M}} \kappa(\boldsymbol{\theta}) d\boldsymbol{\theta} &= \log \int_{\mathcal{M}} \exp \left( -\frac{1}{2} \boldsymbol{\theta}^\top \text{diag} \left( \frac{1}{\boldsymbol{\sigma}} \right) \boldsymbol{\theta} \right) d\boldsymbol{\theta} \\ &= \frac{D}{2} \log 2\pi + \frac{1}{2} \log |\text{diag}(\boldsymbol{\sigma})| \\ &\quad + \log \int_{\mathcal{M}} \exp \left( -\frac{D}{2} \log 2\pi - \frac{1}{2} \log |\text{diag}(\boldsymbol{\sigma})| - \frac{1}{2} \boldsymbol{\theta}^\top \text{diag} \left( \frac{1}{\boldsymbol{\sigma}} \right) \boldsymbol{\theta} \right) d\boldsymbol{\theta} \\ &= \frac{D}{2} \log 2\pi + \frac{1}{2} \log |\text{diag}(\boldsymbol{\sigma})| + \log 1 = \frac{D}{2} \log 2\pi + \frac{1}{2} \log |\text{diag}(\boldsymbol{\sigma})|. \end{aligned}$$

The third (last) term on the RHS is

$$\begin{aligned} & -\log \int_{\mathcal{M}} \kappa(\boldsymbol{\theta}) \exp \left( -\frac{N}{2} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathfrak{J}(\hat{\boldsymbol{\theta}}) (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right) d\boldsymbol{\theta} \\ &= -\log \int_{\mathcal{M}} \exp \left( -\frac{1}{2} \boldsymbol{\theta}^\top \text{diag} \left( \frac{1}{\boldsymbol{\sigma}} \right) \boldsymbol{\theta} - \frac{N}{2} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathfrak{J}(\hat{\boldsymbol{\theta}}) (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right) d\boldsymbol{\theta} \\ &= -\log \int_{\mathcal{M}} \exp \left( -\frac{1}{2} \boldsymbol{\theta}^\top \mathbf{A} \boldsymbol{\theta} + \mathbf{b}^\top \boldsymbol{\theta} + c \right) d\boldsymbol{\theta}, \end{aligned}$$

where

$$\mathbf{A} = N \mathfrak{J}(\hat{\boldsymbol{\theta}}) + \text{diag} \left( \frac{1}{\boldsymbol{\sigma}} \right) \succ 0, \quad \mathbf{b} = N \mathfrak{J}(\hat{\boldsymbol{\theta}}) \hat{\boldsymbol{\theta}}, \quad c = -\frac{N}{2} \hat{\boldsymbol{\theta}}^\top \mathfrak{J}(\hat{\boldsymbol{\theta}}) \hat{\boldsymbol{\theta}}.$$

Then,

$$\begin{aligned}
& -\log \int_{\mathcal{M}} \kappa(\boldsymbol{\theta}) \exp \left( -\frac{N}{2} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathfrak{J}(\hat{\boldsymbol{\theta}}) (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right) d\boldsymbol{\theta} \\
&= -\log \int_{\mathcal{M}} \exp \left( -\frac{1}{2} (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})^\top \mathbf{A} (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}) + c + \frac{1}{2} \bar{\boldsymbol{\theta}}^\top \mathbf{A} \bar{\boldsymbol{\theta}} \right) d\boldsymbol{\theta} \\
&= -\frac{D}{2} \log 2\pi + \frac{1}{2} \log |\mathbf{A}| - c - \frac{1}{2} \bar{\boldsymbol{\theta}}^\top \mathbf{A} \bar{\boldsymbol{\theta}} \\
&\quad -\log \int_{\mathcal{M}} \exp \left( -\frac{D}{2} \log 2\pi + \frac{1}{2} \log |\mathbf{A}| - \frac{1}{2} (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})^\top \mathbf{A} (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}) \right) d\boldsymbol{\theta} \\
&= -\frac{D}{2} \log 2\pi + \frac{1}{2} \log |\mathbf{A}| - c - \frac{1}{2} \bar{\boldsymbol{\theta}}^\top \mathbf{A} \bar{\boldsymbol{\theta}},
\end{aligned}$$

where  $\mathbf{A}\bar{\boldsymbol{\theta}} = \mathbf{b}$ . To sum up,

$$\begin{aligned}
\mathcal{O}_G &= -\log p(\mathbf{X} | \hat{\boldsymbol{\theta}}) + \frac{D}{2} \log 2\pi + \frac{1}{2} \log |\text{diag}(\boldsymbol{\sigma})| \\
&\quad -\frac{D}{2} \log 2\pi + \frac{1}{2} \log |\mathbf{A}| - c - \frac{1}{2} \bar{\boldsymbol{\theta}}^\top \mathbf{A} \bar{\boldsymbol{\theta}} \\
&= -\log p(\mathbf{X} | \hat{\boldsymbol{\theta}}) + \frac{1}{2} \log |\text{diag}(\boldsymbol{\sigma})| + \frac{1}{2} \log |\mathbf{A}| - c - \frac{1}{2} \bar{\boldsymbol{\theta}}^\top \mathbf{A} \bar{\boldsymbol{\theta}}, \\
&= -\log p(\mathbf{X} | \hat{\boldsymbol{\theta}}) + \frac{1}{2} \log |\text{diag}(\boldsymbol{\sigma})| + \frac{1}{2} \log |N\mathfrak{J}(\hat{\boldsymbol{\theta}}) + \text{diag}\left(\frac{1}{\boldsymbol{\sigma}}\right)| \\
&\quad + \frac{N}{2} \hat{\boldsymbol{\theta}}^\top \mathfrak{J}(\hat{\boldsymbol{\theta}}) \hat{\boldsymbol{\theta}} - \frac{1}{2} \left( N\mathfrak{J}(\hat{\boldsymbol{\theta}}) \hat{\boldsymbol{\theta}} \right)^\top \left( N\mathfrak{J}(\hat{\boldsymbol{\theta}}) + \text{diag}\left(\frac{1}{\boldsymbol{\sigma}}\right) \right)^{-1} N\mathfrak{J}(\hat{\boldsymbol{\theta}}) \hat{\boldsymbol{\theta}} \\
&= -\log p(\mathbf{X} | \hat{\boldsymbol{\theta}}) + \frac{1}{2} \log |N\mathfrak{J}(\hat{\boldsymbol{\theta}}) \text{diag}(\boldsymbol{\sigma}) + \mathbf{I}| \\
&\quad + \frac{1}{2} \hat{\boldsymbol{\theta}}^\top \left( \mathfrak{J}(\hat{\boldsymbol{\theta}}) \right)^\top \left( \mathfrak{J}(\hat{\boldsymbol{\theta}}) + \frac{1}{N} \text{diag}\left(\frac{1}{\boldsymbol{\sigma}}\right) \right)^{-1} \text{diag}\left(\frac{1}{\boldsymbol{\sigma}}\right) \hat{\boldsymbol{\theta}} \\
&= -\log p(\mathbf{X} | \hat{\boldsymbol{\theta}}) + \frac{1}{2} \log |N\mathfrak{J}(\hat{\boldsymbol{\theta}}) \text{diag}(\boldsymbol{\sigma}) + \mathbf{I}| + \frac{1}{2} \hat{\boldsymbol{\theta}}^\top \mathfrak{J}(\hat{\boldsymbol{\theta}}) \left( \text{diag}(\boldsymbol{\sigma}) \mathfrak{J}(\hat{\boldsymbol{\theta}}) + \frac{1}{N} \mathbf{I} \right)^{-1} \hat{\boldsymbol{\theta}}.
\end{aligned}$$

The last term does not scale with  $N$  and has a smaller order as compared to other terms. Indeed,

$$\lim_{N \rightarrow \infty} \mathfrak{J}(\hat{\boldsymbol{\theta}}) \left( \mathfrak{J}(\hat{\boldsymbol{\theta}}) + \frac{1}{N} \text{diag}\left(\frac{1}{\boldsymbol{\sigma}}\right) \right)^{-1} = \mathfrak{J}(\hat{\boldsymbol{\theta}}) \mathfrak{J}(\hat{\boldsymbol{\theta}})^+,$$

where  $\mathfrak{J}(\hat{\boldsymbol{\theta}})^+$  is the Moore-Penrose inverse of  $\mathfrak{J}(\hat{\boldsymbol{\theta}})$ . Hence, as  $N \rightarrow \infty$ ,

$$\begin{aligned}
\frac{1}{2} \hat{\boldsymbol{\theta}}^\top \mathfrak{J}(\hat{\boldsymbol{\theta}}) \left( \text{diag}(\boldsymbol{\sigma}) \mathfrak{J}(\hat{\boldsymbol{\theta}}) + \frac{1}{N} \mathbf{I} \right)^{-1} \hat{\boldsymbol{\theta}} &\rightarrow \frac{1}{2} \hat{\boldsymbol{\theta}}^\top \mathfrak{J}(\hat{\boldsymbol{\theta}}) \mathfrak{J}(\hat{\boldsymbol{\theta}})^+ \text{diag}\left(\frac{1}{\boldsymbol{\sigma}}\right) \hat{\boldsymbol{\theta}} \\
&\leq \frac{1}{2} \hat{\boldsymbol{\theta}}^\top \text{diag}\left(\frac{1}{\boldsymbol{\sigma}}\right) \hat{\boldsymbol{\theta}}.
\end{aligned}$$

By assumption (A5), the RHS is  $O(1)$ . This term is therefore dropped. We get

$$\mathcal{O}_G = -\log p(\mathbf{X} | \hat{\boldsymbol{\theta}}) + \frac{1}{2} \log |N\mathfrak{J}(\hat{\boldsymbol{\theta}}) \text{diag}(\boldsymbol{\sigma}) + \mathbf{I}| + O(1).$$

Note that  $\text{rank}(\mathfrak{J}(\hat{\boldsymbol{\theta}})) \leq D$ , and the matrix  $\mathfrak{J}(\hat{\boldsymbol{\theta}})\text{diag}(\boldsymbol{\sigma})$  has the same rank as  $\mathfrak{J}(\hat{\boldsymbol{\theta}})$ . We can write  $\mathfrak{J}(\hat{\boldsymbol{\theta}}) = \mathbf{L}(\hat{\boldsymbol{\theta}})\mathbf{L}(\hat{\boldsymbol{\theta}})^\top$ , where  $\mathbf{L}(\hat{\boldsymbol{\theta}})$  has shape  $D \times \text{rank}(\mathfrak{J}(\hat{\boldsymbol{\theta}}))$ . We abuse  $\mathbf{I}$  to denote both the identity matrix of shape  $D \times D$  and the identity matrix of shape  $\text{rank}(\mathfrak{J}(\hat{\boldsymbol{\theta}})) \times \text{rank}(\mathfrak{J}(\hat{\boldsymbol{\theta}}))$ . By the Weinstein–Aronszajn identity,

$$\begin{aligned}\mathcal{O}_G &= -\log p(\mathbf{X} | \hat{\boldsymbol{\theta}}) + \frac{1}{2} \log \left| N \mathbf{L}(\hat{\boldsymbol{\theta}})\mathbf{L}(\hat{\boldsymbol{\theta}})^\top \text{diag}(\boldsymbol{\sigma}) + \mathbf{I} \right| + O(1) \\ &= -\log p(\mathbf{X} | \hat{\boldsymbol{\theta}}) + \frac{1}{2} \log \left| N \mathbf{L}(\hat{\boldsymbol{\theta}})^\top \text{diag}(\boldsymbol{\sigma}) \mathbf{L}(\hat{\boldsymbol{\theta}}) + \mathbf{I} \right| + O(1) \\ &= -\log p(\mathbf{X} | \hat{\boldsymbol{\theta}}) + \frac{\text{rank}(\mathfrak{J}(\hat{\boldsymbol{\theta}}))}{2} \log N + \frac{1}{2} \log \left| \mathbf{L}(\hat{\boldsymbol{\theta}})^\top \text{diag}(\boldsymbol{\sigma}) \mathbf{L}(\hat{\boldsymbol{\theta}}) + \frac{1}{N} \mathbf{I} \right| + O(1).\end{aligned}$$

Note  $\mathbf{L}(\hat{\boldsymbol{\theta}})^\top \text{diag}(\boldsymbol{\sigma}) \mathbf{L}(\hat{\boldsymbol{\theta}})$  has the same set of non-zero eigenvalues as  $\mathbf{L}(\hat{\boldsymbol{\theta}})\mathbf{L}(\hat{\boldsymbol{\theta}})^\top \text{diag}(\boldsymbol{\sigma}) = \mathfrak{J}(\hat{\boldsymbol{\theta}})\text{diag}(\boldsymbol{\sigma})$ , which we denote as  $\lambda_i^+(\mathfrak{J}(\hat{\boldsymbol{\theta}})\text{diag}(\boldsymbol{\sigma}))$ . Then,

$$\begin{aligned}\mathcal{O}_G &= -\log p(\mathbf{X} | \hat{\boldsymbol{\theta}}) + \frac{\text{rank}(\mathfrak{J}(\hat{\boldsymbol{\theta}}))}{2} \log N \\ &\quad + \frac{1}{2} \sum_{i=1}^{\text{rank}(\mathfrak{J}(\hat{\boldsymbol{\theta}}))} \log \left( \lambda_i^+(\mathfrak{J}(\hat{\boldsymbol{\theta}})\text{diag}(\boldsymbol{\sigma})) + \frac{1}{N} \right) + O(1).\end{aligned}$$

Denote the largest and smallest elements of  $\boldsymbol{\sigma}$  as  $\sigma_{\max}$  and  $\sigma_{\min}$ , respectively. Then,

$$\mathbf{L}(\hat{\boldsymbol{\theta}})^\top \text{diag}(\boldsymbol{\sigma}) \mathbf{L}(\hat{\boldsymbol{\theta}}) \preceq \sigma_{\max} \mathbf{L}(\hat{\boldsymbol{\theta}})^\top \mathbf{L}(\hat{\boldsymbol{\theta}}).$$

Hence,

$$\begin{aligned}\frac{1}{2} \log \left| \mathbf{L}(\hat{\boldsymbol{\theta}})^\top \text{diag}(\boldsymbol{\sigma}) \mathbf{L}(\hat{\boldsymbol{\theta}}) + \frac{1}{N} \mathbf{I} \right| &\leq \frac{1}{2} \log \left| \sigma_{\max} \mathbf{L}(\hat{\boldsymbol{\theta}})^\top \mathbf{L}(\hat{\boldsymbol{\theta}}) + \frac{1}{N} \mathbf{I} \right| \\ &= \frac{1}{2} \sum_{i=1}^{\text{rank}(\mathfrak{J}(\hat{\boldsymbol{\theta}}))} \log \left( \sigma_{\max} \lambda_i^+(\mathfrak{J}(\hat{\boldsymbol{\theta}})) + \frac{1}{N} \right).\end{aligned}$$

Similarly,

$$\frac{1}{2} \log \left| \mathbf{L}(\hat{\boldsymbol{\theta}})^\top \text{diag}(\boldsymbol{\sigma}) \mathbf{L}(\hat{\boldsymbol{\theta}}) + \frac{1}{N} \mathbf{I} \right| \geq \frac{1}{2} \sum_{i=1}^{\text{rank}(\mathfrak{J}(\hat{\boldsymbol{\theta}}))} \log \left( \sigma_{\min} \lambda_i^+(\mathfrak{J}(\hat{\boldsymbol{\theta}})) + \frac{1}{N} \right).$$

If  $\boldsymbol{\sigma} = \sigma \mathbf{1}$ , then  $\sigma_{\max} = \sigma_{\min} = \sigma$ . Both “ $\leq$ ” and “ $\geq$ ” in the above inequalities become tight.

## Appendix J Probability Measures on $\mathcal{M}$

Probability measures are not defined on the lightlike  $\mathcal{M}$ , because along the lightlike geodesics, the distance is zero. To compute the integral of a given function  $f(\boldsymbol{\theta})$  on  $\mathcal{M}$  one has to first choose a proper Riemannian submanifold  $\mathcal{M}^s \subset \mathcal{M}$  specified by an embedding  $\boldsymbol{\theta}(\boldsymbol{\theta}^s)$ , whose metric is not singular. Then, the integral on  $\mathcal{M}^s$  can be defined as  $\int_{\mathcal{M}^s} f(\boldsymbol{\theta}(\boldsymbol{\theta}^s)) d\boldsymbol{\theta}^s$ , where  $\mathcal{M}^s$

is the sub-manifold associated with the frame  $\boldsymbol{\theta}^s = (\theta^1, \dots, \theta^d)$ , so that  $\mathcal{TM}^s = \mathcal{S}(\mathcal{TM})$ , and the induced Riemannian volume element as

$$\begin{aligned} d\boldsymbol{\theta}^s &= \sqrt{|\mathcal{I}(\boldsymbol{\theta}^s)|} d\theta^1 \wedge d\theta^2 \wedge \dots \wedge d\theta^d \\ &= \sqrt{|\mathcal{I}(\boldsymbol{\theta}^s)|} d_E \boldsymbol{\theta}^s, \end{aligned} \quad (25)$$

where  $d_E \boldsymbol{\theta}$  is the Euclidean volume element. We artificially shift  $\boldsymbol{\theta}$  to be positive definite and define the volume element as

$$\begin{aligned} d\boldsymbol{\theta} &:= \sqrt{|\mathcal{I}(\boldsymbol{\theta}) + \varepsilon_1 \mathbf{I}|} d\theta^1 \wedge d\theta^2 \wedge \dots \wedge d\theta^D \\ &= \sqrt{|\mathcal{I}(\boldsymbol{\theta}) + \varepsilon_1 \mathbf{I}|} d_E \boldsymbol{\theta}, \end{aligned} \quad (26)$$

where  $\varepsilon_1 > 0$  is a very small value as compared to the scale of  $\mathcal{I}(\boldsymbol{\theta})$  given by  $\frac{1}{D} \text{tr}(\mathcal{I}(\boldsymbol{\theta}))$ , *i.e.* the average of its eigenvalues. Notice this element will vary with  $\boldsymbol{\theta}$ : different coordinate systems will yield different volumes. Therefore it depends on how  $\boldsymbol{\theta}$  can be uniquely specified. This is roughly guaranteed by our assumption that the  $\boldsymbol{\theta}$ -coordinates correspond to the input coordinates (weights and biases) up to an orthogonal transformation. Despite that eq. (26) is a loose mathematical definition, it makes intuitive sense and is convenient for making derivations. Then, we can integrate functions

$$\int_{\mathcal{M}} f(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int f(\boldsymbol{\theta}) \sqrt{|\mathcal{I}(\boldsymbol{\theta}) + \varepsilon_1 \mathbf{I}|} d_E \boldsymbol{\theta}, \quad (27)$$

where the RHS is an integration over  $\mathbb{R}^D$ , assuming  $\boldsymbol{\theta}$  is real-valued.

Using this tool, we first consider Jeffreys' non-informative prior on a sub-manifold  $\mathcal{M}^s$ , given by

$$p_J(\boldsymbol{\theta}^s) = \frac{\sqrt{|\mathcal{I}(\boldsymbol{\theta}^s)|}}{\int_{\mathcal{M}^s} \sqrt{|\mathcal{I}(\boldsymbol{\theta}^s)|} d_E \boldsymbol{\theta}^s}. \quad (28)$$

It is easy to check  $\int_{\mathcal{M}^s} p_J(\boldsymbol{\theta}^s) d_E \boldsymbol{\theta}^s = 1$ . This prior may lead to similar results as [7, 67], *i.e.* a “razor” of the model  $\mathcal{M}^s$ . However, we will instead use a Gaussian-like prior, because Jeffreys' prior is not well defined on  $\mathcal{M}$ . Moreover, the integral  $\int_{\mathcal{M}^s} \sqrt{|\mathcal{I}(\boldsymbol{\theta}^s)|} d_E \boldsymbol{\theta}^s$  is likely to diverge based on our revised volume element in eq. (26). If the parameter space is real-valued, one can easily check that, the volume based on eq. (26) along the lightlike dimensions will diverge. The zero-centered Gaussian prior corresponds to a better *code*, because it is commonly acknowledged that one can achieve the same training error and generalization without using large weights. For example, regularizing the norm of the weights is widely used in deep learning. By using such an informative prior, one can have the same training error in the first term in eq. (2), while having a smaller “complexity” in the rest of the terms, because we only encode such models with constrained weights. Given the DNN, we define an *informative prior* on the lightlike neuromanifold

$$p(\boldsymbol{\theta}) = \frac{1}{V} \exp\left(-\frac{1}{2\varepsilon_2^2} \|\boldsymbol{\theta}\|^2\right) \sqrt{|\mathcal{I}(\boldsymbol{\theta}) + \varepsilon_1 \mathbf{I}|}, \quad (29)$$

where  $\varepsilon_2 > 0$  is a scale parameter of  $\boldsymbol{\theta}$ , and  $V$  is a normalizing constant to ensure  $\int p(\boldsymbol{\theta}) d_E \boldsymbol{\theta} = 1$ . Here, the base measure is the Euclidean volume element  $d_E \boldsymbol{\theta}$ , as  $\sqrt{|\mathcal{I}(\boldsymbol{\theta}) + \varepsilon_1 \mathbf{I}|}$  already appeared in  $p(\boldsymbol{\theta})$ . Keep in mind, again, that this  $p(\boldsymbol{\theta})$  is defined in a special coordinate system, and is not invariant to re-parametrization. This distribution is also isotropic in the input coordinate system, which agrees with initialization techniques<sup>7</sup>.

<sup>7</sup>Different layers, or weights and biases, may use different variance in their initialization. This minor issue can be solved by a simple re-scaling re-parameterization.

This bi-parametric prior connects Jeffreys' prior (that is widely used in MDL) and a Gaussian prior (that is widely used in deep learning). If  $\varepsilon_2 \rightarrow \infty$ ,  $\varepsilon_1 \rightarrow 0$ , it coincides with Jeffreys' prior (if it is well defined and  $\mathcal{I}(\boldsymbol{\theta})$  has full rank); if  $\varepsilon_1$  is large, the metric  $(\mathcal{I}(\boldsymbol{\theta}) + \varepsilon_1 \mathbf{I})$  becomes spherical, and eq. (29) becomes a Gaussian prior. We refer the reader to [34, 74] for other extensions of Jeffreys' prior.

The normalizing constant of eq. (29) is an information volume measure of  $\mathcal{M}$ , given by

$$V := \int_{\mathcal{M}} \exp\left(-\frac{1}{2\varepsilon_2^2} \|\boldsymbol{\theta}\|^2\right) d\boldsymbol{\theta}. \quad (30)$$

Unlike Jeffreys' prior whose information volume (the 3rd term on the RHS of eq. (2)) can be unbounded, this volume can be bounded as stated in the following theorem.

**Theorem 7.**

$$(\sqrt{2\pi\varepsilon_1\varepsilon_2})^D \leq V \leq (\sqrt{2\pi(\varepsilon_1 + \lambda_m)\varepsilon_2})^D, \quad (31)$$

where  $\lambda_m$  is the largest eigenvalue of the FIM  $\mathcal{I}(\boldsymbol{\theta})$ .

Notice  $\lambda_m$  may not exist, as the integration is taken over  $\boldsymbol{\theta} \in \mathcal{M}$ . Intuitively,  $V$  is a weighted volume w.r.t. a Gaussian-like prior distribution on  $\mathcal{M}$ , while the 3rd term on the RHS of eq. (2) is an unweighted volume. The larger the radius  $\varepsilon_2$ , the more “number” or possibilities of DNNs are included; the larger the parameter  $\varepsilon_1$ , the larger the local volume element in eq. (26) is measured, and therefore the total volume is measured larger.  $\log V$  is an  $O(D)$  terms, meaning the volume grows with the number of dimensions.

### J.1 Proof of Theorem 7

By definition,

$$V = \int_{\mathcal{M}} \exp\left(-\frac{1}{2\varepsilon_2^2} \|\boldsymbol{\theta}\|^2\right) d\boldsymbol{\theta} = \int \exp\left(-\frac{1}{2\varepsilon_2^2} \|\boldsymbol{\theta}\|^2\right) \sqrt{|\mathcal{I}(\boldsymbol{\theta}) + \varepsilon_1 \mathbf{I}|} d_{\mathbf{E}} \boldsymbol{\theta}.$$

By our assumption,  $\boldsymbol{\theta}$  is an orthogonal transformation of the neural network weights and biases, and therefore  $\boldsymbol{\theta} \in \mathbb{R}^D$ . We have

$$\sqrt{|\mathcal{I}(\boldsymbol{\theta}) + \varepsilon_1 \mathbf{I}|} \geq \sqrt{|\varepsilon_1 \mathbf{I}|} = \varepsilon_1^{\frac{D}{2}}.$$

Hence

$$\begin{aligned} V &\geq \int \exp\left(-\frac{1}{2\varepsilon_2^2} \|\boldsymbol{\theta}\|^2\right) \varepsilon_1^{\frac{D}{2}} d_{\mathbf{E}} \boldsymbol{\theta} \\ &= (2\pi)^{\frac{D}{2}} \varepsilon_2^D \varepsilon_1^{\frac{D}{2}} \int \exp\left(-\frac{D}{2} \log 2\pi - \frac{1}{2} \log |\varepsilon_2^2 \mathbf{I}| - \frac{1}{2\varepsilon_2^2} \|\boldsymbol{\theta}\|^2\right) d_{\mathbf{E}} \boldsymbol{\theta} \\ &= (2\pi)^{\frac{D}{2}} \varepsilon_2^D \varepsilon_1^{\frac{D}{2}} = (\sqrt{2\pi\varepsilon_1\varepsilon_2})^D. \end{aligned}$$

For the upper bound, we prove a stronger result as follows.

$$\sqrt{|\mathcal{I}(\boldsymbol{\theta}) + \varepsilon_1 \mathbf{I}|} = \left( \prod_{i=1}^D (\lambda_i + \varepsilon_1)^{\frac{1}{D}} \right)^{\frac{D}{2}} \leq \left( \frac{1}{D} \text{tr}(\mathcal{I}(\boldsymbol{\theta})) + \varepsilon_1 \right)^{\frac{D}{2}}.$$

Therefore

$$V \leq (\sqrt{2\pi\varepsilon_2})^D \left( \frac{1}{D} \text{tr}(\mathcal{I}(\boldsymbol{\theta})) + \varepsilon_1 \right)^{\frac{D}{2}}.$$



If one applies  $\frac{1}{D} \text{tr}(\mathcal{I}(\boldsymbol{\theta})) \leq \lambda_m$  to the RHS, the upper bound is further relaxed as

$$V \leq \left( \sqrt{2\pi\varepsilon_2} \right)^D (\lambda_m + \varepsilon_1)^{\frac{D}{2}} = \left( \sqrt{2\pi(\varepsilon_1 + \lambda_m)\varepsilon_2} \right)^D.$$

## Appendix K An Alternative Derivation of the Razor

In this section, we provide an alternative derivation of the propose razor  $\mathcal{O}$  based on a different prior. The main observations on the negative complexity are consistent with the cases of Gaussian and Jeffreys' priors.

We plug in the expression of  $p(\boldsymbol{\theta})$  in eq. (29) and get

$$\begin{aligned} -\log p(\mathbf{X}) &\approx -\log p(\mathbf{X} | \hat{\boldsymbol{\theta}}) + \log V \\ &\quad - \log \int_{\mathcal{M}} \left( -\frac{\|\boldsymbol{\theta}\|^2}{2\varepsilon_2^2} - \frac{N}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathfrak{J}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right) d\boldsymbol{\theta}. \end{aligned}$$

In the last term on the RHS, inside the parentheses is a quadratic function w.r.t.  $\boldsymbol{\theta}$ . However the integration is w.r.t. to the non-Euclidean volume element  $d\boldsymbol{\theta}$  and therefore does not have closed form. We need to assume

**(A7)**  $N$  is large enough so that  $|\mathcal{I}(\boldsymbol{\theta}) + \varepsilon_1 \mathbf{I}| \approx |\mathcal{I}(\hat{\boldsymbol{\theta}}) + \varepsilon_1 \mathbf{I}|$ .

This means the quadratic function will be sharp enough to make the volume element  $d\boldsymbol{\theta}$  to be roughly constant. Along the lightlike dimensions (zero eigenvalues of  $\mathcal{I}(\boldsymbol{\theta})$ ) this is trivial.

Plug eq. (29) into eq. (13), the following three terms

$$\frac{1}{V}, \quad \sqrt{|\mathcal{I}(\boldsymbol{\theta}) + \varepsilon_1 \mathbf{I}|} \approx \sqrt{|\mathcal{I}(\hat{\boldsymbol{\theta}}) + \varepsilon_1 \mathbf{I}|}, \quad \exp \left( \log p(\mathbf{X} | \hat{\boldsymbol{\theta}}) \right) = p(\mathbf{X} | \hat{\boldsymbol{\theta}})$$

can all be taken out of the integration as constant scalars, as they do not depend on  $\boldsymbol{\theta}$ . The main difficulty is to perform the integration

$$\begin{aligned} &\int \exp \left( -\frac{\|\boldsymbol{\theta}\|^2}{2\varepsilon_2^2} - \frac{N}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathfrak{J}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right) d_{\mathbf{E}}\boldsymbol{\theta} \\ &= \int \exp \left( -\frac{1}{2}\boldsymbol{\theta}^\top \mathbf{A}\boldsymbol{\theta} + \mathbf{b}^\top \boldsymbol{\theta} + c \right) d_{\mathbf{E}}\boldsymbol{\theta} \\ &= \int \exp \left( -\frac{1}{2}(\boldsymbol{\theta} - \mathbf{A}^{-1}\mathbf{b})^\top \mathbf{A}(\boldsymbol{\theta} - \mathbf{A}^{-1}\mathbf{b}) + \frac{1}{2}\mathbf{b}^\top \mathbf{A}^{-1}\mathbf{b} + c \right) d_{\mathbf{E}}\boldsymbol{\theta} \\ &= \exp \left( \frac{1}{2}\mathbf{b}^\top \mathbf{A}^{-1}\mathbf{b} + c \right) \int \exp \left( -\frac{1}{2}(\boldsymbol{\theta} - \mathbf{A}^{-1}\mathbf{b})^\top \mathbf{A}(\boldsymbol{\theta} - \mathbf{A}^{-1}\mathbf{b}) \right) d_{\mathbf{E}}\boldsymbol{\theta} \\ &= \exp \left( \frac{1}{2}\mathbf{b}^\top \mathbf{A}^{-1}\mathbf{b} + c \right) \exp \left( \frac{D}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{A}| \right) \\ &= \exp \left( \frac{1}{2}\mathbf{b}^\top \mathbf{A}^{-1}\mathbf{b} + c + \frac{D}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{A}| \right). \end{aligned}$$

where

$$\mathbf{A} = N\mathfrak{J}(\hat{\boldsymbol{\theta}}) + \frac{1}{\varepsilon_2^2}\mathbf{I}, \quad \mathbf{b} = N\mathfrak{J}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{\theta}}, \quad c = -\frac{1}{2}\hat{\boldsymbol{\theta}}^\top N\mathfrak{J}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{\theta}}.$$

The rest of the derivations are straightforward. Note  $R = -c - \frac{1}{2}\mathbf{b}^\top \mathbf{A}^{-1}\mathbf{b}$ .

After derivations and simplifications, we get

$$\begin{aligned} -\log p(\mathbf{X}) &\approx -\log p(\mathbf{X} | \hat{\boldsymbol{\theta}}) + \frac{D}{2} \log \frac{N}{2\pi} + \log V \\ &+ \frac{1}{2} \log \left| \mathfrak{J}(\hat{\boldsymbol{\theta}}) + \frac{1}{N\epsilon_2^2} \mathbf{I} \right| - \frac{1}{2} \log \left| \mathcal{I}(\hat{\boldsymbol{\theta}}) + \epsilon_1 \mathbf{I} \right| + R. \end{aligned} \quad (32)$$

The remainder term is given by

$$R = \frac{1}{2} \hat{\boldsymbol{\theta}}^\top \left[ N \mathfrak{J}(\hat{\boldsymbol{\theta}}) - N \mathfrak{J}(\hat{\boldsymbol{\theta}}) \left( N \mathfrak{J}(\hat{\boldsymbol{\theta}}) + \frac{1}{\epsilon_2^2} \mathbf{I} \right)^{-1} N \mathfrak{J}(\hat{\boldsymbol{\theta}}) \right] \hat{\boldsymbol{\theta}}. \quad (33)$$

We need to analyze the order of this  $R$  term. Assume the largest eigenvalue of  $\mathfrak{J}(\hat{\boldsymbol{\theta}})$  is  $\lambda_m$ , then

$$|R| \leq \frac{N \lambda_m}{\epsilon_2^2 N \lambda_m + 1} \|\hat{\boldsymbol{\theta}}\|^2. \quad (34)$$

We assume

(A8) The ratio between the scale of each dimension of the MLE  $\hat{\boldsymbol{\theta}}$  to  $\epsilon_2$ , *i.e.*  $\frac{\hat{\theta}_i}{\epsilon_2}$  ( $i = 1, \dots, D$ ) is in the order  $O(1)$ .

Intuitively, the scale parameter  $\epsilon_2$  in our prior  $p(\boldsymbol{\theta})$  in eq. (29) is chosen to “cover” the good models. Therefore, the order of  $R$  is  $O(D)$ . As  $N$  turns large,  $R$  will be dominated by the 2nd  $O(D \log N)$  term. We will therefore discard  $R$  for simplicity. It could be useful for a more delicate analysis. In conclusion, we arrive at the following expression

$$\mathcal{O} := -\log p(\mathbf{X} | \hat{\boldsymbol{\theta}}) + \frac{D}{2} \log \frac{N}{2\pi} + \log V + \frac{1}{2} \log \frac{\left| \mathfrak{J}(\hat{\boldsymbol{\theta}}) + \frac{1}{N\epsilon_2^2} \mathbf{I} \right|}{\left| \mathcal{I}(\hat{\boldsymbol{\theta}}) + \epsilon_1 \mathbf{I} \right|}. \quad (35)$$

Notice the similarity with eq. (2), where the first two terms on the RHS are exactly the same. The 3rd term is an  $O(D)$  term, similar to the 3rd term in eq. (2). It is bounded according to theorem 7, while the 3rd term in eq. (2) could be unbounded. Our last term is in a similar form to the last term in eq. (2), except it is well defined on lightlike manifold. If we let  $\epsilon_2 \rightarrow \infty$ ,  $\epsilon_1 \rightarrow 0$ , we get exactly eq. (2) and in this case  $\mathcal{O} = \chi$ . As the number of parameters  $D$  turns large, both the 2nd and 3rd terms will grow linearly w.r.t.  $D$ , meaning that they contribute positively to the model complexity. Interestingly, the fourth term is a “*negative complexity*”. Regard  $\frac{1}{N\epsilon_2^2}$  and  $\epsilon_1$  as small positive values. The fourth term essentially is a log-ratio from the observed FIM to the true FIM. For small models, they coincide, because the sample size  $N$  is large based on the model size. In this case, the effect of this term is minor. For DNNs, the sample size  $N$  is very limited based on the huge model size  $D$ . Along a dimension  $\theta_i$ ,  $\mathfrak{J}(\boldsymbol{\theta})$  is likely to be singular as stated in proposition 2, even if  $\mathcal{I}$  has a very small positive value. In this case, their log-ratio will be negative. Therefore, the razor  $\mathcal{O}$  favors DNNs with their Fisher-spectrum clustered around 0.

In fig. 2, model C displays the concepts of a DNN, where there are many good local optima. The performance is not sensitive to specific values of model parameters. On the lightlike neuromanifold  $\mathcal{M}$ , there are many directions that are very close to being lightlike. When a DNN model varies along these directions, the model slightly changes in terms of  $\mathcal{I}(\boldsymbol{\theta})$ , but their prediction on the samples measured by  $\mathfrak{J}(\boldsymbol{\theta})$  are invariant. These directions count *negatively* towards the complexity, because these extra freedoms (dimensions of  $\boldsymbol{\theta}$ ) occupy almost zero volume in the geometric sense, and are helpful to give a shorter code to future unseen samples.

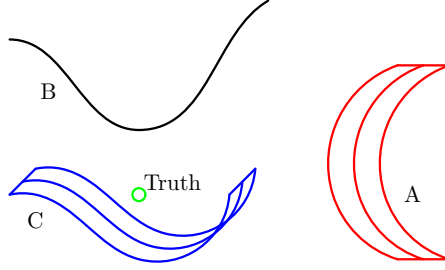


Figure 2: A: a model far from the truth (underlying distribution of observed data); B: close to the truth but sensitive to parameter; C (deep learning): close to the truth with many good local optima.

To obtain a simpler expression, we consider the case that  $\mathcal{I}(\boldsymbol{\theta}) \equiv \mathcal{I}(\hat{\boldsymbol{\theta}})$  is both constant and diagonal in the interested region defined by eq. (29). In this case,

$$\log V \approx \frac{D}{2} \log 2\pi + D \log \varepsilon_2 + \frac{1}{2} \log |\mathcal{I}(\hat{\boldsymbol{\theta}}) + \varepsilon_1 \mathbf{I}|. \quad (36)$$

On the other hand, as  $D \rightarrow \infty$ , the spectrum of the FIM  $\mathcal{I}(\boldsymbol{\theta})$  will follow the density  $\rho_{\mathcal{I}}(\boldsymbol{\theta})$ . We plug these expressions into eq. (35), discard all lower-order terms, and get a simplified version of the razor

$$\mathcal{O} \approx -\log p(\mathbf{X} | \hat{\boldsymbol{\theta}}) + \frac{D}{2} \log N + \frac{D}{2} \int_0^\infty \rho_{\mathcal{I}}(\lambda) \log \left( \lambda + \frac{1}{N\varepsilon_2^2} \right) d\lambda, \quad (37)$$

where  $\rho_{\mathcal{I}}$  denotes the spectral density of the Fisher information matrix.

## References

- [1] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.*, 19(6):716–723, 1974.
- [2] Guillaume Alain, Nicolas Le Roux, and Pierre-Antoine Manzagol. Negative eigenvalues of the Hessian in deep neural networks. In *ICLR’18 workshop*, 2018. arXiv:1902.02366 [cs.LG].
- [3] Shun-ichi Amari. *Information Geometry and Its Applications*, volume 194 of *Applied Mathematical Sciences*. Springer, Japan, 2016.
- [4] Shun-ichi Amari, Tomoko Ozeki, Ryo Karakida, Yuki Yoshida, and Masato Okada. Dynamics of learning in MLP: Natural gradient and singularity revisited. *Neural Computation*, 30(1):1–33, 2018.
- [5] Toshiki Aoki and Katsuhiko Kuribayashi. On the category of stratifolds. *Cahiers de Topologie et Géométrie Différentielle Catégoriques*, LVIII(2):131–160, 2017. arXiv:1605.04142 [math.CT].
- [6] Oguzhan Bahadir and Mukut Mani Tripathi. Geometry of lightlike hypersurfaces of a statistical manifold, 2019. arXiv:1901.09251 [math.DG].
- [7] Vijay Balasubramanian. MDL, Bayesian inference and the geometry of the space of probability distributions. In *Advances in Minimum Description Length: Theory and Applications*, pages 81–98. MIT Press, Cambridge, Massachusetts, 2005.

- [8] A. Barron, J. Rissanen, and Bin Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, 1998.
- [9] A.R. Barron and T.M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37(4):1034–1054, 1991.
- [10] Léonard Blier and Yann Ollivier. The description length of deep learning models. In *Advances in Neural Information Processing Systems 31*, pages 2216–2226. Curran Associates, Inc., NY 12571, USA, 2018.
- [11] A. Blum and J. Langford. PAC-MDL bounds. In *Proc. Sixteenth Conf. Learning Theory (COLT’ 03)*, pages 344–357, 2003.
- [12] Ovidiu Calin. *Deep learning architectures*. Springer, London, 2020.
- [13] Ovidiu Calin and Constantin Udrişte. *Geometric modeling in probability and statistics*. Springer, Cham, 2014.
- [14] Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. Model compression and acceleration for deep neural networks: The principles, progress, and challenges. *IEEE Signal Processing Magazine*, 35(1):126–136, 2018.
- [15] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1019–1028, 2017.
- [16] Krishan Duggal. A review on unique existence theorems in lightlike geometry. *Geometry*, 2014, 2014. Article ID 835394.
- [17] Krishan Duggal and Aurel Bejancu. *Lightlike Submanifolds of Semi-Riemannian Manifolds and Applications*, volume 364 of *Mathematics and Its Applications*. Springer, Netherlands, 1996.
- [18] Pascal Mattia Esser and Frank Nielsen. Towards modeling and resolving singular parameter spaces using stratifolds. *arXiv preprint arXiv:2112.03734*, 2021.
- [19] Xinlong Feng and Zhinan Zhang. The rank of a random matrix. *Applied Mathematics and Computation*, 185(1):689–694, 2007.
- [20] Adam Gaier and David Ha. Weight agnostic neural networks. In *Advances in Neural Information Processing Systems 32*, pages 5365–5379. Curran Associates, Inc., NY 12571, USA, 2019.
- [21] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 249–256, 2010.
- [22] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, 2011.
- [23] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, Cambridge, Massachusetts, 2016.

- [24] Peter Grünwald and Teemu Roos. Minimum description length revisited. *International Journal of Mathematics for Industry*, 11(01):1930001, 2019.
- [25] Peter D. Grünwald. *The Minimum Description Length Principle*. Adaptive Computation and Machine Learning series. The MIT Press, Cambridge, Massachusetts, 2007.
- [26] Peter D. Grünwald and Nishant A. Mehta. A tight excess risk bound via a unified PAC-Bayesian–Rademacher–Shtarkov–MDL complexity. In Aurélien Garivier and Satyen Kale, editors, *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 433–465, 2019.
- [27] Tomohiro Hayase and Ryo Karakida. The spectrum of Fisher information of deep networks achieving dynamical isometry. In *International Conference on Artificial Intelligence and Statistics*, pages 334–342, 2021.
- [28] Masahito Hayashi. Large deviation theory for non-regular location shift family. *Annals of the Institute of Statistical Mathematics*, 63(4):689–716, 2011.
- [29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.
- [30] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.
- [31] Harold Hotelling. Spaces of statistical parameters. *Bull. Amer. Math. Soc.*, 36:191, 1930.
- [32] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In *Advances in Neural Information Processing Systems 29*, pages 4107–4115. Curran Associates, Inc., NY 12571, USA, 2016.
- [33] Varun Jain, Amrinder Pal Singh, and Rakesh Kumar. On the geometry of lightlike submanifolds of indefinite statistical manifolds, 2019. arXiv:1903.07387 [math.DG].
- [34] Ruichao Jiang, Javad Tavakoli, and Yiqiang Zhao. Weyl prior and Bayesian statistics. *Entropy*, 22(4), 2020.
- [35] Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal statistics of Fisher information in deep neural networks: Mean field approach. In *International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1032–1041, 2019.
- [36] Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Pathological Spectra of the Fisher Information Metric and Its Variants in Deep Neural Networks. *Neural Computation*, 33(8):2274–2307, 2021.
- [37] David C Kay. *Schaum’s outline of theory and problems of tensor calculus*. McGraw-Hill, New York, 1988.
- [38] Andreï Nikolaevich Kolmogorov. *Sur la notion de la moyenne*. G. Bardi, tip. della R. Accad. dei Lincei, Rome, Italy, 1930.
- [39] Osamu Komori and Shinto Eguchi. A unified formulation of  $k$ -Means, fuzzy  $c$ -Means and Gaussian mixture model by the Kolmogorov–Nagumo average. *Entropy*, 23(5):518, 2021.

- [40] Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical Fisher approximation for natural gradient descent. In *Advances in Neural Information Processing Systems 32*, pages 4158–4169. Curran Associates, Inc., NY 12571, USA, 2019.
- [41] D.N. Kupeli. *Singular Semi-Riemannian Geometry*, volume 366 of *Mathematics and Its Applications*. Springer, Netherlands, 1996.
- [42] Stefan L Lauritzen. Statistical manifolds. *Differential geometry in statistical inference*, 10:163–216, 1987.
- [43] Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations (ICLR)*, 2018.
- [44] Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-Rao metric, geometry, and complexity of neural networks. In *International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 888–896, 2019.
- [45] Wu Lin, Valentin Duruisseaux, Melvin Leok, Frank Nielsen, Mohammad Emtiyaz Khan, and Mark Schmidt. Simplifying momentum-based positive-definite submanifold optimization with applications to deep learning. In *International Conference on Machine Learning*, pages 21026–21050. PMLR, 2023.
- [46] David J.C. MacKay. *Bayesian methods for adaptive models*. PhD thesis, California Institute of Technology, 1992.
- [47] James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.
- [48] James A. Mingo and Roland Speicher. *Free Probability and Random Matrices*, volume 35 of *Fields Institute Monographs*. Springer, New York, 2017.
- [49] Noboru Murata, Shuji Yoshizawa, and Shun-ichi Amari. Network information criterion-determining the number of hidden units for an artificial neural network model. *IEEE transactions on neural networks*, 5(6):865–872, 1994.
- [50] In Jae Myung, Vijay Balasubramanian, and Mark A. Pitt. Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences*, 97(21):11170–11175, 2000.
- [51] Mitio Nagumo. Über eine Klasse der Mittelwerte. In *Japanese journal of mathematics: transactions and abstracts*, volume 7, pages 71–79. The Mathematical Society of Japan, 1930.
- [52] Naomichi Nakajima and Toru Ohmoto. The dually flat structure for singular models. *Information Geometry*, 4(1):31–64, 2021.
- [53] Behnam Neyshabur, Srinadh Bhojanapalli, David Mcallester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems 30*, pages 5947–5956. Curran Associates, Inc., NY 12571, USA, 2017.
- [54] Katsumi Nomizu, Nomizu Katsumi, and Takeshi Sasaki. *Affine differential geometry: geometry of affine immersions*. Cambridge Tracts in Mathematics. Cambridge university press, Cambridge, United Kingdom, 1994.

- [55] A Emin Orhan and Xaq Pitkow. Skip connections eliminate singularities. In *International Conference on Learning Representations (ICLR)*, 2018.
- [56] Vardan Papayan. Traces of class/cross-class structure pervade deep learning spectra. *Journal of Machine Learning Research*, 21(252):1–64, 2020.
- [57] Razvan Pascanu and Yoshua Bengio. Revisiting natural gradient for deep networks. In *International Conference on Learning Representations (ICLR)*, 2014.
- [58] Jeffrey Pennington and Yasaman Bahri. Geometry of neural network loss surfaces via random matrix theory. In *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2798–2806, 2017.
- [59] Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli. The emergence of spectral universality in deep networks. In *International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1924–1932, 2018.
- [60] Jeffrey Pennington and Pratik Worah. The spectrum of the Fisher information matrix of a single-hidden-layer neural network. In *Advances in Neural Information Processing Systems 31*, pages 5410–5419. Curran Associates, Inc., NY 12571, USA, 2018.
- [61] David Pollard. A note on insufficiency and the preservation of Fisher information. In *From Probability to Statistics and Back: High-Dimensional Models and Processes—A Festschrift in Honor of Jon A. Wellner*, pages 266–275. Institute of Mathematical Statistics, Beachwood, Ohio, 2013.
- [62] Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. In *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2847–2854, 2017.
- [63] Calyampudi Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of Cal. Math. Soc.*, 37(3):81–91, 1945.
- [64] Calyampudi Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. In *Breakthroughs in statistics*, pages 235–247. Springer, New York, NY, 1992.
- [65] J. Rissanen. Strong optimality of the normalized ml models as universal codes and information in data. *IEEE Transactions on Information Theory*, 47(5):1712–1717, 2001.
- [66] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.
- [67] Jorma Rissanen. Fisher information and stochastic complexity. *IEEE Trans. Inf. Theory*, 42(1):40–47, 1996.
- [68] Levent Sagun, Utku Evci, V. Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the Hessian of over-parametrized neural networks. In *ICLR’18 workshop*, 2018. arXiv:1706.04454 [cs.LG].
- [69] Salem Said, Hatem Hajri, Lionel Bombrun, and Baba C Vemuri. Gaussian distributions on Riemannian symmetric spaces: statistical learning with structured covariance matrices. *IEEE Transactions on Information Theory*, 64(2):752–772, 2017.
- [70] Gideon Schwarz. Estimating the dimension of a model. *Ann. Stat.*, 6(2):461–464, 1978.



- [71] Y. M. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23(3):3–17, 1987.
- [72] Alexander Soen and Ke Sun. On the variance of the Fisher information for deep learning. In *Advances in Neural Information Processing Systems 34*, pages 5708–5719, NY 12571, USA, 2021. Curran Associates, Inc.
- [73] Ke Sun and Frank Nielsen. Relative Fisher information and natural gradient for learning large modular models. In *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3289–3298, 2017.
- [74] Junnichi Takeuchi and S-I Amari.  $\alpha$ -parallel prior and its properties. *IEEE Transactions on Information Theory*, 51(3):1011–1023, 2005.
- [75] Philip Thomas. Genga: A generalization of natural gradient ascent with positive and negative convergence results. In *International Conference on Machine Learning*, volume 32 (2) of *Proceedings of Machine Learning Research*, pages 1575–1583, 2014.
- [76] Guillermo Valle-Pérez, Chico Q. Camargo, and Ard A. Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. In *International Conference on Learning Representations (ICLR)*, 2019.
- [77] Christopher Stewart Wallace and D. M. Boulton. An information measure for classification. *Computer Journal*, 11(2):185–194, 1968.
- [78] Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory*, volume 25 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, United Kingdom, 2009.
- [79] Haikun Wei, Jun Zhang, Florent Cousseau, Tomoko Ozeki, and Shun-ichi Amari. Dynamics of learning near singularities in layered networks. *Neural computation*, 20(3):813–843, 2008.
- [80] Keisuke Yamazaki and Sumio Watanabe. Singularities in mixture models and upper bounds of stochastic complexity. *Neural networks*, 16(7):1029–1038, 2003.
- [81] Yuki Yoshida, Ryo Karakida, Masato Okada, and Shun-ichi Amari. Statistical mechanical analysis of learning dynamics of two-layer perceptron with multiple output units. *Journal of Physics A: Mathematical and Theoretical*, 2019.
- [82] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.
- [83] Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006.