

Fisher-Rao Metric, Geometry, and Complexity of Neural Networks

Tengyuan Liang*

University of Chicago

Tomaso Poggio†

Massachusetts Institute of Technology

Alexander Rakhlin‡

University of Pennsylvania

James Stokes§

University of Pennsylvania

Abstract.

We study the relationship between geometry and capacity measures for deep neural networks from an invariance viewpoint. We introduce a new notion of capacity—the Fisher-Rao norm—that possesses desirable invariance properties and is motivated by Information Geometry. We discover an analytical characterization of the new capacity measure, through which we establish norm-comparison inequalities and further show that the new measure serves as an umbrella for several existing norm-based complexity measures. We discuss upper bounds on the generalization error induced by the proposed measure. Extensive numerical experiments on CIFAR-10 support our theoretical findings. Our theoretical analysis rests on a key structural lemma about partial derivatives of multi-layer rectifier networks.

Key words and phrases: deep learning, statistical learning theory, information geometry, Fisher-Rao metric, invariance, ReLU activation, natural gradient, capacity control, generalization error.

1. INTRODUCTION

Beyond their remarkable representation and memorization ability, deep neural networks empirically perform well in out-of-sample prediction. This intriguing out-of-sample generalization property poses two fundamental theoretical questions:

- What are the complexity notions that control the generalization aspects of neural networks?

* (e-mail: tengyuan.liang@chicagobooth.edu)

† (e-mail: tp@ai.mit.edu)

‡ (e-mail: rakhlin@wharton.upenn.edu)

§ (e-mail: stokesj@sas.upenn.edu)

- Why does stochastic gradient descent, or other variants, find parameters with small complexity?

In this paper we approach the generalization question for deep neural networks from a geometric invariance vantage point. The motivation behind invariance is twofold: (1) The specific parametrization of the neural network is arbitrary and should not impact its generalization power. As pointed out in [Neyshabur et al., 2015a], for example, there are many continuous operations on the parameters of ReLU nets that will result in exactly the same prediction and thus generalization can only depend on the equivalence class obtained by identifying parameters under these transformations. (2) Although flatness of the loss function has been linked to generalization [Hochreiter and Schmidhuber, 1997], existing definitions of flatness are neither invariant to nodewise re-scalings of ReLU nets nor general coordinate transformations [Dinh et al., 2017] of the parameter space, which calls into question their utility for describing generalization.

It is thus natural to argue for a purely geometric characterization of generalization that is invariant under the aforementioned transformations and additionally resolves the conflict between flat minima and the requirement of invariance. Information geometry is concerned with the study of geometric invariances arising in the space of probability distributions, so we will leverage it to motivate a particular geometric notion of complexity — the Fisher-Rao norm. From an algorithmic point of view the steepest descent induced by this geometry is precisely the natural gradient [Amari, 1998]. From the generalization viewpoint, the Fisher-Rao norm naturally incorporates distributional aspects of the data and harmoniously unites elements of flatness and norm which have been argued to be crucial for explaining generalization [Neyshabur et al., 2017].

Statistical learning theory equips us with many tools to analyze out-of-sample performance. The Vapnik-Chervonenkis dimension is one possible complexity notion, yet it may be too large to explain generalization in over-parametrized models, since it scales with the size (dimension) of the network. In contrast, under additional distributional assumptions of a margin, Perceptron (a one-layer network) enjoys a dimension-free error guarantee, with an ℓ_2 norm playing the role of “capacity”. These observations (going back to the 60’s) have led the theory of large-margin classifiers, applied to kernel methods, boosting, and neural networks [Anthony and Bartlett, 1999]. In particular, the analysis of Koltchinskii and Panchenko [2002] combines the empirical margin distribution (quantifying how well the data can be separated) and the Rademacher complexity of a restricted subset of functions. This in turn raises the capacity control question: what is a good notion of the restrictive subset of parameter space for neural networks? Norm-based capacity control provides a possible answer and is being actively studied for deep networks [Krogh and Hertz, 1992, Neyshabur et al., 2015b,a, Bartlett et al., 2017, Neyshabur et al., 2017], yet the invariances are not always reflected in these capacity notions. In general, it is very difficult to answer the question of which capacity measure is superior. Nevertheless, we will show that our proposed Fisher-Rao norm serves as an umbrella for the previously considered norm-based capacity measures, and it appears to shed light on possible answers to the above question.

Much of the difficulty in analyzing neural networks stems from their unwieldy recursive definition interleaved with nonlinear maps. In analyzing the Fisher-Rao

norm, we proved an identity for the partial derivatives of the neural network that appears to open the door to some of the geometric analysis. In particular, we prove that any stationary point of the empirical objective with hinge loss that perfectly separates the data must also have a large margin. Such an automatic large-margin property of stationary points may link the algorithmic facet of the problem with the generalization property. The same identity gives us a handle on the Fisher-Rao norm and allows us to prove a number of facts about it. Since we expect that the identity may be useful in deep network analysis, we start by stating this result and its implications in the next section. In Section 3 we introduce the Fisher-Rao norm and establish through norm-comparison inequalities that it serves as an umbrella for existing norm-based measures of capacity. Using these norm-comparison inequalities we bound the generalization error of various geometrically distinct subsets of the Fisher-Rao ball and provide a rigorous proof of generalization for deep linear networks. Extensive numerical experiments are performed in Section 5 demonstrating the superior properties of the Fisher-Rao norm.

2. GEOMETRY OF DEEP RECTIFIED NETWORKS

DEFINITION 1. The function class \mathcal{H}_L realized by the *feedforward neural network architecture* of depth L with coordinate-wise activation functions $\sigma_l : \mathbb{R} \rightarrow \mathbb{R}$ is defined as set of functions $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ ($\mathcal{X} \subseteq \mathbb{R}^p$ and $\mathcal{Y} \subseteq \mathbb{R}^K$)¹ with

$$(2.1) \quad f_\theta(x) = \sigma_{L+1}(\sigma_L(\dots\sigma_2(\sigma_1(x^T W^0)W^1)W^2)\dots)W^L) ,$$

where the parameter vector $\theta \in \Theta_L \subseteq \mathbb{R}^d$ ($d = pk_1 + \sum_{i=1}^{L-1} k_i k_{i+1} + k_L K$) and

$$\Theta_L = \{W^0 \in \mathbb{R}^{p \times k_1}, W^1 \in \mathbb{R}^{k_1 \times k_2}, \dots, W^{L-1} \in \mathbb{R}^{k_{L-1} \times k_L}, W^L \in \mathbb{R}^{k_L \times K}\} .$$

For simplicity of calculations, we have set all bias terms to zero². We also assume throughout the paper that

$$(2.2) \quad \sigma(z) = \sigma'(z)z.$$

for all the activation functions, which includes ReLU $\sigma(z) = \max\{0, z\}$, “leaky” ReLU $\sigma(z) = \max\{\alpha z, z\}$, and linear activations as special cases.

To make the exposition of the structural results concise, we define the following intermediate functions in the definition (2.1). The output value of the t -th layer hidden node is denoted as $O^t(x) \in \mathbb{R}^{k_t}$, and the corresponding input value as $N^t(x) \in \mathbb{R}^{k_t}$, with $O^t(x) = \sigma_t(N^t(x))$. By definition, $O^0(x) = x \in \mathbb{R}^p$, and the final output $O^{L+1}(x) = f_\theta(x) \in \mathbb{R}^K$. For any N_i^t, O_i^t , the subscript i denotes the i -th coordinate of the vector.

Given a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$, the statistical learning problem can be phrased as optimizing the unobserved population loss:

$$(2.3) \quad L(\theta) := \mathbb{E}_{(X,Y) \sim \mathcal{P}} \ell(f_\theta(X), Y) ,$$

¹It is possible to generalize the above architecture to include linear pre-processing operations such as zero-padding and average pooling.

²In practice, we found that setting the bias to zero does not significantly impact results on image classification tasks such as MNIST and CIFAR-10.

based on i.i.d. samples $\{(X_i, Y_i)\}_{i=1}^N$ from the unknown joint distribution \mathcal{P} . The unregularized empirical objective function is denoted by

$$(2.4) \quad \hat{L}(\theta) := \widehat{\mathbb{E}}\ell(f_\theta(X), Y) = \frac{1}{N} \sum_{i=1}^N \ell(f_\theta(X_i), Y_i) .$$

We first establish the following structural result for neural networks. It will be clear in the later sections that the lemma is motivated by the study of the Fisher-Rao norm, formally defined in Eqn. (3.1) below, and information geometry. For the moment, however, let us provide a different viewpoint. For linear functions $f_\theta(x) = \langle \theta, x \rangle$, we clearly have that $\langle \partial f / \partial \theta, \theta \rangle = f_\theta(x)$. Remarkably, a direct analogue of this simple statement holds for neural networks, even if over-parametrized.

LEMMA 2.1 (Structure in Gradient). *Given a single data input $x \in \mathbb{R}^p$, consider the feedforward neural network in Definition 1 with activations satisfying (2.2). Then for any $0 \leq t \leq s \leq L$, one has the identity*

$$(2.5) \quad \sum_{i \in [k_t], j \in [k_{t+1}]} \frac{\partial O^{s+1}}{\partial W_{ij}^t} W_{ij}^t = O^{s+1}(x) .$$

In addition, it holds that

$$(2.6) \quad \sum_{t=0}^L \sum_{i \in [k_t], j \in [k_{t+1}]} \frac{\partial O^{L+1}}{\partial W_{ij}^t} W_{ij}^t = (L+1)O^{L+1}(x) .$$

Lemma 2.1 reveals the structural constraints in the gradients of rectified networks. In particular, even though the gradients lie in an over-parametrized high-dimensional space, many equality constraints are induced by the network architecture. Before we unveil the surprising connection between Lemma 2.1 and the proposed Fisher-Rao norm, let us take a look at a few immediate corollaries of this result. The first corollary establishes a large-margin property of stationary points that separate the data.

COROLLARY 2.1 (Large Margin Stationary Points). *Consider the binary classification problem with $\mathcal{Y} = \{-1, +1\}$, and a neural network where the output layer has only one unit. Choose the hinge loss $\ell(f, y) = \max\{0, 1 - yf\}$. If a certain parameter θ satisfies two properties*

1. θ is a stationary point for $\hat{L}(\theta)$ in the sense $\nabla_\theta \hat{L}(\theta) = \mathbf{0}$;
2. θ separates the data in the sense that $Y_i f_\theta(X_i) > 0$ for all $i \in [N]$,

then it must be that θ is a large margin solution: for all $i \in [N]$,

$$Y_i f_\theta(X_i) \geq 1 .$$

The same result holds for the population criteria $L(\theta)$, in which case (2) is stated as $\mathbb{P}(Y f_\theta(X) > 0) = 1$, and the conclusion is $\mathbb{P}(Y f_\theta(X) \geq 1) = 1$.

PROOF. Observe that $\frac{\partial \ell(f, Y)}{\partial f} = -y$ if $yf < 1$, and $\frac{\partial \ell(f, Y)}{\partial f} = 0$ if $yf \geq 1$. Using Eqn. (2.6) when the output layer has only one unit, we find

$$\begin{aligned} \langle \nabla_{\theta} \widehat{L}(\theta), \theta \rangle &= (L+1) \widehat{\mathbb{E}} \left[\frac{\partial \ell(f_{\theta}(X), Y)}{\partial f_{\theta}(X)} f_{\theta}(X) \right] , \\ &= (L+1) \widehat{\mathbb{E}} \left[-Y f_{\theta}(X) \mathbf{1}_{Y f_{\theta}(X) < 1} \right] . \end{aligned}$$

For a stationary point θ , we have $\nabla_{\theta} \widehat{L}(\theta) = \mathbf{0}$, which implies the LHS of the above equation is 0. Now recall that the second condition that θ separates the data implies $-Y f_{\theta}(X) < 0$ for any point in the data set. In this case, the RHS equals zero if and only if $Y f_{\theta}(X) \geq 1$. \square

Granted, the above corollary can be proved from first principles without the use of Lemma 2.1, but the proof reveals a quantitative statement about stationary points along arbitrary directions θ .

In the second corollary, we consider linear networks.

COROLLARY 2.2 (Stationary Points for Deep Linear Networks). *Consider linear neural networks with $\sigma(x) = x$ and square loss function. Then all stationary points $\theta = \{W^0, W^1, \dots, W^L\}$ that satisfy*

$$\nabla_{\theta} \widehat{L}(\theta) = \nabla_{\theta} \widehat{\mathbb{E}} \left[\frac{1}{2} (f_{\theta}(X) - Y)^2 \right] = 0 ,$$

must also satisfy

$$\langle w(\theta), \mathbf{X}^T \mathbf{X} w(\theta) - \mathbf{X}^T \mathbf{Y} \rangle = 0 ,$$

where $w(\theta) = \prod_{t=0}^L W^t \in \mathbb{R}^p$, $\mathbf{X} \in \mathbb{R}^{N \times p}$ and $\mathbf{Y} \in \mathbb{R}^N$ are the data matrices.

PROOF. The proof follows from applying Lemma 2.1

$$0 = \theta^T \nabla_{\theta} \widehat{L}(\theta) = (L+1) \widehat{\mathbb{E}} \left[(Y - X^T \prod_{t=0}^L W^t) X^T \prod_{t=0}^L W^t \right] ,$$

which means $\langle w(\theta), \mathbf{X}^T \mathbf{X} w(\theta) - \mathbf{X}^T \mathbf{Y} \rangle = 0$. \square

REMARK 2.1. This simple Lemma is not quite asserting that all stationary points are global optima, since global optima satisfy $\mathbf{X}^T \mathbf{X} w(\theta) - \mathbf{X}^T \mathbf{Y} = \mathbf{0}$, while we only proved that the stationary points satisfy $\langle w(\theta), \mathbf{X}^T \mathbf{X} w(\theta) - \mathbf{X}^T \mathbf{Y} \rangle = 0$.

3. FISHER-RAO NORM AND GEOMETRY

In this section, we propose a new notion of complexity of neural networks that can be motivated by geometrical invariance considerations, specifically the Fisher-Rao metric of information geometry. We postpone this motivation to Section 3.3 and instead start with the definition and some properties. Detailed comparison with the known norm-based capacity measures and generalization results are delayed to Section 4.

3.1 An analytical formula

DEFINITION 2. *The Fisher-Rao norm for a parameter θ is defined as the following quadratic form*

$$(3.1) \quad \|\theta\|_{\text{fr}}^2 := \langle \theta, \mathbf{I}(\theta)\theta \rangle, \quad \text{where } \mathbf{I}(\theta) = \mathbb{E}[\nabla_{\theta} \ell(f_{\theta}(X), Y) \otimes \nabla_{\theta} \ell(f_{\theta}(X), Y)].$$

The underlying distribution for the expectation in the above definition has been left ambiguous because it will be useful to specialize to different distributions depending on the context. Even though we call the above quantity the ‘‘Fisher-Rao norm,’’ it should be noted that it does not satisfy the triangle inequality. The following Theorem unveils a surprising identity for the Fisher-Rao norm.

THEOREM 3.1 (Fisher-Rao norm). *Assume the loss function $\ell(\cdot, \cdot)$ is smooth in the first argument. The following identity holds for a feedforward neural network (Definition 1) with L hidden layers and activations satisfying (2.2):*

$$(3.2) \quad \|\theta\|_{\text{fr}}^2 = (L + 1)^2 \mathbb{E} \left[\left\langle \frac{\partial \ell(f_{\theta}(X), Y)}{\partial f_{\theta}(X)}, f_{\theta}(X) \right\rangle^2 \right].$$

The proof of the Theorem relies mainly on the geometric Lemma 2.1 that describes the gradient structure of multi-layer rectified networks.

REMARK 3.1. In the case when the output layer has only one node, Theorem 3.1 reduces to the simple formula

$$(3.3) \quad \|\theta\|_{\text{fr}}^2 = (L + 1)^2 \mathbb{E} \left[\left(\frac{\partial \ell(f_{\theta}(X), Y)}{\partial f_{\theta}(X)} \right)^2 f_{\theta}(X)^2 \right].$$

PROOF OF THEOREM 3.1. Using the definition of the Fisher-Rao norm,

$$\begin{aligned} \|\theta\|_{\text{fr}}^2 &= \mathbb{E} [\langle \theta, \nabla_{\theta} \ell(f_{\theta}(X), Y) \rangle^2], \\ &= \mathbb{E} \left[\left\langle \nabla_{\theta} f_{\theta}(X) \frac{\partial \ell(f_{\theta}(X), Y)}{\partial f_{\theta}(X)}, \theta \right\rangle^2 \right], \\ &= \mathbb{E} \left[\left\langle \frac{\partial \ell(f_{\theta}(X), Y)}{\partial f_{\theta}(X)}, \nabla_{\theta} f_{\theta}(X)^T \theta \right\rangle^2 \right]. \end{aligned}$$

By Lemma 2.1,

$$\begin{aligned} \nabla_{\theta} f_{\theta}(X)^T \theta &= \nabla_{\theta} O^{L+1}(x)^T \theta, \\ &= \sum_{t=0}^L \sum_{i \in [k_t], j \in [k_{t+1}]} \frac{\partial O^{L+1}}{\partial W_{ij}^t} W_{ij}^t, \\ &= (L + 1) O^{L+1} = (L + 1) f_{\theta}(X). \end{aligned}$$

Combining the above equalities, we obtain

$$\|\theta\|_{\text{fr}}^2 = (L + 1)^2 \mathbb{E} \left[\left\langle \frac{\partial \ell(f_{\theta}(X), Y)}{\partial f_{\theta}(X)}, f_{\theta}(X) \right\rangle^2 \right].$$

□

Before illustrating how the explicit formula in Theorem 3.1 can be viewed as a unified “umbrella” for many of the known norm-based capacity measures, let us point out one simple invariance property of the Fisher-Rao norm, which follows as a direct consequence of Thm. 3.1. This property is not satisfied for ℓ_2 norm, spectral norm, path norm, or group norm.

COROLLARY 3.1 (Invariance). *If there are two parameters $\theta_1, \theta_2 \in \Theta_L$ such that they are equivalent, in the sense that $f_{\theta_1} = f_{\theta_2}$, then their Fisher-Rao norms are equal, i.e.,*

$$\|\theta_1\|_{\text{fr}} = \|\theta_2\|_{\text{fr}} .$$

3.2 Norms and geometry

In this section we will employ Theorem 3.1 to reveal the relationship among different norms and their corresponding geometries. Norm-based capacity control is an active field of research for understanding why deep learning generalizes well, including ℓ_2 norm (weight decay) in [Krogh and Hertz, 1992, Krizhevsky et al., 2012], path norm in [Neyshabur et al., 2015a], group-norm in [Neyshabur et al., 2015b], and spectral norm in [Bartlett et al., 2017]. All these norms are closely related to the Fisher-Rao norm, despite the fact that they capture distinct inductive biases and different geometries.

For simplicity, we will showcase the derivation with the absolute loss function $\ell(f, y) = |f - y|$ and when the output layer has only one node ($k_{L+1} = 1$). The argument can be readily adopted to the general setting. We will show that the Fisher-Rao norm serves as a lower bound for all the norms considered in the literature, with some pre-factor whose meaning will be clear in Section 4.1. In addition, the Fisher-Rao norm enjoys an interesting umbrella property: by considering a more constrained geometry (motivated from algebraic norm comparison inequalities) the Fisher-Rao norm motivates new norm-based capacity control methods.

The main theorem we will prove is informally stated as follows.

THEOREM 3.2 (Norm comparison, informal). *Denoting $\|\cdot\|$ as any one of: (1) spectral norm, (2) matrix induced norm, (3) group norm, or (4) path norm, we have*

$$\frac{1}{L+1} \|\theta\|_{\text{fr}} \leq \|\theta\| ,$$

for any $\theta \in \Theta_L = \{W^0, W^1, \dots, W^L\}$. The specific norms (1)-(4) are formally introduced in Definitions 3-6.

The detailed proof of the above theorem will be the main focus of Section 4.1. Here we will give a sketch on how the results are proved.

LEMMA 3.1 (Matrix form).

$$(3.4) \quad f_{\theta}(x) = x^T W^0 D^1(x) W^1 D^2(x) \cdots D^L W^L D^{L+1}(x) ,$$

where $D^t(x) = \text{diag}[\sigma'(N^t(x))] \in \mathbb{R}^{k_t \times k_t}$, for $0 < t \leq L+1$. In addition, $D^t(x)$ is a diagonal matrix with diagonal elements being either 0 or 1.

PROOF OF LEMMA 3.1. Since $O^0(x)W^0 = x^T W^0 = N^1(x) \in \mathbb{R}^{1 \times k_1}$, we have $N^1(x)D^1 = N^1(x)\text{diag}(\sigma'(N^1(x))) = O^1(x)$. Proof is completed via induction. \square

For the absolute loss, one has $(\partial \ell(f_\theta(X), Y) / \partial f_\theta(X))^2 = 1$ and therefore Theorem 3.1 simplifies to,

$$(3.5) \quad \|\theta\|_{\text{fr}}^2 = (L+1)^2 \mathbb{E}_{X \sim \mathcal{P}} [v(\theta, X)^T X X^T v(\theta, X)] \quad ,$$

where $v(\theta, x) := W^0 D^1(x) W^1 D^2(x) \cdots D^L(x) W^L D^{L+1}(x) \in \mathbb{R}^p$. The norm comparison results are thus established through a careful decomposition of the data-dependent vector $v(\theta, X)$, in distinct ways according to the comparing norm/geometry.

3.3 Motivation and invariance

In this section, we will provide the original intuition and motivation for our proposed Fisher-Rao norm from the viewpoint of geometric invariance.

Information geometry and the Fisher-Rao metric

Information geometry provides a window into geometric invariances when we adopt a generative framework where the data generating process belongs to the parametric family $\mathcal{P} \in \{\mathcal{P}_\theta \mid \theta \in \Theta_L\}$ indexed by the parameters of the neural network architecture. The Fisher-Rao metric on $\{\mathcal{P}_\theta\}$ is defined in terms of a local inner product for each value of $\theta \in \Theta_L$ as follows. For each $\alpha, \beta \in \mathbb{R}^d$ define the corresponding tangent vectors $\bar{\alpha} := dp_{\theta+t\alpha}/dt|_{t=0}$, $\bar{\beta} := dp_{\theta+t\beta}/dt|_{t=0}$. Then for all $\theta \in \Theta_L$ and $\alpha, \beta \in \mathbb{R}^d$ we define the local inner product

$$(3.6) \quad \langle \bar{\alpha}, \bar{\beta} \rangle_{p_\theta} := \int_M \frac{\bar{\alpha}}{p_\theta} \frac{\bar{\beta}}{p_\theta} p_\theta \quad ,$$

where $M = \mathcal{X} \times \mathcal{Y}$. The above inner product extends to a Riemannian metric on the space of positive densities $\text{Prob}(M)$ called the Fisher-Rao metric³. The relationship between the Fisher-Rao metric and the Fisher information matrix $\mathcal{I}(\theta)$ in statistics literature follows from the identity,

$$(3.7) \quad \langle \bar{\alpha}, \bar{\beta} \rangle_{p_\theta} = \langle \alpha, \mathcal{I}(\theta)\beta \rangle \quad .$$

Notice that the Fisher information matrix induces a *semi*-inner product $(\alpha, \beta) \mapsto \langle \alpha, \mathcal{I}(\theta)\beta \rangle$ unlike the Fisher-Rao metric which is non-degenerate⁴. If we make the additional modeling assumption that $p_\theta(x, y) = p(x)p_\theta(y|x)$ then the Fisher information becomes,

$$(3.8) \quad \mathcal{I}(\theta) = \mathbb{E}_{(X,Y) \sim p_\theta} [\nabla_\theta \log p_\theta(Y|X) \otimes \nabla_\theta \log p_\theta(Y|X)] \quad .$$

If we now identify our loss function as $\ell(f_\theta(x), y) = -\log p_\theta(y|x)$ then the Fisher-Rao metric coincides with the Fisher-Rao norm when $\alpha = \beta = \theta$. In fact, our Fisher-norm encompasses the Fisher-Rao metric and generalizes it to the case when the model is misspecified $\mathcal{P} \notin \{\mathcal{P}_\theta\}$.

Flatness

³Bauer et al. [2016] showed that it is essentially the the unique metric that is invariant under the diffeomorphism group of M .

⁴The null space of $\mathcal{I}(\theta)$ is mapped to the origin under $\alpha \mapsto dp_{\theta+t\alpha}/dt|_{t=0}$.

Having identified the geometric origin of Fisher-Rao norm, let us study the implications for generalization of flat minima. Dinh et al. [2017] argued by way of counter-example that the existing measures of flatness are inadequate for explaining the generalization capability of multi-layer neural networks. Specifically, by utilizing the invariance property of multi-layer rectified networks under non-negative nodewise rescalings, they proved that the Hessian eigenvalues of the loss function can be made arbitrarily large, thereby weakening the connection between flat minima and generalization. They also identified a more general problem which afflicts Hessian-based measures of generalization for any network architecture and activation function: the Hessian is sensitive to network parametrization whereas generalization should be invariant under general coordinate transformations. Our proposal can be motivated from the following fact⁵ which relates flatness to geometry (under appropriate regularity conditions)

$$(3.9) \quad \mathbb{E}_{(X,Y) \sim \mathcal{P}_\theta} \langle \theta, \text{Hess}_\theta [\ell(f_\theta(X), Y)] \theta \rangle = \mathbb{E}_{(X,Y) \sim \mathcal{P}_\theta} \langle \theta, \nabla_\theta \ell(f_\theta(X), Y) \rangle^2 = \|\theta\|_{\text{fr}}^2 .$$

In other words, the Fisher-Rao norm evades the node-wise rescaling issue because it is exactly invariant under linear re-parametrizations. The Fisher-Rao norm moreover possesses an “infinitesimal invariance” property under non-linear coordinate transformations, which can be seen by passing to the infinitesimal form where non-linear coordinate invariance is realized exactly by the following infinitesimal line element,

$$(3.10) \quad ds^2 = \sum_{i,j \in [d]} [\mathcal{I}(\theta)]_{ij} d\theta_i d\theta_j .$$

Comparing $\|\theta\|_{\text{fr}}$ with the above line element reveals the geometric interpretation of the Fisher-Rao norm as the approximate geodesic distance from the origin. It is important to realize that our definition of flatness (3.9) differs from [Dinh et al., 2017] who employed the Hessian loss $\text{Hess}_\theta [\hat{L}(\theta)]$. Unlike the Fisher-Rao norm, the norm induced by the Hessian loss does not enjoy the infinitesimal invariance property (it only holds at critical points).

Natural gradient

There exists a close relationship between the Fisher-Rao norm and the natural gradient. In particular, the natural gradient descent is simply the steepest descent direction induced by the Fisher-Rao geometry of $\{\mathcal{P}_\theta\}$. Indeed, the natural gradient can be expressed as a semi-norm-penalized iterative optimization scheme as follows,

$$(3.11) \quad \theta_{t+1} = \arg \min_{\theta \in \mathbb{R}^d} \left[\langle \theta - \theta_t, \nabla \hat{L}(\theta_t) \rangle + \frac{1}{2\eta_t} \|\theta - \theta_t\|_{\mathbf{I}(\theta_t)}^2 \right] = \theta_t - \eta_t \mathbf{I}(\theta_t)^+ \nabla \hat{L}(\theta_t) .$$

We remark that the positive semi-definite matrix $\mathbf{I}(\theta_t)$ changes with different t . We emphasize an “invariance” property of *natural gradient* under re-parametrization and an “approximate invariance” property under over-parametrization, which is not satisfied for the classic gradient descent. The formal statement and its proof are deferred to Lemma 6.1 in Section 6.2. The invariance property is desirable:

⁵Set $\ell(f_\theta(x), y) = -\log p_\theta(y|x)$ and recall the fact that Fisher information can be viewed as variance as well as the curvature.

in multi-layer ReLU networks, there are many equivalent re-parametrizations of the problem, such as nodewise rescalings, which may slow down the optimization process. The advantage of natural gradient is also illustrated empirically in Section 5.5.

4. CAPACITY CONTROL AND GENERALIZATION

In this section, we discuss in full detail the questions of geometry, capacity measures, and generalization. First, let us define empirical *Rademacher complexity* for the parameter space Θ , conditioned on data $\{X_i, i \in [N]\}$, as

$$(4.1) \quad \mathcal{R}_N(\Theta) = \mathbb{E} \sup_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^N \epsilon_i f_{\theta}(X_i) ,$$

where $\epsilon_i, i \in [N]$ are i.i.d. Rademacher random variables.

4.1 Norm Comparison

Let us collect some definitions before stating each norm comparison result. For a vector v , the vector ℓ_p norm is denoted $\|v\|_p := (\sum_i |v_i|^p)^{1/p}$, $p > 0$. For a matrix M , $\|M\|_{\sigma} := \max_{v \neq 0} \|v^T M\| / \|v\|$ denotes the spectral norm; $\|M\|_{p \rightarrow q} = \max_{v \neq 0} \|v^T M\|_q / \|v\|_p$ denotes the matrix induced norm, for $p, q \geq 1$; $\|M\|_{p,q} = [\sum_j (\sum_i |M_{ij}|^p)^{q/p}]^{1/q}$ denotes the matrix group norm, for $p, q \geq 1$.

4.1.1 Spectral norm.

DEFINITION 3 (Spectral norm). Define the following ‘‘spectral norm’’ ball:

$$(4.2) \quad \|\theta\|_{\sigma} := \left[\mathbb{E} \left(\|X\|^2 \prod_{t=1}^{L+1} \|D^t(X)\|_{\sigma}^2 \right) \right]^{1/2} \prod_{t=0}^L \|W^t\|_{\sigma} .$$

We have the following norm comparison Lemma.

LEMMA 4.1 (Spectral norm).

$$\frac{1}{L+1} \|\theta\|_{\text{fr}} \leq \|\theta\|_{\sigma} .$$

REMARK 4.1. Spectral norm as a capacity control has been considered in [Bartlett et al., 2017]. Lemma 4.1 shows that spectral norm serves as a more stringent constraint than Fisher-Rao norm. Let us provide an explanation of the pre-factor $[\mathbb{E}(\|X\|^2 \prod_{t=1}^{L+1} \|D^t(X)\|_{\sigma}^2)]^{1/2}$ here. Define the set of parameters induced by the Fisher-Rao norm geometry

$$B_{\text{fr}}(1) := \{\theta : \mathbb{E}[v(\theta, X)^T X X^T v(\theta, X)] \leq 1\} = \{\theta : \frac{1}{L+1} \|\theta\|_{\text{fr}} \leq 1\} .$$

From Lemma 4.1, if the expectation is over the empirical measure $\hat{\mathbb{E}}$, then, because $\|D^t(X)\|_{\sigma} \leq 1$, we obtain

$$\frac{1}{L+1} \|\theta\|_{\text{fr}} \leq \left[\hat{\mathbb{E}} \left(\|X\|^2 \prod_{t=1}^{L+1} \|D^t(X)\|_{\sigma}^2 \right) \right]^{1/2} \prod_{t=0}^L \|W^t\|_{\sigma} \leq [\hat{\mathbb{E}}\|X\|^2]^{1/2} \prod_{t=0}^L \|W^t\|_{\sigma} ,$$

$$\text{which implies } \left\{ \theta : \prod_{t=0}^L \|W^t\|_{\sigma} \leq \frac{1}{[\hat{\mathbb{E}}\|X\|^2]^{1/2}} \right\} \subset B_{\text{fr}}(1) .$$

From Theorem 1.1 in [Bartlett et al., 2017], we know that a subset of the $B_{\text{fr}}(1)$ characterized by the *spectral norm* enjoys the following upper bound on Rademacher complexity under mild conditions: for any $r > 0$

$$(4.3) \quad \mathcal{R}_N \left(\left\{ \theta : \prod_{t=0}^L \|W^t\|_{\sigma} \leq r \right\} \right) \lesssim r \cdot \frac{[\widehat{\mathbb{E}}\|X\|^2]^{1/2} \cdot \text{Polylog}}{N} .$$

Plugging in $r = \frac{1}{[\widehat{\mathbb{E}}\|X\|^2]^{1/2}}$, we have,

$$(4.4) \quad \mathcal{R}_N \left(\left\{ \theta : \prod_{t=0}^L \|W^t\|_{\sigma} \leq \frac{1}{[\widehat{\mathbb{E}}\|X\|^2]^{1/2}} \right\} \right) \lesssim \frac{1}{[\widehat{\mathbb{E}}\|X\|^2]^{1/2}} \cdot \frac{[\widehat{\mathbb{E}}\|X\|^2]^{1/2} \cdot \text{Polylog}}{N} \rightarrow 0 .$$

Interestingly, the additional factor $[\widehat{\mathbb{E}}\|X\|^2]^{1/2}$ in Theorem 1.1 in [Bartlett et al., 2017] exactly cancels with our pre-factor in the norm comparison. The above calculations show that a subset of $B_{\text{fr}}(1)$, induced by the spectral norm geometry, has good generalization error.

4.1.2 Group norm.

DEFINITION 4 (Group norm). Define the following “group norm” ball, for $p \geq 1, q > 0$

$$(4.5) \quad \|\theta\|_{p,q} := \left[\mathbb{E} \left(\|X\|_{p^*}^2 \prod_{t=1}^{L+1} \|D^t(X)\|_{q \rightarrow p^*}^2 \right) \right]^{1/2} \prod_{t=0}^L \|W^t\|_{p,q} ,$$

where $\frac{1}{p} + \frac{1}{p^*} = 1$. Here $\|\cdot\|_{q \rightarrow p^*}$ denote the matrix induced norm.

LEMMA 4.2 (Group norm). *It holds that*

$$(4.6) \quad \frac{1}{L+1} \|\theta\|_{\text{fr}} \leq \|\theta\|_{p,q} .$$

REMARK 4.2. Group norm as a capacity measure has been considered in [Neyshabur et al., 2015b]. Lemma 4.2 shows that group norm serves as a more stringent constraint than Fisher-Rao norm. Again, let us provide an explanation of the pre-factor $[\mathbb{E}(\|X\|_{p^*}^2 \prod_{t=1}^{L+1} \|D^t(X)\|_{q \rightarrow p^*}^2)]^{1/2}$ here.

Note that for all X

$$\prod_{t=1}^{L+1} \|D^t(X)\|_{q \rightarrow p^*} \leq \prod_{t=1}^{L+1} k_t^{[\frac{1}{p^*} - \frac{1}{q}]_+} ,$$

because

$$\|D^t(X)\|_{q \rightarrow p^*} = \max_{v \neq 0} \frac{\|v^T D^t(X)\|_{p^*}}{\|v\|_q} \leq \max_{v \neq 0} \frac{\|v\|_{p^*}}{\|v\|_q} \leq k_t^{[\frac{1}{p^*} - \frac{1}{q}]_+} .$$

From Lemma 4.2, if the expectation is over the empirical measure $\widehat{\mathbb{E}}$, we know that in the case when $k_t = k$ for all $0 < t \leq L$,

$$\begin{aligned} \frac{1}{L+1} \|\theta\|_{\text{fr}} &\leq \left[\widehat{\mathbb{E}} \left(\|X\|_{p^*}^2 \prod_{t=1}^{L+1} \|D^t(X)\|_{q \rightarrow p^*}^2 \right) \right]^{1/2} \prod_{t=0}^L \|W^t\|_{p,q} , \\ &\leq \left(\max_i \|X_i\|_{p^*}^2 \right)^{1/2} \left(k^{\lfloor \frac{1}{p^*} - \frac{1}{q} \rfloor_+} \right)^L \cdot \prod_{t=0}^L \|W^t\|_{p,q} , \end{aligned}$$

$$\text{which implies } \left\{ \theta : \prod_{t=0}^L \|W^t\|_{p,q} \leq \frac{1}{\left(k^{\lfloor \frac{1}{p^*} - \frac{1}{q} \rfloor_+} \right)^L \max_i \|X_i\|_{p^*}} \right\} \subset B_{\text{fr}}(1) .$$

By Theorem 1 in [Neyshabur et al., 2015b], we know that a subset of $B_{\text{fr}}(1)$ (different from the subset induced by spectral geometry) characterized by the *group norm*, satisfies the following upper bound on the Rademacher complexity, for any $r > 0$

$$(4.7) \quad \mathcal{R}_N \left(\left\{ \theta : \prod_{t=0}^L \|W^t\|_{p,q} \leq r \right\} \right) \lesssim r \cdot \frac{\left(2k^{\lfloor \frac{1}{p^*} - \frac{1}{q} \rfloor_+} \right)^L \max_i \|X_i\|_{p^*} \cdot \text{Polylog}}{\sqrt{N}} .$$

Plugging in $r = \frac{1}{\left(k^{\lfloor \frac{1}{p^*} - \frac{1}{q} \rfloor_+} \right)^L \max_i \|X_i\|_{p^*}}$, we have

$$(4.8) \quad \begin{aligned} &\mathcal{R}_N \left(\left\{ \theta : \prod_{t=0}^L \|W^t\|_{p,q} \leq \frac{1}{\left(k^{\lfloor \frac{1}{p^*} - \frac{1}{q} \rfloor_+} \right)^L \max_i \|X_i\|_{p^*}} \right\} \right) \\ &\lesssim \frac{1}{\left(k^{\lfloor \frac{1}{p^*} - \frac{1}{q} \rfloor_+} \right)^L \max_i \|X_i\|_{p^*}} \cdot \frac{2^L \left(k^{\lfloor \frac{1}{p^*} - \frac{1}{q} \rfloor_+} \right)^L \max_i \|X_i\|_{p^*} \cdot \text{Polylog}}{\sqrt{N}} \rightarrow 0 . \end{aligned}$$

Once again, we point out that the intriguing combinatorial factor in Theorem 1 of Neyshabur et al. [2015b] exactly cancels with our pre-factor in the norm comparison. The above calculations show that another subset of $B_{\text{fr}}(1)$, induced by the group norm geometry, has good generalization error (without additional factors).

4.1.3 Path norm.

DEFINITION 5 (Path norm). Define the following “path norm” ball, for $q \geq 1$

$$(4.9) \quad \|\pi(\theta)\|_q := \left[\mathbb{E} \left(\sum_{i_0, i_1, \dots, i_L} |X_{i_0}| \prod_{t=1}^{L+1} |D_{i_t}^t(X)|^{q^*} \right)^{2/q^*} \right]^{1/2} \cdot \left(\sum_{i_0, i_1, \dots, i_L} \prod_{t=0}^L |W_{i_t i_{t+1}}^t|^q \right)^{1/q} ,$$

where $\frac{1}{q} + \frac{1}{q^*} = 1$, indices set $i_0 \in [p], i_1 \in [k_1], \dots, i_L \in [k_L], i_{L+1} = 1$. Here $\pi(\theta)$ is a notation for all the paths (from input to output) of the weights θ .

LEMMA 4.3 (Path- q norm). *The following inequality holds for any $q \geq 1$,*

$$(4.10) \quad \frac{1}{L+1} \|\theta\|_{\text{fr}} \leq \|\pi(\theta)\|_q .$$

REMARK 4.3. Path norm has been investigated in [Neyshabur et al., 2015a], where the definition is

$$\left(\sum_{i_0, i_1, \dots, i_L} \prod_{t=0}^L |W_{i_t i_{t+1}}^t|^q \right)^{1/q} .$$

Again, let us provide an intuitive explanation for our pre-factor

$$\left[\mathbb{E} \left(\sum_{i_0, i_1, \dots, i_L} |X_{i_0}| \prod_{t=1}^{L+1} |D_{i_t}^t(X)|^{q^*} \right)^{2/q^*} \right]^{1/2} ,$$

here for the case $q = 1$. Due to Lemma 4.3, when the expectation is over empirical measure,

$$\begin{aligned} \frac{1}{L+1} \|\theta\|_{\text{fr}} &\leq \left[\widehat{\mathbb{E}} \left(\sum_{i_0, i_1, \dots, i_L} |X_{i_0}| \prod_{t=1}^{L+1} |D_{i_t}^t(X)|^{q^*} \right)^{2/q^*} \right]^{1/2} \cdot \left(\sum_{i_0, i_1, \dots, i_L} \prod_{t=0}^L |W_{i_t i_{t+1}}^t|^q \right)^{1/q} , \\ &\leq \max_i \|X_i\|_\infty \cdot \left(\sum_{i_0, i_1, \dots, i_L} \prod_{t=0}^L |W_{i_t i_{t+1}}^t| \right) , \end{aligned}$$

$$\text{which implies } \left\{ \theta : \sum_{i_0, i_1, \dots, i_L} \prod_{t=0}^L |W_{i_t i_{t+1}}^t| \leq \frac{1}{\max_i \|X_i\|_\infty} \right\} \subset B_{\text{fr}}(1) .$$

By Corollary 7 in [Neyshabur et al., 2015b], we know that for any $r > 0$, the Rademacher complexity of path-1 norm ball satisfies

$$\mathcal{R}_N \left(\left\{ \theta : \sum_{i_0, i_1, \dots, i_L} \prod_{t=0}^L |W_{i_t i_{t+1}}^t| \leq r \right\} \right) \lesssim r \cdot \frac{2^L \max_i \|X_i\|_\infty \cdot \text{Polylog}}{\sqrt{N}} .$$

Plugging in $r = \frac{1}{\max_i \|X_i\|_\infty}$, we find that the subset of Fisher-Rao norm ball $B_{\text{fr}}(1)$ induced by path-1 norm geometry, satisfies

$$\mathcal{R}_N \left(\left\{ \theta : \sum_{i_0, i_1, \dots, i_L} \prod_{t=0}^L |W_{i_t i_{t+1}}^t| \leq \frac{1}{\max_i \|X_i\|_\infty} \right\} \right) \lesssim \frac{1}{\max_i \|X_i\|_\infty} \cdot \frac{2^L \max_i \|X_i\|_\infty \cdot \text{Polylog}}{\sqrt{N}} \rightarrow 0 .$$

Once again, the additional factor appearing in the Rademacher complexity bound in [Neyshabur et al., 2015b], cancels with our pre-factor in the norm comparison.

4.1.4 Matrix induced norm.

DEFINITION 6 (Induced norm). Define the following ‘‘matrix induced norm’’ ball, for $p, q > 0$, as

$$(4.11) \quad \|\theta\|_{p \rightarrow q} := \left[\mathbb{E} \left(\|X\|_p^2 \prod_{t=1}^{L+1} \|D^t(X)\|_{q \rightarrow p}^2 \right) \right]^{1/2} \prod_{t=0}^L \|W^t\|_{p \rightarrow q} .$$

LEMMA 4.4 (Matrix induced norm). *For any $p, q > 0$, the following inequality holds*

$$\frac{1}{L+1} \|\theta\|_{\text{fr}} \leq \|\theta\|_{p \rightarrow q} .$$

Remark that $\|D^t(X)\|_{q \rightarrow p}^2$ may contain dependence on k when $p \neq q$. This motivates us to consider the following generalization of matrix induced norm, where the norm for each W^t can be different.

DEFINITION 7 (Chain of induced norm). Define the following “chain of induced norm” ball, for a chain of $P = (p_0, p_1, \dots, p_{L+1}), p_i > 0$

$$(4.12) \quad \|\theta\|_P := \left[\mathbb{E} \left(\|X\|_{p_0}^2 \prod_{t=1}^{L+1} \|D^t(X)\|_{p_t \rightarrow p_t}^2 \right) \right]^{1/2} \prod_{t=0}^L \|W^t\|_{p_t \rightarrow p_{t+1}} .$$

LEMMA 4.5 (Chain of induced norm). *It holds that*

$$\frac{1}{L+1} \|\theta\|_{\text{fr}} \leq \|\theta\|_P .$$

REMARK 4.4. Lemma 4.5 exhibits a new flexible norm that dominates the Fisher-Rao norm. The example shows that one can motivate a variety of new norms (and their corresponding geometry) as subsets of the Fisher-Rao norm ball.

We will conclude this section with two geometric observations about the Fisher-Rao norm with absolute loss function $\ell(f, y) = |f - y|$ and one output node. In this case, even though $\{\theta : \frac{1}{L+1} \|\theta\|_{\text{fr}} \leq 1\}$ is non-convex, it is star-shaped.

LEMMA 4.6 (Star shape). *For any $\theta \in \Theta$, let $\{r\theta, r > 0\}$ denote the line connecting between 0 and θ to infinity. Then one has,*

$$\frac{d}{dr} \|r\theta\|_{\text{fr}}^2 = \frac{2(L+1)}{r} \|r\theta\|_{\text{fr}}^2 .$$

This also implies

$$\|r\theta\|_{\text{fr}} = r^{L+1} \|\theta\|_{\text{fr}} .$$

Despite the non-convexity of set of parameters with a bound on the Fisher-Rao norm, there is certain “convexity” in the function space:

LEMMA 4.7 (Convexity in f_θ). *For any $\theta_1, \theta_2 \in \Theta_L$ that satisfy*

$$\frac{1}{L+1} \|\theta_i\|_{\text{fr}} \leq 1, \quad \text{for } i = 1 \text{ or } 2,$$

we have for any $0 < \lambda < 1$, the convex combination $\lambda f_{\theta_1} + (1 - \lambda) f_{\theta_2}$ can be realized by a parameter $\theta' \in \Theta_{L+1}$ in the sense

$$f_{\theta'} = \lambda f_{\theta_1} + (1 - \lambda) f_{\theta_2} ,$$

and satisfies

$$\frac{1}{(L+1)+1} \|\theta'\|_{\text{fr}} \leq 1 .$$

4.2 Generalization

In this section, we will investigate the generalization puzzle for deep learning through the lens of the Fisher-Rao norm. We will first introduce a rigorous proof in the case of multi-layer linear networks, that capacity control with Fisher-Rao norm ensures good generalization. Then we will provide a heuristic argument why Fisher-Rao norm seems to be the right norm-based capacity control for rectified neural networks, via norm comparison in Section 4.1. We complement our heuristic argument with extensive numerical investigations in Section 5.

THEOREM 4.1 (Deep Linear Networks). *Consider multi-layer linear networks with $\sigma(x) = x$, L hidden layers, input dimension p and single output unit, and parameters $\theta \in \Theta_L = \{W^0, W^1, \dots, W^L\}$. Define the Fisher-Rao norm ball as in Eqn. (3.5)*

$$B_{\text{fr}}(\gamma) = \left\{ \theta : \frac{1}{L+1} \|\theta\|_{\text{fr}} \leq \gamma \right\} .$$

Then we have

$$(4.13) \quad \mathbb{E} \mathcal{R}_N(B_{\text{fr}}(\gamma)) \leq \gamma \sqrt{\frac{p}{N}} ,$$

assuming the Gram matrix $\mathbb{E}[XX^T] \in \mathbb{R}^{p \times p}$ is full rank.⁶

REMARK 4.5. Combining the above Theorem with classic symmetrization and margin bounds [Koltchinskii and Panchenko, 2002], one can deduce that for binary classification, the following generalization guarantee holds (for any margin parameter $\alpha > 0$),

$$\mathbb{E} \mathbf{1}[f_{\theta}(X)Y < 0] \leq \frac{1}{N} \sum_i \mathbf{1}[f_{\theta}(X_i)Y_i \leq \alpha] + \frac{C}{\alpha} \mathcal{R}_N(B_{\text{fr}}(\gamma)) + C \sqrt{\frac{\log 1/\delta}{N}} ,$$

for any $\theta \in B_{\text{fr}}(\gamma)$ with probability at least $1 - \delta$, where $C > 0$ is some constant.

We would like to emphasize that to explain generalization in this over-parametrized multi-layer linear network, it is indeed desirable that the generalization error in Theorem 4.1 only depends on the Fisher-Rao norm and the intrinsic input dimension p , without additional dependence on other network parameters (such as width, depth) and the X dependent factor.

In the case of ReLU networks, it turns out that bounding $\mathcal{R}_N(B_{\text{fr}}(\gamma))$ in terms of the Fisher-Rao norm is a very challenging task. Instead, we provide heuristic arguments via bounding the Rademacher complexity for various subsets of $B_{\text{fr}}(\gamma)$. As discussed in Remarks 4.1-4.3, the norms considered (spectral, group, and path norm) can be viewed as a subset of unit Fisher-Rao norm ball induced by distinct

⁶This assumption is to simplify exposition, and can be removed.

geometry. To remind ourselves, we have shown

$$\begin{aligned} \text{spectral norm } B_{\|\cdot\|_\sigma} &:= \left\{ \theta : \prod_{t=0}^L \|W^t\|_\sigma \leq \frac{\gamma}{\left[\widehat{\mathbb{E}}\|X\|^2\right]^{1/2}} \right\} \subset B_{\text{fr}}(\gamma) , \\ \text{group } p, q \text{ norm } B_{\|\cdot\|_{p,q}} &:= \left\{ \theta : \prod_{t=0}^L \|W^t\|_{p,q} \leq \frac{\gamma}{\left(k^{\left[\frac{1}{p^*} - \frac{1}{q}\right]_+}\right)^L \max_i \|X_i\|_{p^*}} \right\} \subset B_{\text{fr}}(\gamma) , \\ \text{path-1 norm } B_{\|\pi(\cdot)\|_1} &:= \left\{ \theta : \sum_{i_0, i_1, \dots, i_L} \prod_{t=0}^L |W_{i_t i_{t+1}}^t| \leq \frac{\gamma}{\max_i \|X_i\|_\infty} \right\} \subset B_{\text{fr}}(\gamma) , \end{aligned}$$

from which the following bounds follow,

$$\begin{aligned} \mathcal{R}_N(B_{\|\cdot\|_\sigma}) &\leq \gamma \cdot \frac{\text{Polylog}}{N} , \\ \mathcal{R}_N(B_{\|\cdot\|_{p,q}}) &\leq \gamma \cdot \frac{2^L \text{Polylog}}{\sqrt{N}} , \\ \mathcal{R}_N(B_{\|\pi(\cdot)\|_1}) &\leq \gamma \cdot \frac{2^L \text{Polylog}}{\sqrt{N}} . \end{aligned}$$

The surprising fact is that despite the distinct geometry of the subsets $B_{\|\cdot\|_\sigma}$, $B_{\|\cdot\|_{p,q}}$ and $B_{\|\pi(\cdot)\|_1}$ (which are described by different norms), the Rademacher complexity of these sets all depend on the “enveloping” Fisher-Rao norm explicitly without either the intriguing combinatorial factor or the X dependent factor. We believe this envelope property sheds light on how to compare different norm-based capacity measures.

Before concluding this section, we present the contour plot of Fisher-Rao norm and path-2 norm in a simple two layer ReLU network in Fig. 1, to better illustrate the geometry of Fisher-Rao norm and the subset induced by other norm. We choose two weights as x, y -axis and plot the levelsets of the norms.

5. EXPERIMENTS

5.1 Experimental details

In the realistic K -class classification context there is no activation function on the K -dimensional output layer of the network ($\sigma_{L+1}(x) = x$) and we focus on ReLU activation $\sigma(x) = \max\{0, x\}$ for the intermediate layers. The loss function is taken to be the cross entropy $\ell(y', y) = -\langle e_y, \log g(y') \rangle$, where $e_y \in \mathbb{R}^K$ denotes the one-hot-encoded class label and $g(z)$ is the softmax function defined by,

$$g(z) = \left(\frac{\exp(z_1)}{\sum_{k=1}^K \exp(z_k)}, \dots, \frac{\exp(z_K)}{\sum_{k=1}^K \exp(z_k)} \right)^T .$$

It can be shown that the gradient of the loss function with respect to the output of the neural network is $\nabla \ell(f, y) = -\nabla \langle e_y, \log g(f) \rangle = g(f) - e_y$, so plugging into the general expression for the Fisher-Rao norm we obtain,

$$(5.1) \quad \|\theta\|_{\text{fr}}^2 = (L+1)^2 \mathbb{E}[\{\langle g(f_\theta(X)), f_\theta(X) \rangle - f_\theta(X)_Y\}^2].$$

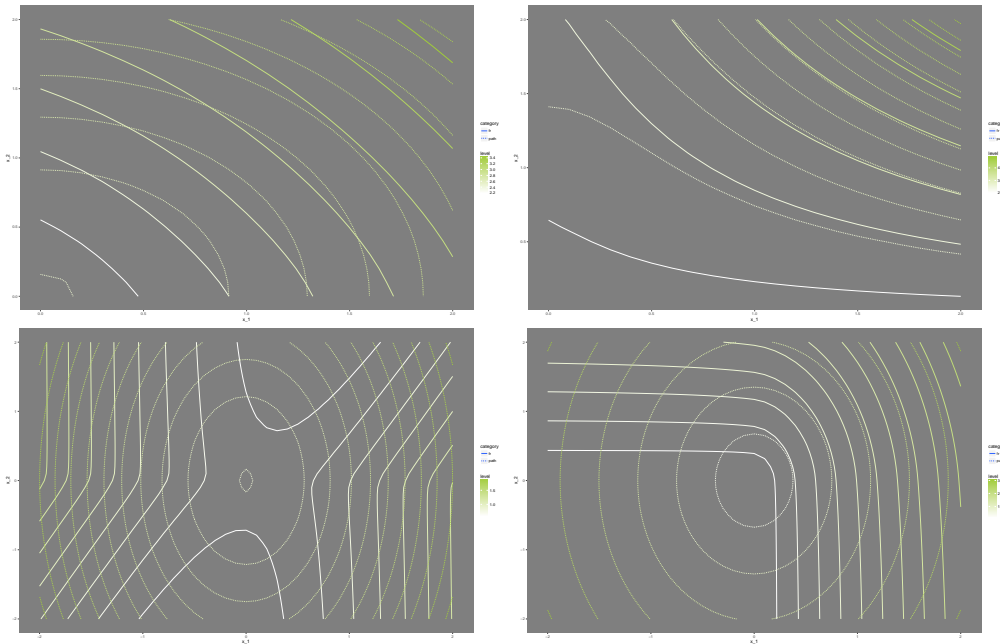


FIG 1. The levelsets of Fisher-Rao norm (solid) and path-2 norm (dotted). The color denotes the value of the norm.

In practice, since we do not have access to the population density $p(x)$ of the covariates, we estimate the Fisher-Rao norm by sampling from a test set of size m , leading to our final formulas

$$(5.2) \quad \|\theta\|_{\text{fr}}^2 = (L+1)^2 \frac{1}{m} \sum_{i=1}^m \sum_{y=1}^K g(f_{\theta}(x_i))_y [\langle g(f_{\theta}(x_i)), f_{\theta}(x_i) \rangle - f_{\theta}(x_i)_y]^2,$$

$$(5.3) \quad \|\theta\|_{\text{fr,emp}}^2 = (L+1)^2 \frac{1}{m} \sum_{i=1}^m [\langle g(f_{\theta}(x_i)), f_{\theta}(x_i) \rangle - f_{\theta}(x_i)_{y_i}]^2.$$

5.2 Over-parametrization with Hidden Units

In order to understand the effect of network over-parametrization we investigated the relationship between different proposals for capacity control and the number of parameters $d = pk_1 + \sum_{i=1}^{L-1} k_i k_{i+1} + k_L K$ of the neural network. For simplicity we focused on a fully connected architecture consisting of L hidden layers with k neurons per hidden layer so that the expression simplifies to $d = k[p + k(L-1) + K]$. The network parameters were learned by minimizing the cross-entropy loss on the CIFAR-10 image classification dataset with no explicit regularization nor data augmentation. The cross-entropy loss was optimized using 200 epochs of minibatch gradient descent utilizing minibatches of size 50 and otherwise identical experimental conditions described in [Zhang et al., 2016]. The same experiment was repeated using minibatch natural gradient descent employing the Kronecker-factored approximate curvature (K-FAC) method [Martens and Grosse, 2015] with the same learning rate and momentum schedules. The first fact we observe is that the Fisher-Rao norm remains approximately constant (or decreasing) when the network is overparametrized by increasing the width k at fixed depth $L = 2$ (see Fig. 2). If we vary the depth L of the network at fixed

width $k = 500$ then we find that the Fisher-Rao norm is essentially constant when measured in its ‘natural units’ of $L + 1$ (see Fig. 3). Finally, if we compare each proposal based on its absolute magnitude, the Fisher-Rao norm is distinguished as the minimum-value norm, and becomes $O(1)$ when evaluated using the model distribution. This self-normalizing property can be understood as a consequence of the relationship to flatness discussed in section 3.3, which holds when the expectation is taken with respect to the model.

5.3 Corruption with Random Labels

Over-parametrized neural networks tend to exhibit good generalization despite perfectly fitting the training set [Zhang et al., 2016]. In order to pinpoint the “correct” notion of complexity which drives generalization error, we conducted a series of experiments in which we changed both the network size and the signal-to-noise ratio of the datasets. In particular, we focus on the set of neural architectures obtained by varying the hidden layer width k at fixed depth $L = 2$ and moreover for each training/test example we assign a random label with probability α .

It can be seen from the last two panels of Fig. 5 and 4 that for non-random labels ($\alpha = 0$), the empirical Fisher-Rao norm actually decreases with increasing k , in tandem with the generalization error and moreover this correlation seems to persist when we vary the label randomization. Overall the Fisher-Rao norm is distinguished from other measures of capacity by the fact that its empirical version seems to track the generalization gap and moreover this trend does not appear to be sensitive to the choice of optimization.

It is also interesting to note that the Fisher-Rao norm has a stability property with respect to increasing k which suggests that a formal $k \rightarrow \infty$ limit might exist. Finally, we note that unlike the vanilla gradient, the natural gradient differentiates the different architectures by their Fisher-Rao norm. Although we don’t completely understand this phenomenon, it is likely a consequence of the fact that the natural gradient is iteratively minimizing the Fisher-Rao semi-norm.

5.4 Margin Story

Bartlett et al. [2017] adopted the margin story to explain generalization. They investigated the spectrally-normalized margin to explain why CIFAR-10 with random labels is a harder dataset (generalize poorly) than the uncorrupted CIFAR-10 (which generalize well). Here we adopt the same idea in this experiment, where we plot margin normalized by the empirical Fisher-Rao norm, in comparison to the spectral norm, based on the model trained either by vanilla gradient and natural gradient. It can be seen from Fig. 6 that the Fisher-Rao-normalized margin also accounts for the generalization gap between random and original CIFAR-10. In addition, Table 1 shows that the empirical Fisher-Rao norm improves the normalized margin relative to the spectral norm. These results were obtained by optimizing with the natural gradient but are not sensitive to the choice of optimizer.

5.5 Natural Gradient and Pre-conditioning

It was shown in [Shalev-Shwartz et al., 2017] that multi-layer networks struggle to learn certain piecewise-linear curves because the problem instances are poorly-conditioned. The failure was attributed to the fact that simply using a black-box model without a deeper analytical understanding of the problem structure could

	Model Fisher-Rao	Empirical Fisher-Rao	Spectral
$\alpha = 0$	1.61	22.68	136.67
$\alpha = 1$	2.12	35.98	205.56
Ratio	0.76	0.63	0.66

TABLE 1

Comparison of Fisher-Rao norm and spectral norm after training with natural gradient using original dataset ($\alpha = 0$) and with random labels ($\alpha = 1$). Qualitatively similar results holds for GD+momentum.

be computationally sub-optimal. Our results suggest that the problem can be overcome within the confines of black-box optimization by using natural gradient. In other words, the natural gradient automatically pre-conditions the problem and appears to achieve similar performance as that attained by hard-coded convolutions [Shalev-Shwartz et al., 2017], within the same number of iterations (see Fig. 7).

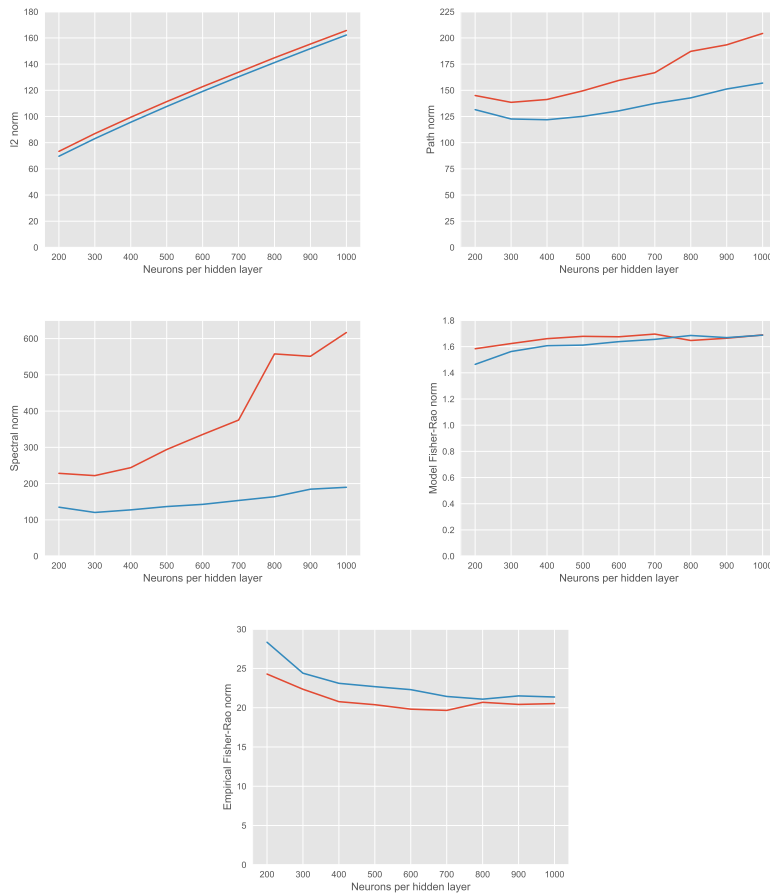


FIG 2. Dependence of different norms on width k of hidden layers ($L = 2$) after optimizing with vanilla gradient descent (red) and natural gradient descent (blue).

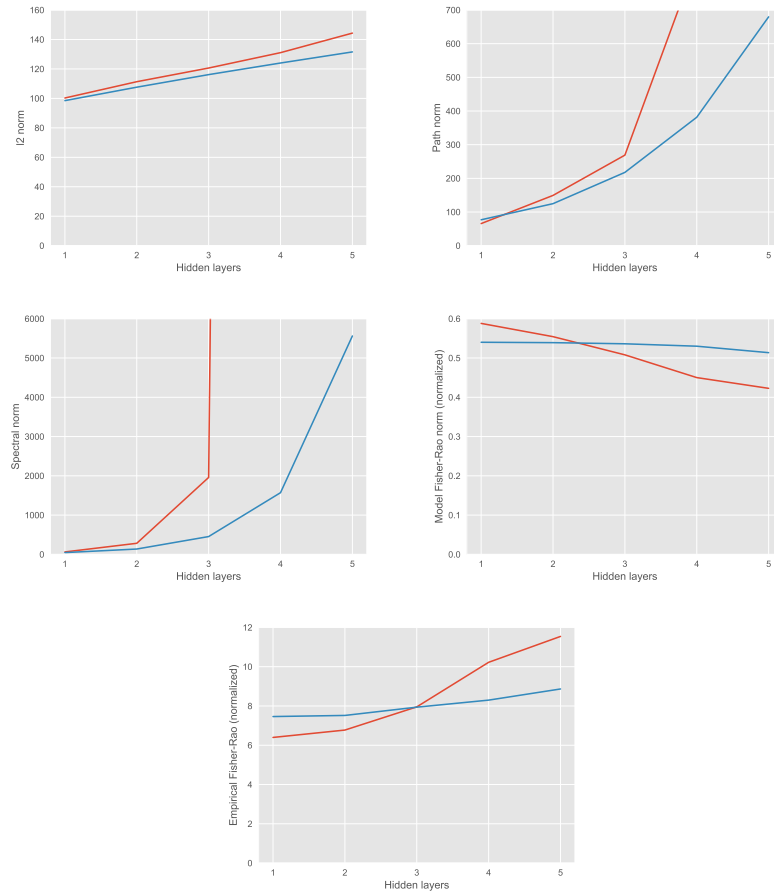


FIG 3. Dependence of different norms on depth L ($k = 500$) after optimizing with vanilla gradient descent (red) and natural gradient descent (blue). The Fisher-Rao norms are normalized by $L+1$.

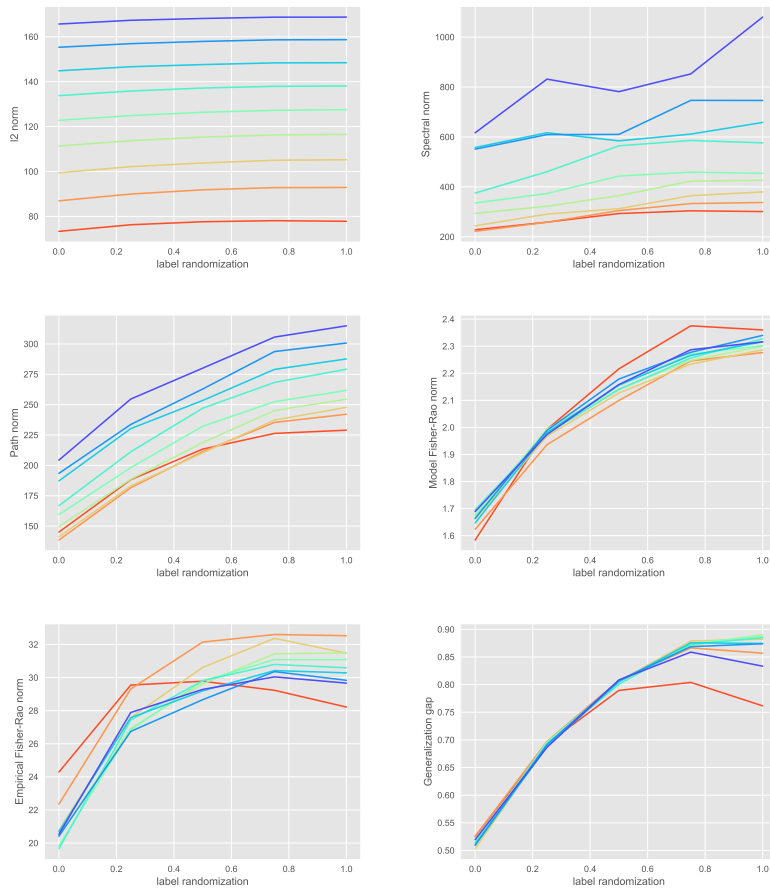


FIG 4. Dependence of capacity measures on label randomization after optimizing with vanilla gradient descent. The colors show the effect of varying network width from $k = 200$ (red) to $k = 1000$ (blue) in increments of 100.

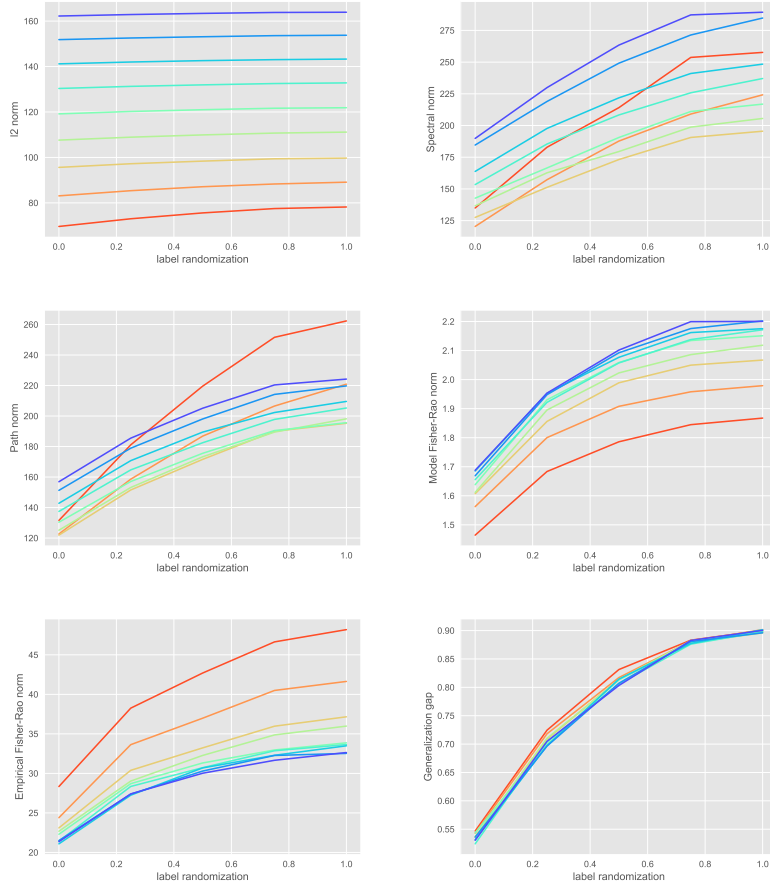


FIG 5. Dependence of capacity measures on label randomization after optimizing with the natural gradient descent. The colors show the effect of varying network width from $k = 200$ (red) to $k = 1000$ (blue) in increments of 100. The natural gradient optimization clearly distinguishes the network architectures according to their Fisher-Rao norm.

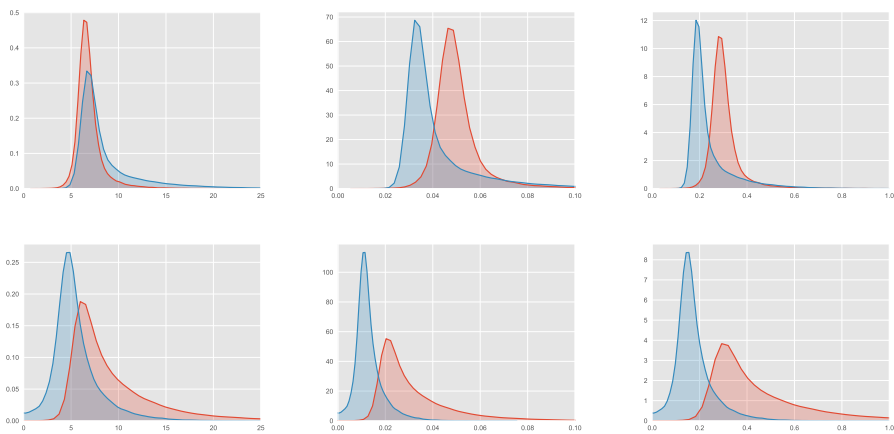


FIG 6. Distribution of margins found by natural gradient (top) and vanilla gradient (bottom) before rescaling (left) and after rescaling by spectral norm (center) and empirical Fisher-Rao norm (right).

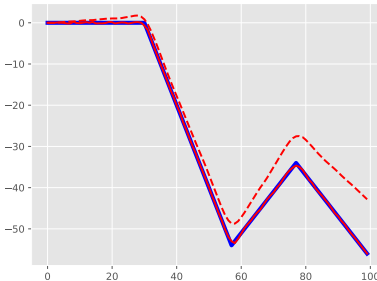


FIG 7. *Reproduction of conditioning experiment from [Shalev-Shwartz et al., 2017] after 10^4 iterations of Adam (dashed) and K-FAC (red).*

6. FURTHER DISCUSSION

In this paper we studied the generalization puzzle of deep learning from an invariance point of view. The notions of invariance come from several angles: invariance in information geometry, invariance of non-linear local transformation, invariance under function equivalence, algorithmic invariance of parametrization, “flat” minima invariance of linear transformations, among many others. We proposed a new non-convex capacity measure using the Fisher-Rao norm. We demonstrated the good properties of Fisher-Rao norm as a capacity measure both from the theoretical and the empirical side.

6.1 Parameter identifiability

Let us briefly discuss the aforementioned parameter identifiability issue in deep networks. The function classes considered in this paper admit various group actions, which leave the function output invariant⁷. This means that our hypothesis class is in bijection with the equivalence class $\mathcal{H}_L \cong \Theta / \sim$ where we identify $\theta \sim \theta'$ if and only if $f_\theta \equiv f_{\theta'}$. Unlike previously considered norms, the capacity measure introduced in this paper respects all of the symmetries of \mathcal{H}_L .

Non-negative homogeneity In the example of deep linear networks, where $\sigma(x) = x$, we have the following non-Abelian Lie group symmetry acting on the network weights for all $\Lambda_1, \dots, \Lambda_L \in GL(k_1, \mathbb{R}) \times \dots \times GL(k_L, \mathbb{R})$,

$$\theta \longrightarrow (W^0 \Lambda_1, \Lambda_1^{-1} W^1 \Lambda_2 \dots, \Lambda_l^{-1} W^l \Lambda_{l+1}, \dots, \Lambda_L^{-1} W^L) .$$

It is convenient to express these transformations in terms of the Lie algebra of real-valued matrices $M_1, \dots, M_L \in M_{k_1}(\mathbb{R}) \times \dots \times M_{k_L}(\mathbb{R})$,

$$\theta \longrightarrow (W^0 e^{M_1}, e^{-M_1} W^1 e^{M_2} \dots, e^{-M_l} W^l e^{M_{l+1}}, \dots, e^{-M_L} W^L) .$$

If $\sigma(x) = \max\{0, x\}$ (deep rectified network) then the symmetry is broken to the abelian subalgebra $\mathbf{v}_1, \dots, \mathbf{v}_L \in \mathbb{R}^{k_1} \times \dots \times \mathbb{R}^{k_L}$,

$$\Lambda_1 = e^{\text{diag}(\mathbf{v}_1)}, \dots, \Lambda_L = e^{\text{diag}(\mathbf{v}_L)} .$$

Dead neurons. For certain choices of activation function the symmetry group is enhanced at some $\theta \in \Theta$. For example, if $\sigma(x) = \max\{0, x\}$ and the parameter

⁷In the statistics literature this is referred to as a non-identifiable function class.

vector θ is such that all the weights and biases feeding into some hidden unit $v \in V$ are negative, then f_θ is invariant with respect to all of the outgoing weights of v . **Permutation symmetry.** In addition to the continuous symmetries there is discrete group of permutation symmetries. In the case of a single hidden layer with k units, this discrete symmetry gives rise to $k!$ equivalent weights for a given θ . If in addition the activation function satisfies $\sigma(-x) = -\sigma(x)$ (such as tanh) then we obtain an additional degeneracy factor of 2^k .

6.2 Invariance of natural gradient

Consider the continuous-time analog of natural gradient flow,

$$(6.1) \quad d\theta_t = -\mathbf{I}(\theta_t)^{-1} \nabla_{\theta} L(\theta_t) dt,$$

where $\theta \in \mathbb{R}^p$. Consider a differentiable transformation from one parametrization to another $\theta \mapsto \xi \in \mathbb{R}^q$ denoted by $\xi(\theta) : \mathbb{R}^p \rightarrow \mathbb{R}^q$. Denote the Jacobian $\mathbf{J}_\xi(\theta) = \frac{\partial(\xi_1, \xi_2, \dots, \xi_q)}{\partial(\theta_1, \theta_2, \dots, \theta_p)} \in \mathbb{R}^{q \times p}$. Define the loss function $\tilde{L} : \mathbb{R} \rightarrow \mathbb{R}$ that satisfies

$$L(\theta) = \tilde{L}(\xi(\theta)) = \tilde{L} \circ \xi(\theta),$$

and denote $\tilde{\mathbf{I}}(\xi)$ as the Fisher Information on ξ associated with \tilde{L} . Consider also the natural gradient flow on the ξ parametrization,

$$(6.2) \quad d\xi_t = -\tilde{\mathbf{I}}(\xi_t)^{-1} \nabla_{\xi} \tilde{L}(\xi_t) dt.$$

Intuitively, one can show that the natural gradient flow is “invariant” to the specific parametrization of the problem.

LEMMA 6.1 (Parametrization invariance). *Denote $\theta \in \mathbb{R}^p$, and the differentiable transformation from one parametrization to another $\theta \mapsto \xi \in \mathbb{R}^q$ as $\xi(\theta) : \mathbb{R}^p \rightarrow \mathbb{R}^q$. Assume $\mathbf{I}(\theta)$, $\tilde{\mathbf{I}}(\xi)$ are invertible, and consider two natural gradient flows $\{\theta_t, t > 0\}$ and $\{\xi_t, t > 0\}$ defined in Eqn. (6.1) and (6.2) on θ and ξ respectively.*

(1) *Re-parametrization: if $q = p$, and assume $\mathbf{J}_\xi(\theta)$ is invertible, then natural gradient flow on the two parameterizations satisfies,*

$$\xi(\theta_t) = \xi_t, \quad \forall t,$$

if the initial locations θ_0, ξ_0 are equivalent in the sense $\xi(\theta_0) = \xi_0$.

(2) *Over-parametrization: If $q > p$ and $\xi_t = \xi(\theta_t)$ at some fixed time t , then the infinitesimal change satisfies*

$$\xi(\theta_{t+dt}) - \xi(\theta_t) = M_t(\xi_{t+dt} - \xi_t), \quad M_t \text{ has eigenvalues either } 0 \text{ or } 1$$

where $M_t = \mathbf{I}(\xi_t)^{-1/2}(I_q - U_\perp U_\perp^T)\mathbf{I}(\xi_t)^{1/2}$, and U_\perp denotes the null space of $\mathbf{I}(\xi)^{1/2}\mathbf{J}_\xi(\theta)$.

7. PROOFS

PROOF OF LEMMA 2.1. Recall the property of the activation function in (2.2). Let us prove for any $0 \leq t \leq s \leq L$, and any $l \in [k_{s+1}]$

$$(7.1) \quad \sum_{i \in [k_t], j \in [k_{t+1}]} \frac{\partial O_l^{s+1}}{\partial W_{ij}^t} W_{ij}^t = O_l^{s+1}(x).$$

We prove this statement via induction on the non-negative gap $s - t$. Starting with $s - t = 0$, we have

$$\begin{aligned}\frac{\partial O_l^{t+1}}{\partial W_{il}^t} &= \frac{\partial O_l^{t+1}}{\partial N_l^{t+1}} \frac{\partial N_l^{t+1}}{\partial W_{il}^t} = \sigma'(N_l^{t+1}(x)) O_i^t(x), \\ \frac{\partial O_l^{t+1}}{\partial W_{ij}^t} &= 0, \quad \text{for } j \neq l,\end{aligned}$$

and, therefore,

$$(7.2) \quad \sum_{i \in [k_t], j \in [k_{t+1}]} \frac{\partial O_l^{t+1}}{\partial W_{ij}^t} W_{ij}^t = \sum_{i \in [k_t]} \sigma'(N_l^{t+1}(x)) O_i^t(x) W_{il}^t = \sigma'(N_l^{t+1}(x)) N_l^{t+1}(x) = O_l^{t+1}(x).$$

This solves the base case when $s - t = 0$.

Let us assume for general $s - t \leq h$ the induction hypothesis ($h \geq 0$), and let us prove it for $s - t = h + 1$. Due to chain-rule in the back-propagation updates

$$(7.3) \quad \frac{\partial O_l^{s+1}}{\partial W_{ij}^t} = \frac{\partial O_l^{s+1}}{\partial N_l^{s+1}} \sum_{k \in [k_s]} \frac{\partial N_l^{s+1}}{\partial O_k^s} \frac{\partial O_k^s}{\partial W_{ij}^t}.$$

Using the induction on $\frac{\partial O_k^s}{\partial W_{ij}^t}$ as $(s - 1) - t = h$

$$(7.4) \quad \sum_{i \in [k_t], j \in [k_{t+1}]} \frac{\partial O_k^s}{\partial W_{ij}^t} W_{ij}^t = O_k^s(x),$$

and, therefore,

$$\begin{aligned}& \sum_{i \in [k_t], j \in [k_{t+1}]} \frac{\partial O_l^{s+1}}{\partial W_{ij}^t} W_{ij}^t \\ &= \sum_{i \in [k_t], j \in [k_{t+1}]} \frac{\partial O_l^{s+1}}{\partial N_l^{s+1}} \sum_{k \in [k_s]} \frac{\partial N_l^{s+1}}{\partial O_k^s} \frac{\partial O_k^s}{\partial W_{ij}^t} W_{ij}^t \\ &= \frac{\partial O_l^{s+1}}{\partial N_l^{s+1}} \sum_{k \in [k_s]} \frac{\partial N_l^{s+1}}{\partial O_k^s} \sum_{i \in [k_t], j \in [k_{t+1}]} \frac{\partial O_k^s}{\partial W_{ij}^t} W_{ij}^t \\ &= \sigma'(N_l^{s+1}(x)) \sum_{k \in [k_s]} W_{kl}^s O_k^s(x) = O_l^{s+1}(x).\end{aligned}$$

This completes the induction argument. In other words, we have proved for any t, s that $t \leq s$, and l is any hidden unit in layer s

$$(7.5) \quad \sum_{i, j \in \dim(W^t)} \frac{\partial O_l^{s+1}}{\partial W_{ij}^t} W_{ij}^t = O_l^{s+1}(x).$$

Remark that in the case when there are hard-coded zero weights, the proof still goes through exactly. The reason is, for the base case $s = t$,

$$\sum_{i \in [k_t], j \in [k_{t+1}]} \frac{\partial O_l^{t+1}}{\partial W_{ij}^t} W_{ij}^t = \sum_{i \in [k_t]} \sigma'(N_l^{t+1}(x)) O_i^t(x) W_{il}^t \mathbf{1}(W_{il}^t \neq 0) = \sigma'(N_l^{t+1}(x)) N_l^{t+1}(x) = O_l^{t+1}(x).$$

and for the induction step,

$$\sum_{i \in [k_t], j \in [k_{t+1}]} \frac{\partial O_l^{s+1}}{\partial W_{ij}^t} W_{ij}^t = \sigma'(N_l^{s+1}(x)) \sum_{k \in [k_s]} W_{kl}^s O_k^s(x) \mathbf{1}(W_{kl}^s \neq 0) = O_l^{s+1}(x).$$

□

PROOF OF LEMMA 4.1. The proof follows from a peeling argument from the right hand side. Recall $O^t \in \mathbb{R}^{1 \times k_t}$, one has

$$\begin{aligned} \frac{1}{(L+1)^2} \|\theta\|_{\text{fr}}^2 &= \mathbb{E} [|O^L W^L D^{L+1}|^2] \quad \text{because } |O^L W^L| \leq \|W^L\|_{\sigma} \cdot \|O^L\|_2 \\ &\leq \mathbb{E} [\|W^L\|_{\sigma}^2 \cdot \|O^L\|_2^2 \cdot |D^{L+1}(X)|^2] \\ &= \mathbb{E} [|D^{L+1}(X)|^2 \cdot \|W^L\|_{\sigma}^2 \cdot \|O^{L-1} W^{L-1} D^L\|_2^2] \\ &\leq \mathbb{E} [|D^{L+1}(X)|^2 \cdot \|W^L\|_{\sigma}^2 \cdot \|O^{L-1} W^{L-1}\|_2^2 \cdot \|D^L\|_{\sigma}^2] \\ &\leq \mathbb{E} [\|D^L\|_{\sigma}^2 |D^{L+1}(X)|^2 \cdot \|W^L\|_{\sigma}^2 \|W^{L-1}\|_{\sigma}^2 \cdot \|O^{L-1}\|_2^2] \\ &\leq \mathbb{E} [\|D^L\|_{\sigma}^2 \|D^{L+1}(X)\|_{\sigma}^2 \|O^{L-1}\|_2^2] \cdot \|W^{L-1}\|_{\sigma}^2 \|W^L\|_{\sigma}^2 \\ &\dots \quad \text{repeat the process to bound } \|O^{L-1}\|_2 \\ &\leq \mathbb{E} \left(\|X\|^2 \prod_{t=1}^{L+1} \|D^t(X)\|_{\sigma}^2 \right) \prod_{t=0}^L \|W^t\|_{\sigma}^2 = \|\theta\|_{\sigma}^2. \end{aligned}$$

□

PROOF OF LEMMA 4.2. The proof still follows a peeling argument from the right. We know that

$$\begin{aligned} \frac{1}{(L+1)^2} \|\theta\|_{\text{fr}}^2 &= \mathbb{E} [|O^L W^L D^{L+1}|^2] \\ &\leq \mathbb{E} [\|W^L\|_{p,q}^2 \cdot \|O^L\|_{p^*}^2 \cdot |D^{L+1}(X)|^2] \quad \text{use (7.6)} \\ &= \mathbb{E} [|D^{L+1}(X)|^2 \cdot \|W^L\|_{p,q}^2 \cdot \|O^{L-1} W^{L-1} D^L\|_{p^*}^2] \\ &\leq \mathbb{E} [|D^{L+1}(X)|^2 \cdot \|W^L\|_{p,q}^2 \cdot \|O^{L-1} W^{L-1}\|_q^2 \cdot \|D^L\|_{q \rightarrow p^*}^2] \\ &\leq \mathbb{E} [\|D^L\|_{q \rightarrow p^*}^2 \|D^{L+1}(X)\|_{p,q}^2 \cdot \|W^L\|_{p,q}^2 \|W^{L-1}\|_{p,q}^2 \cdot \|O^{L-1}\|_{p^*}^2] \quad \text{use (7.8)} \\ &= \mathbb{E} [\|D^L\|_{q \rightarrow p^*}^2 \|D^{L+1}(X)\|_{p,q}^2 \cdot \|O^{L-1}\|_{p^*}^2] \cdot \|W^{L-1}\|_{p,q}^2 \|W^L\|_{p,q}^2 \\ &\leq \dots \quad \text{repeat the process to bound } \|O^{L-1}\|_{p^*} \\ &\leq \mathbb{E} \left(\|X\|_{p^*}^2 \prod_{t=1}^{L+1} \|D^t(X)\|_{q \rightarrow p^*}^2 \right) \prod_{t=0}^L \|W^t\|_{p,q}^2 = \|\theta\|_{p,q}^2 \end{aligned}$$

In the proof the first inequality we use Holder's inequality

$$(7.6) \quad \langle w, v \rangle \leq \|w\|_p \|v\|_{p^*}$$

where $\frac{1}{p} + \frac{1}{p^*} = 1$. Let's prove for $v \in \mathbb{R}^n$, $M \in \mathbb{R}^{n \times m}$, we have

$$(7.7) \quad \|v^T M\|_q \leq \|v\|_{p^*} \|M\|_{p,q}.$$

Denote each column of M as $M_{.j}$, for $1 \leq j \leq m$,

$$(7.8) \quad \|v^T M\|_q = \left(\sum_{j=1}^m |v^T M_{.j}|^q \right)^{1/q} \leq \left(\sum_{j=1}^m \|v\|_{p^*}^q \|M_{.j}\|_p^q \right)^{1/q} = \|v\|_{p^*} \|M\|_{p,q}.$$

□

PROOF. Proof of Lemma 4.3 The proof is due to Holder's inequality, for any $x \in \mathbb{R}^p$

$$\begin{aligned} & \left| \sum_{i_0, i_1, \dots, i_L} x_{i_0} W_{i_0 i_1}^0 D_{i_1}^1(x) W_{i_1 i_2}^1 \cdots D_{i_L}^L(x) W_{i_L}^L D^{L+1}(x) \right| \\ & \leq \left(\sum_{i_0, i_1, \dots, i_L} |x_{i_0} D_{i_1}^1(x) \cdots D_{i_L}^L(x) D^{L+1}(x)|^{q^*} \right)^{1/q^*} \cdot \left(\sum_{i_0, i_1, \dots, i_L} |W_{i_0 i_1}^0 W_{i_1 i_2}^1 W_{i_2 i_3}^2 \cdots W_{i_L}^L|^q \right)^{1/q}. \end{aligned}$$

Therefore we have

$$\begin{aligned} \frac{1}{(L+1)^2} \|\theta\|_{\text{fr}}^2 &= \mathbb{E} \left| \sum_{i_0, i_1, \dots, i_L} X_{i_0} W_{i_0 i_1}^0 D_{i_1}^1(X) W_{i_1 i_2}^1 \cdots W_{i_L}^L D^{L+1}(X) \right|^2 \\ &\leq \left(\sum_{i_0, i_1, \dots, i_L} |W_{i_0 i_1}^0 W_{i_1 i_2}^1 W_{i_2 i_3}^2 \cdots W_{i_L}^L|^q \right)^{2/q} \cdot \mathbb{E} \left(\sum_{i_0, i_1, \dots, i_L} |X_{i_0} D_{i_1}^1(X) \cdots D_{i_L}^L(X) D^{L+1}(X)|^{q^*} \right)^{2/q^*}, \end{aligned}$$

which is

$$\frac{1}{L+1} \|\theta\|_{\text{fr}} \leq \left[\mathbb{E} \left(\sum_{i_0, i_1, \dots, i_L} |X_{i_0} \prod_{t=1}^{L+1} D_{i_t}^t(X)|^{q^*} \right)^{1/2} \cdot \left(\sum_{i_0, i_1, \dots, i_L} \prod_{t=0}^L |W_{i_t i_{t+1}}^t|^q \right)^{1/q} \right] = \|\pi(\theta)\|_q.$$

□

PROOF OF LEMMA 4.4. The proof follows from the recursive use of the inequality,

$$\|M\|_{p \rightarrow q} \|v\|_p \geq \|v^T M\|_q.$$

We have

$$\begin{aligned} \|\theta\|_{\text{fr}}^2 &= \mathbb{E} [|O^L W^L D^{L+1}|^2] \\ &\leq \mathbb{E} [\|W^L\|_{p \rightarrow q}^2 \cdot \|O^L\|_p^2 \cdot |D^{L+1}(X)|^2] \\ &\leq \mathbb{E} [|D^{L+1}(X)|^2 \cdot \|W^L\|_{p \rightarrow q}^2 \cdot \|O^{L-1} W^{L-1} D^L\|_p^2] \\ &\leq \mathbb{E} [|D^{L+1}(X)|^2 \cdot \|W^L\|_{p \rightarrow q}^2 \cdot \|O^{L-1} W^{L-1}\|_q^2 \cdot \|D^L\|_{q \rightarrow p}^2] \\ &\leq \mathbb{E} [\|D^L\|_{q \rightarrow p}^2 \|D^{L+1}(X)\|_{q \rightarrow p}^2 \cdot \|W^L\|_{p \rightarrow q}^2 \|W^{L-1}\|_{p \rightarrow q}^2 \cdot \|O^{L-1}\|_p^2] \\ &\leq \dots \text{ repeat the process to bound } \|O^{L-1}\|_p \\ &\leq \mathbb{E} \left(\|X\|_p^2 \prod_{t=1}^{L+1} \|D^t(X)\|_{q \rightarrow p}^2 \right) \prod_{t=0}^L \|W^t\|_{p \rightarrow q}^2 = \|\theta\|_{p \rightarrow q}^2, \end{aligned}$$

where third to last line is because $D^{L+1}(X) \in \mathbb{R}^1$, $|D^{L+1}(X)| = \|D^{L+1}(X)\|_{q \rightarrow p}$.

□

PROOF OF LEMMA 4.5. The proof follows from a different strategy of peeling the terms from the right hand side, as follows,

$$\begin{aligned}
\|\theta\|_{\text{fr}}^2 &= \mathbb{E} \left[|O^L W^L D^{L+1}|^2 \right] \\
&\leq \mathbb{E} \left[\|W^L\|_{p_L \rightarrow p_{L+1}}^2 \cdot \|O^L\|_{p_L}^2 \cdot |D^{L+1}(X)|^2 \right] \\
&\leq \mathbb{E} \left[|D^{L+1}(X)|^2 \cdot \|W^L\|_{p_L \rightarrow p_{L+1}}^2 \cdot \|O^{L-1} W^{L-1} D^L\|_{p_L}^2 \right] \\
&\leq \mathbb{E} \left[|D^{L+1}(X)|^2 \cdot \|W^L\|_{p_L \rightarrow p_{L+1}}^2 \cdot \|O^{L-1} W^{L-1}\|_{p_L} \|D^L\|_{p_L \rightarrow p_L}^2 \right] \\
&\leq \mathbb{E} \left[\|D^L\|_{p_L \rightarrow p_L}^2 |D^{L+1}(X)|^2 \cdot \|W^L\|_{p_L \rightarrow p_{L+1}}^2 \|W^{L-1}\|_{p_{L-1} \rightarrow p_L}^2 \cdot \|O^{L-1}\|_{p_{L-1}}^2 \right] \\
&\leq \mathbb{E} \left(\|X\|_{p_0}^2 \prod_{t=1}^{L+1} \|D^t(X)\|_{p_t \rightarrow p_t}^2 \right) \prod_{t=0}^L \|W^t\|_{p_t \rightarrow p_{t+1}}^2 = \|\theta\|_P^2.
\end{aligned}$$

□

PROOF OF LEMMA 4.6.

$$\begin{aligned}
\frac{d}{dr} \|r\theta\|_{\text{fr}}^2 &= \mathbb{E} [2\theta \nabla f_{r\theta}(X) f_{r\theta}(X)] \\
&= \mathbb{E} \left[\frac{2(L+1)}{r} f_{r\theta}(X) f_{r\theta}(X) \right] \quad \text{use Lemma 2.1} \\
&= \frac{2(L+1)}{r} \|r\theta\|_{\text{fr}}^2
\end{aligned}$$

The last claim can be proved through solving the simple ODE. □

PROOF OF LEMMA 4.7. Let us first construct $\theta' \in \Theta_{L+1}$ that realizes $\lambda f_{\theta_1} + (1-\lambda)f_{\theta_2}$. The idea is very simple: we put θ_1 and θ_2 networks side-by-side, then construct an additional output layer with weights $\lambda, 1-\lambda$ on the output of f_{θ_1} and f_{θ_2} , and the final output layer is passed through $\sigma(x) = x$. One can easily see that our key Lemma 2.1 still holds for this network: the interaction weights between f_{θ_1} and f_{θ_2} are always hard-coded as 0. Therefore we have constructed a $\theta' \in \Theta_{L+1}$ that realizes $\lambda f_{\theta_1} + (1-\lambda)f_{\theta_2}$.

Now recall that

$$\begin{aligned}
\frac{1}{L+2} \|\theta'\|_{\text{fr}} &= (\mathbb{E} f_{\theta'}^2)^{1/2} \\
&= (\mathbb{E} (\lambda f_{\theta_1} + (1-\lambda)f_{\theta_2})^2)^{1/2} \\
&\leq \lambda (\mathbb{E} f_{\theta_1}^2)^{1/2} + (1-\lambda) (\mathbb{E} f_{\theta_2}^2)^{1/2} \leq 1
\end{aligned}$$

because $\mathbb{E}[f_{\theta_1} f_{\theta_2}] \leq (\mathbb{E} f_{\theta_1}^2)^{1/2} (\mathbb{E} f_{\theta_2}^2)^{1/2}$. □

PROOF OF THEOREM 4.1. Due to Eqn. (3.5), one has

$$\begin{aligned}
\frac{1}{(L+1)^2} \|\theta\|_{\text{fr}}^2 &= \mathbb{E} [v(\theta, X)^T X X^T v(\theta, X)] \\
&= v(\theta)^T \mathbb{E} [X X^T] v(\theta)
\end{aligned}$$

because in the linear case $v(\theta, X) = W^0 D^1(x) W^1 D^2(x) \cdots D^L(x) W^L D^{L+1}(x) = \prod_{t=0}^L W^t =: v(\theta) \in \mathbb{R}^p$. Therefore

$$\begin{aligned}
\mathcal{R}_N(B_{\text{fr}}(\gamma)) &= \mathbb{E} \sup_{\epsilon \in B_{\text{fr}}(\gamma)} \frac{1}{N} \sum_{i=1}^N \epsilon_i f_{\theta}(X_i) \\
&= \mathbb{E} \sup_{\epsilon \in B_{\text{fr}}(\gamma)} \frac{1}{N} \sum_{i=1}^N \epsilon_i X_i^T v(\theta) \\
&= \mathbb{E} \sup_{\epsilon \in B_{\text{fr}}(\gamma)} \frac{1}{N} \left\langle \sum_{i=1}^N \epsilon_i X_i, v(\theta) \right\rangle \\
&\leq \gamma \mathbb{E} \frac{1}{N} \left\| \sum_{i=1}^N \epsilon_i X_i \right\|_{[\mathbb{E}(X X^T)]^{-1}} \\
&\leq \gamma \frac{1}{\sqrt{N}} \sqrt{\frac{1}{N} \mathbb{E} \left\| \sum_{i=1}^N \epsilon_i X_i \right\|_{[\mathbb{E}(X X^T)]^{-1}}^2} \\
&= \gamma \frac{1}{\sqrt{N}} \sqrt{\left\langle \frac{1}{N} \sum_{i=1}^N X_i X_i^T, [\mathbb{E}(X X^T)]^{-1} \right\rangle}.
\end{aligned}$$

Therefore

$$\mathbb{E} \mathcal{R}_N(B_{\text{fr}}(\gamma)) \leq \gamma \frac{1}{\sqrt{N}} \sqrt{\mathbb{E} \left\langle \frac{1}{N} \sum_{i=1}^N X_i X_i^T, [\mathbb{E}(X X^T)]^{-1} \right\rangle} = \gamma \sqrt{\frac{p}{N}}.$$

□

PROOF OF LEMMA 6.1. From basic calculus, one has

$$\begin{aligned}
\nabla_{\theta} L(\theta) &= \mathbf{J}_{\xi}(\theta)^T \nabla_{\xi} \tilde{L}(\xi) \\
\mathbf{I}(\theta) &= \mathbf{J}_{\xi}(\theta)^T \tilde{\mathbf{I}}(\xi) \mathbf{J}_{\xi}(\theta)
\end{aligned}$$

Therefore, plugging in the above expression into the natural gradient flow in θ

$$\begin{aligned}
d\theta_t &= -\mathbf{I}(\theta_t)^{-1} \nabla_{\theta} L(\theta_t) dt \\
&= -[\mathbf{J}_{\xi}(\theta_t)^T \tilde{\mathbf{I}}(\xi(\theta_t)) \mathbf{J}_{\xi}(\theta_t)]^{-1} \mathbf{J}_{\xi}(\theta_t)^T \nabla_{\xi} \tilde{L}(\xi(\theta_t)) dt.
\end{aligned}$$

In the re-parametrization case, $\mathbf{J}_{\xi}(\theta)$ is invertible, and assuming $\xi_t = \xi(\theta_t)$,

$$\begin{aligned}
d\theta_t &= -[\mathbf{J}_{\xi}(\theta_t)^T \tilde{\mathbf{I}}(\xi(\theta_t)) \mathbf{J}_{\xi}(\theta_t)]^{-1} \mathbf{J}_{\xi}(\theta_t)^T \nabla_{\xi} \tilde{L}(\xi(\theta_t)) dt \\
&= -\mathbf{J}_{\xi}(\theta_t)^{-1} \tilde{\mathbf{I}}(\xi(\theta_t))^{-1} \nabla_{\xi} \tilde{L}(\xi(\theta_t)) dt \\
\mathbf{J}_{\xi}(\theta_t) d\theta_t &= -\tilde{\mathbf{I}}(\xi(\theta_t))^{-1} \nabla_{\xi} \tilde{L}(\xi(\theta_t)) dt \\
d\xi(\theta_t) &= -\tilde{\mathbf{I}}(\xi(\theta_t))^{-1} \nabla_{\xi} \tilde{L}(\xi(\theta_t)) dt = -\tilde{\mathbf{I}}(\xi_t)^{-1} \nabla_{\xi} \tilde{L}(\xi_t) dt.
\end{aligned}$$

What we have shown is that under $\xi_t = \xi(\theta_t)$, $\xi(\theta_{t+dt}) = \xi_{t+dt}$. Therefore, if $\xi_0 = \xi(\theta_0)$, we have that $\xi_t = \xi(\theta_t)$.

In the over-parametrization case, $\mathbf{J}_\xi(\theta) \in \mathbb{R}^{q \times p}$ is a non-square matrix. For simplicity of derivation, abbreviate $B := \mathbf{J}_\xi(\theta) \in \mathbb{R}^{q \times p}$. We have

$$\begin{aligned} d\theta_t &= \theta_{t+dt} - \theta_t = -\mathbf{I}(\theta_t)^{-1} \nabla_\theta L(\theta_t) dt \\ &= -[B^T \tilde{\mathbf{I}}(\xi) B]^{-1} B^T \nabla_\xi \tilde{L}(\xi(\theta_t)) dt \\ B(\theta_{t+dt} - \theta_t) &= -B [B^T \tilde{\mathbf{I}}(\xi) B]^{-1} B^T \tilde{L}(\xi(\theta_t)) dt. \end{aligned}$$

Via the Sherman-Morrison-Woodbury formula

$$\left[I_q + \frac{1}{\epsilon} \tilde{\mathbf{I}}(\xi)^{1/2} B B^T \tilde{\mathbf{I}}(\xi)^{1/2} \right]^{-1} = I_q - \tilde{\mathbf{I}}(\xi)^{1/2} B (\epsilon I_p + B^T \tilde{\mathbf{I}}(\xi) B)^{-1} B^T \tilde{\mathbf{I}}(\xi)^{1/2}$$

Denoting $\tilde{\mathbf{I}}(\xi)^{1/2} B B^T \tilde{\mathbf{I}}(\xi)^{1/2} = U \Lambda U^T$, we have that $\text{rank}(\Lambda) \leq p < q$. Therefore, the LHS as

$$\begin{aligned} \left[I_q + \frac{1}{\epsilon} \tilde{\mathbf{I}}(\xi)^{1/2} B B^T \tilde{\mathbf{I}}(\xi)^{1/2} \right]^{-1} &= U \left[I_q + \frac{1}{\epsilon} \Lambda \right]^{-1} U^T \\ \lim_{\epsilon \rightarrow 0} \left[I_q + \frac{1}{\epsilon} \tilde{\mathbf{I}}(\xi)^{1/2} B B^T \tilde{\mathbf{I}}(\xi)^{1/2} \right]^{-1} &= U_\perp U_\perp^T \end{aligned}$$

where U_\perp corresponding to the space associated with zero eigenvalue of $\tilde{\mathbf{I}}(\xi)^{1/2} B B^T \tilde{\mathbf{I}}(\xi)^{1/2}$. Therefore taking $\epsilon \rightarrow 0$, we have

$$\begin{aligned} \lim_{\epsilon \rightarrow 0} \left[I_q + \frac{1}{\epsilon} \tilde{\mathbf{I}}(\xi)^{1/2} B B^T \tilde{\mathbf{I}}(\xi)^{1/2} \right]^{-1} &= \lim_{\epsilon \rightarrow 0} I_q - \tilde{\mathbf{I}}(\xi)^{1/2} B (\epsilon I_p + B^T \tilde{\mathbf{I}}(\xi) B)^{-1} B^T \tilde{\mathbf{I}}(\xi)^{1/2} \\ \tilde{\mathbf{I}}(\xi)^{-1/2} U_\perp U_\perp^T \tilde{\mathbf{I}}(\xi)^{-1/2} &= \tilde{\mathbf{I}}(\xi)^{-1} - B (B^T \tilde{\mathbf{I}}(\xi) B)^{-1} B^T \end{aligned}$$

where only the last step uses the fact $\tilde{\mathbf{I}}(\xi)$ is invertible. Therefore

$$\begin{aligned} \xi(\theta_{t+dt}) - \xi(\theta_t) &= B(\theta_{t+dt} - \theta_t) \\ &= -B [B^T \tilde{\mathbf{I}}(\xi) B]^{-1} B^T \nabla_\xi \tilde{L}(\xi) dt \\ &= -\eta \mathbf{I}(\xi)^{-1/2} (I_d - U_\perp U_\perp^T) \mathbf{I}(\xi)^{-1/2} \nabla_\xi \tilde{L}(\xi) dt \\ &= \mathbf{I}(\xi)^{-1/2} (I_d - U_\perp U_\perp^T) \mathbf{I}(\xi)^{1/2} \left\{ \mathbf{I}(\xi)^{-1} \nabla_\xi \tilde{L}(\xi) dt \right\} \\ &= M_t (\xi_{t+dt} - \xi_t). \end{aligned}$$

The above claim asserts that in the over-parametrized setting, running natural gradient in the over-parametrized is nearly ‘‘invariant’’ in the following sense: if $\xi(\theta_t) = \xi_t$, then

$$\begin{aligned} \xi(\theta_{t+dt}) - \xi(\theta_t) &= M_t (\xi_{t+dt} - \xi_t) \\ M_t &= \mathbf{I}(\xi_t)^{-1/2} (I_q - U_\perp U_\perp^T) \mathbf{I}(\xi_t)^{1/2} \end{aligned}$$

and we know M_t has eigenvalue either 1 or 0. In the case when $p = q$ and $\mathbf{J}_\xi(\theta)$ has full rank, it holds that $M_t = I$ is the identity matrix, reducing the problem to the re-parametrized case. \square

REFERENCES

- Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2): 251–276, 1998.
- Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Cambridge university press, 1999.
- Peter Bartlett, Dylan J Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1706.08498*, 2017.
- Martin Bauer, Martins Bruveris, and Peter W Michor. Uniqueness of the fisher–rao metric on the space of smooth densities. *Bulletin of the London Mathematical Society*, 48(3):499–506, 2016.
- Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. *arXiv preprint arXiv:1703.04933*, 2017.
- Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.
- Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, pages 1–50, 2002.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- Anders Krogh and John A Hertz. A simple weight decay can improve generalization. In *Advances in neural information processing systems*, pages 950–957, 1992.
- James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International Conference on Machine Learning*, pages 2408–2417, 2015.
- Behnam Neyshabur, Ruslan R Salakhutdinov, and Nati Srebro. Path-sgd: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2422–2430, 2015a.
- Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pages 1376–1401, 2015b.
- Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring generalization in deep learning. *arXiv preprint arXiv:1706.08947*, 2017.
- Shai Shalev-Shwartz, Ohad Shamir, and Shaked Shammah. Failures of deep learning. *arXiv preprint arXiv:1703.07950*, 2017.
- Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.