INFORMATION GEOMETRY OF ORTHOGONAL INITIALIZATIONS AND TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently mean field theory has been successfully used to analyze properties of wide, random neural networks. It gave rise to a prescriptive theory for initializing feed-forward neural networks with orthogonal weights, which ensures that both the forward propagated activations and the backpropagated gradients are near ℓ_2 isometries and as a consequence training is orders of magnitude faster. Despite strong empirical performance, the mechanisms by which critical initializations confer an advantage in the optimization of deep neural networks are poorly understood. Here we show a novel connection between the maximum curvature of the optimization landscape (gradient smoothness) as measured by the Fisher information matrix (FIM) and the spectral radius of the input-output Jacobian, which partially explains why more isometric networks can train much faster. Furthermore, given that orthogonal weights are necessary to ensure that gradient norms are approximately preserved at initialization, we experimentally investigate the benefits of maintaining orthogonality throughout training, and we conclude that manifold optimization of weights performs well regardless of the smoothness of the gradients. Moreover, we observe a surprising yet robust behavior of highly isometric initializations — even though such networks have a lower FIM condition number *at initialization*, and therefore by analogy to convex functions should be easier to optimize, experimentally they prove to be much harder to train with stochastic gradient descent. We propose an explanation for this phenomenon by exploting connections between Fisher geometry and the recently introduced Neural Tangent Kernel.

1 INTRODUCTION

Deep neural networks (DNN) have shown tremendous success in computer vision problems, speech recognition, amortized probabilistic inference, and the modelling of neural data. Despite their performance, DNNs face obstacles in their practical application, which stem from both the excessive computational cost of running gradient descent for a large number of epochs, as well as the inherent brittleness of gradient descent applied to very deep models. A number of heuristic approaches such as batch normalization, weight normalization and residual connections (He et al., 2015; Ioffe & Szegedy, 2015; Salimans & Kingma, 2016) have emerged in an attempt to address these trainability issues.

Recently mean field theory has been successful in developing a more principled analysis of gradients of neural networks, and has become the basis for a new random initialization principle. The mean field approach postulates that in the limit of infinitely wide random weight matrices, the distribution of pre-activations converges weakly to a Gaussian. Using this approach, a series of works proposed to initialize the networks in such a way that for each layer the input-output Jacobian has mean singular values of 1 (Schoenholz et al., 2016). This requirement was further strengthened to suggest that the spectrum of singular values of the input-output Jacobian should concentrate on 1, and it was shown that this can only be achieved with random weight matrices.

Under these conditions the backpropagated gradients are bounded in ℓ_2 norm (Pennington et al., 2017) irrespective of depth, i.e., they neither vanish nor explode. It was shown experimentally in (Pennington et al., 2017; Xiao et al., 2018b; Chen et al., 2018) that networks with these *critical* initial conditions train orders of magnitude faster than networks with arbitrary initializations. The empirical success invites questions from an optimization perspective on how the spectrum of the hidden layer input-output Jacobian relates to notions of curvature of the parameters space, and subsequently to convergence rate. The largest effective (initial) step size η_0 for stochastic gradient descent is inversely

proportional to the local gradient smoothness M (Bottou et al., 2016; Boyd & Vandenberghe, 2004). Intuitively, the gradient step can be at most as large as the fastest change in the parameter landscape. Recent attempts have been made to analyze the mean field geometry of the optimization using the Fisher information matrix (FIM) (Amari et al., 2018; Karakida et al., 2018). The theoretical and practical appeal of measuring curvature with the FIM is due to among other reasons the fact that the FIM is necessarily positive semidefinite even for non-convex objectives, and due to it its intimate relationship with the Hessian matrix. (Karakida et al., 2018) derived an upper bound on the maximum eigenvalue, however this bound is not satisfactory since it is agnostic of the entire spectrum of singular values and therefore cannot differentiate between Gaussian and orthogonal weight initalizations.

In this paper, we develop a new bound on the parameter curvature M given the maximum eigenvalue of the Fisher information $\lambda_{max}(\bar{\mathbf{G}})$ which holds both Gaussian and orthogonal. We show that this quantity under certain conditions is proportional to the maximum singular value of the input-output Jacobian. We use this result to probe different orthogonal initializations, and observe that, broadly speaking, networks with a smaller initial curvature train faster and generalize better, as expected. However, consistently with a previous report (Pennington et al., 2018), we also observe highly isometric networks perform worse despite having a slowly varying loss landscape (i.e. small initial $\lambda_{max}(\bar{\mathbf{G}})$). We propose a theoretical explanation for this phenomenon using the connections between the FIM and the recently introduced Neural Tangent Kernel (Jacot et al., 2018; Lee et al., 2019). Given that the smallest and largest eigenvalues have an approximately inverse relationship (Karakida et al., 2018), we propose an explanation that the long term optimization behavior is mostly controlled by the smallest eigenvalue m and therefore surprisingly there is a *sweetspot with the condition number being* $\frac{m}{M} > 1$.

We then investigate whether constraining the spectrum of the Jacobian matrix of each layer affects optimization rate. We do so by training networks using Riemannian optimization to constrain their weights to be orthogonal, or nearly orthogonal and we find that manifold constrained networks are insensitive to the maximal curvature at the beginning of training unlike the unconstrained gradient descent (hereafter "Euclidean"). In particular, we observe that the advantage conferred by optimizing over manifolds cannot be explained by the improvement of the gradient smoothness as measured by $\lambda_{max}(\bar{\mathbf{G}})$.

Importantly, we observe that contary to (Bansal et al., 2018)'s results Euclidean networks with a carefully designed initialization reduce the test misclassification error at approximately the same rate as their manifold constrained counterparts, and overall attain a higher accuracy.

2 Background

2.1 Formal Description of the Network

Following (Pennington et al., 2017; 2018; Schoenholz et al., 2016), we consider a feed-forward, fully connected neural network with L hidden layers. Each layer $l \in \{1, \ldots, L\}$ is given as a recursion of the form

$$\mathbf{x}^l = \phi(\mathbf{h}^l), \quad \mathbf{h}^l = \mathbf{W}^l \mathbf{x}^{l-1} + \mathbf{b}^l \tag{1}$$

where $\mathbf{x}^l$ are the activations, $\mathbf{h}^l$ are the pre-activations, $\mathbf{W}^l \in \mathbb{R}^{N^l \times N^{l-1}}$ are the weight matrices, $\mathbf{b}^l$ are the bias vectors, and $\phi(\cdot)$ is the activation function. The input is denoted as $\mathbf{x}^0$. The output layer of the network computes $\hat{\mathbf{y}} = g^{-1}(\mathbf{h}^g)$ where g is the link function and $\mathbf{h}^g = \mathbf{W}^g x^L + \mathbf{b}^g$.
The hidden layer input-output Jacobian matrix $\mathbf{J}_{\mathbf{x}^0}^{\mathbf{x}^L}$ is,

$$\mathbf{J}_{\mathbf{x}^0}^{\mathbf{x}^L} \triangleq \frac{\partial \mathbf{x}^L}{\partial \mathbf{x}^0} = \prod_{l=1}^{L} \mathbf{D}^l \mathbf{W}^l \tag{2}$$

where $\mathbf{D}^l$ is a diagonal matrix with entries $\mathbf{D}^l_{i,i} = \phi'(\mathbf{h}^l_i)$. As pointed out in (Pennington et al., 2017; Schoenholz et al., 2016), the conditioning of the Jacobian matrix affects the conditioning of the back-propagated gradients for all layers.

2.2 CRITICAL INITIALIZATIONS

Extending the classic result on the Gaussian process limit for wide layer width obtained by (Neal, 1996), recent work (Matthews et al., 2018; Lee et al., 2017) has shown that for deep untrained networks with elements of their weight matrices $\mathbf{W}_{i,j}$ drawn from a Gaussian distribution $\mathcal{N}(0, \frac{\sigma_{\mathbf{W}}^2}{N^l})$ the empirical distribution of the pre-activations $\mathbf{h}^l$ converges weakly to a Gaussian distribution $\mathcal{N}(0, q^l\mathbf{I})$ for each layer l in the limit of the width $N \to \infty$. Similarly, it has been postulated that random orthogonal matrices scaled by $\sigma_{\mathbf{W}}$ give rise to the same limit. Under this mean-field condition, the variance of the pre-activation distribution q^l is recursively given by,

$$q^l = \sigma_{\mathbf{W}}^2 \int \phi\left(\sqrt{q^{l-1}}h\right) \mathrm{d}\mu(h) + \sigma_{\mathbf{b}}^2 \tag{3}$$

where $\mu(h)$ denotes the standard Gaussian measure $\int \frac{\mathrm{d}h}{\sqrt{2\pi}} \exp\left(\frac{-h^2}{2}\right)$ and $\sigma_{\mathbf{b}}^2$ denotes the variance of the Gaussian distributed biases (Schoenholz et al., 2016). The variance of the first layer pre-activations q^1 depends on ℓ_2 norm squared of inputs $q^1 = \frac{\sigma_{\mathbf{W}}^2}{N^1}\left\|p\left(\mathbf{x}^0\right)\right\|_2^2 + \sigma_{\mathbf{b}}^2$. The recursion defined in equation 3 has a fixed point

$$q^* = \sigma_{\mathbf{W}}^2 \int \phi\left(\sqrt{q^*}h\right) \mathrm{d}\mu(h) + \sigma_{\mathbf{b}}^2 \tag{4}$$

which can be satisfied for all layers by appropriately choosing $\sigma_{\mathbf{W}}, \sigma_{\mathbf{b}}$ and scaling the input $\mathbf{x}^0$. To permit the mean field analysis of backpropagated signals, the authors (Schoenholz et al., 2016; Pennington et al., 2017; 2018; Karakida et al., 2018) further assume the propagated activations and back propagated gradients to be independent. Specifically,

Assumption 1. *[Mean field assumptions]*
(i) $\lim_{N\to\infty} \mathbf{h} \xrightarrow{d} \mathcal{N}(0, q^*)$
(ii) $\lim_{N\to\infty} \mathrm{Cov}\left[\mathbf{J}^g_{\mathbf{x}^{i+1}}\mathbf{h}^i, \mathbf{J}^g_{\mathbf{x}^{j+1}}\mathbf{h}^j\right] = 0 \quad$ *for all* $i \neq j$

Under this assumption, the authors (Schoenholz et al., 2016; Pennington et al., 2017) analyze distributions of singular values of Jacobian matrices between different layers in terms of a small number of parameters, with the calculations of the backpropagated signals proceeding in a selfsame fashion as calculations for the forward propagation of activations. The corollaries of Assumption 1 and condition in equation 4 is that $\phi'(\mathbf{h}^l)$ for $1 \leq l \leq L$ are i.i.d. In order to ensure that $\mathbf{J}^{\mathbf{x}^L}_{\mathbf{x}^0}$ is well conditioned, (Pennington et al., 2017) require that in addition to the variance of pre-activation being constant for all layers, two additional constraints be met. Firstly, they require that the mean square singular value of $\mathbf{DW}$ for each layer have a certain value in expectation.

$$\chi = \frac{1}{N}\mathbb{E}\left[\,\mathrm{Tr}\left[(\mathbf{DW})^\top\mathbf{DW}\right]\right] = \sigma_{\mathbf{W}}^2 \int \left[\phi'(\sqrt{q^*}h)\right]^2 \mathrm{d}\mu(h) \tag{5}$$

Given that the mean squared singular value of the Jacobian matrix $\mathbf{J}^{\mathbf{x}^L}_{\mathbf{x}^0}$ is $(\chi)^L$, setting $\chi = 1$ corresponds to a critical initialization where the gradients are asymptotically stable as $L \to \infty$. Secondly, they require that the maximal squared singular value s^2_{max} of the Jacobian $\mathbf{J}^{\mathbf{x}^L}_{\mathbf{x}^0}$ be bounded. (Pennington et al., 2017) showed that for weights with Gaussian distributed elements, the maximal singular value increases linearly in depth even if the network is initialized with $\chi = 1$. Fortunately, for orthogonal weights, the maximal singular value s_{max} is bounded even as $L \to \infty$ (Pennington et al., 2018).

3 THEORETICAL RESULTS: RELATING THE SPECTRA OF JACOBIAN AND FISHER INFORMATION MATRICES

To better understand the geometry of the optimization landscape, we wish to put a Lipschitz bound on the gradient, which in turn gives an upper bound on the largest step size of any first order optimization algorithm. For a general objective function f, the condition is equivalent to

$$\|\nabla f(x) - \nabla f(x')\|_2 \leq M\|x - x'\|_2 \quad \text{for all} \quad x, x' \subset \mathcal{S} \subseteq \mathbb{R}^d$$

The Lipschitz constant ensures that the gradient doesn't change arbitrarily fast with respect to $x, \quad x'$, and therefore ∇f defines a descent direction for the objective over a distance M. In general estimating

the Lipschitz constant is NP-hard Kunstner et al. (2019), therefore we seek to find local measures of curvature along the optimization trajectory. As we will show below the approximate gradient smoothness is tractable for randomly initialized neural networks.

The analytical study of Hessians of random neural networks started with Pennington & Bahri (2017), but was limited to shallow architectures. Subsequent work Amari et al. (2018); Karakida et al. (2018) on second order geometry of random networks shares much of the spirit of the current work, in that it proposes to replace the possibly indefinite Hessian with the related Fisher information matrix as a measure of curvature. The Fisher information matrix plays a fundamental role in the geometry of probabilistic models, under the Kullback-Leibler divergence loss. However,because of its relation to the Hessian, it can also be seen as defining an approximate curvature matrix for second order optimization. Recall that the FIM is defined as

Definition. *Fisher Information Matrix*

$$\mathbf{G} \triangleq \mathbb{E}_{p_\theta(\mathbf{y}|\mathbf{x}^0)} \left[\mathbb{E}_{p(\mathbf{x}^0)} \left[\nabla_\theta \log p_\theta(\mathbf{y}|\mathbf{x}^0) \nabla_\theta \log p_\theta(\mathbf{y}|\mathbf{x}^0)^\top\right]\right] \tag{6}$$

$$= \mathbb{E}_{p_\theta(\mathbf{y}|\mathbf{x}^0)} \left[\mathbb{E}_{p(\mathbf{x}^0)} \left[\mathbf{J}_\theta^{h^g\top} \nabla^2_{h^g} \mathcal{L}\, \mathbf{J}_\theta^{h^g}\right]\right] = \mathbb{E}_{p_\theta(\mathbf{y}|\mathbf{x}^0)} \left[\mathbb{E}_{p(\mathbf{x}^0)} \left[\mathbf{H} - \sum_k \nabla_{\mathbf{x}^g} \mathcal{L}_k \, \nabla^2_\theta h^g_k\right]\right] \tag{7}$$

where $\mathcal{L}$ denotes the loss and $\mathbf{h}^g$ is the output layer. The relation between the Hessian and Fisher Information matrices is apparent from equation 7, showing that the Hessian $\mathbf{H}$ is a quadratic form of the Jacobian matrices plus the possibly indefinite matrix of second derivatives with respect to parameters. Our goal is to express the gradient smoothness using the results of the previous section. Given equation 7 we can derive an analytical approximation to the Lipschitz bound using the results from the previous section; i.e. we will express the expected maximum eigenvalue of the random Fisher information matrix in terms of the expected maximum singular value of the Jacobian $\mathbf{J}_{\mathbf{h}^1}^{\mathbf{h}^L}$. To do so, let us consider the output of a multilayer perceptron as defining a conditional probability distribution $p_\theta(\mathbf{y}|\mathbf{x}^0)$, where $\Theta = \{\mathrm{vec}(\mathbf{W}^1), \ldots, \mathrm{vec}(\mathbf{W}^L), \mathbf{b}^1, \ldots, \mathbf{b}^L\}$ is the set of all hidden layer parameters, and θ is the column vector containing the concatenation of all the parameters in Θ. As observed by (Martens & Grosse, 2015) the Fisher of a multilayer network naturally has a block structure, with each corresponding to the weights and biases of each layer. These blocks with respect to parameter vectors $a, b \in \Theta$ can further be expressed as

$$\bar{\mathbf{G}}_{a,b} = \mathbf{J}_a^{\mathbf{h}^g\top} \mathbf{H}_g \mathbf{J}_b^{\mathbf{h}^g} \tag{8}$$

where the final layer Hessian $\mathbf{H}_g$ is defined as $\nabla^2_{\mathbf{h}^g} \log p_\theta(\mathbf{y}|\mathbf{x}^0)$. We can re-express the outer product of the score function $\nabla_{\mathbf{h}^g} \log p_\theta(\mathbf{y}|\mathbf{x}^0)$ as the second derivative of the log-likelihood (see equation 6), provided it satisfies certain technical conditions. What is important for us is that all canonical link function for generalized linear models, like the softmax function and the identity function allow this re-writing, and that this re-writing allows us drop the conditional expectation with respect to $p_\theta\left(\mathbf{y}|\mathbf{x}^0\right)$.

The Jacobians in equation 8 can be computed iteratively. Importantly the Jacobian from the output layer to the a-th parameter block is just the product of diagonal activations and weight matrices multiplied by the Jacobian from the α-th layer to the a-th parameter. We define these matrices of partial derivatives of the α-th layer pre-activations with respect to the layer specific parameters separately for $\mathbf{W}^\alpha$ and $\mathbf{b}^\alpha$ as:

$$\mathbf{J}_a^{\mathbf{h}^\alpha} = \mathbf{x}^{\alpha-1\top} \otimes \mathbf{I} \qquad \text{for } a = \mathrm{vec}(\mathbf{W}^\alpha) \tag{9}$$

$$\mathbf{J}_a^{\mathbf{h}^\alpha} = \mathbf{I} \qquad \text{for } a = \mathbf{b}^\alpha \tag{10}$$

Under the infinitesmally weak correlation assumption (see Assumption 1), we can further simplify the expression for the blocks of the Fisher information matrix equation 8.

Lemma 1. *The expected blocks with respect to weight matrices for all layers $\alpha, \beta \neq 1$ are*

$$\bar{\mathbf{G}}_{\mathrm{vec}(\mathbf{W}^\alpha),\mathrm{vec}(\mathbf{W}^\beta)} = \mathbb{E}\left[\mathbf{x}^{\alpha-1}\mathbf{x}^{\beta-1\top}\right] \otimes \mathbb{E}\left[\mathbf{J}_{\mathbf{h}^\alpha}^{\mathbf{h}^g\top} \mathbf{H}_g \mathbf{J}_{\mathbf{h}^\beta}^{\mathbf{h}}\right] \tag{11}$$

Lemma 2. *The expected blocks with respect to a weight matrix $\mathbf{W}^\alpha$ and a bias vector $\mathbf{b}^\beta$ are*

$$\bar{\mathbf{G}}_{\mathrm{vec}(\mathbf{W}^\alpha),\mathbf{b}^\beta} = \mathbb{E}\left[\mathbf{x}^{\alpha-1\top} \otimes \mathbf{I}\right] \mathbb{E}\left[\mathbf{J}_{\mathbf{h}^\alpha}^{\mathbf{h}^g\top} \mathbf{H}_g \mathbf{J}_{\mathbf{h}^\beta}^{\mathbf{h}^g}\right] \tag{12}$$

The crucial observation here is that in the mean-field limit the cross-covariance for the activations $\mathbf{x}^{\alpha-1}, \quad \mathbf{x}^{\beta-1}$ is either zero or rank 1 for activations in different layers. The case when both activations are in the same layer is trivially taken care of by our mean-field assumptions — the covariance is proportional to the identity the identity plus potentially a rank one matrix. These rank 1 terms come from the fact that the expectation of $\mathbb{E}\left[\phi(h)^2\right]$ under a Gaussian distribution need not be zero.

Now, leveraging lemmas 1 and 2 we derive a block diagonal approximation which in turn allows us to bound the maximum eigenvalue $\lambda_{max}(\bar{\mathbf{G}})$. In doing so we will use a corollary of the block Gershgorin theorem.

Proposition 1 ((informal) Block Gershgorin theorem)**.** *The maximum eigenvalue $\lambda_{max}(\bar{\mathbf{G}})$ is contained in a union of disks centered around the maximal eigenvalue of each diagonal block with radia equal to the sum of the singular values of the off-diagonal terms.*

For a rigorous statement of the theorem see Appendix A.2. It is noteworthy that block-diagonal approximations have been crucial to the application of Fisher Information matrices as preconditioners in stochastic second order methods (Botev et al., 2017; Martens & Grosse, 2015). These methods were motivated by practical performance, in their choice of number of diagonal blocks used for preconditioning. Under the mean-field assumptions we are able to show computable bounds on the error in approximating the eigen spectrum of the Fisher Information matrix.

The proposition 1 suggest a simple, easily computable way to bound the expected maximal eigenvalue of the Fisher information matrix—choose the block with the largest eigenvalue and calculate the expected spectral radia for the corresponding off diagonal terms. We do so by making an auxiliary assumption:

Assumption 2. *The maximum singular value of $\mathbf{J}^{\mathbf{h}^g}_{\mathbf{h}^\alpha}$ monotonically increases as $\alpha \downarrow 1$.*

We motivate this assumption in a twofold fashion: firstly the work done by (Pennington et al., 2017; 2018) shows that the spectral edge, i.e. the maximal, non-negative singular value in the support of of the spectral distribution increases with depth, secondly it has been commonly observed in numerical experiments that very deep neural networks have ill conditioned gradients.

Under this assumption it is sufficient to study the maximal singular value of blocks of the Fisher information matrix with respect to $\mathrm{vec}(\mathbf{W}^1), b^1$ and the spectral norms of its corresponding off-diagonal blocks. We define functions Σ_{max} of each block as upper bounds on the spectral bounds of the respective block. The specific values are given in the following Lemma:

Lemma 3. *The maximum expected singular values of the off-diagonal blocks $\forall \beta \neq 1$ are bounded by Σ_{max}:*

$$\mathbb{E}\left[\sigma_{max}\left(\mathbf{G}_{\mathrm{vec}(\mathbf{W}^1),\mathrm{vec}(\mathbf{W}^\beta)}\right)\right] \leq \Sigma_{max}\left(\mathbf{G}_{\mathrm{vec}(\mathbf{W}^1),\mathrm{vec}(\mathbf{W}^\beta)}\right) \tag{13}$$

$$\triangleq \sqrt{N^\beta}\left|\mathbb{E}\left[\phi(h)\right]\right| \left\|\mathbb{E}\left[\mathbf{x}^0\right]\right\|_2 \mathbb{E}\left[\sigma_{max}\left(\mathbf{J}^{\mathbf{h}^g\top}_{\mathbf{h}^1}\right)\right] \mathbb{E}\left[\sigma_{max}\left(\mathbf{H}_g\right)\right] \mathbb{E}\left[\sigma_{max}\left(\mathbf{J}^{\mathbf{h}^g}_{\mathbf{h}^\beta}\right)\right] \tag{14}$$

$$\mathbb{E}\left[\sigma_{max}\left(\mathbf{G}_{\mathrm{vec}(\mathbf{W}^1),b^\beta}\right)\right] \leq \Sigma_{max}\left(\mathbf{G}_{\mathrm{vec}(\mathbf{W}^1),b^\beta}\right) \tag{15}$$

$$\triangleq \left|\mathbb{E}\left[\phi(h)\right]\right| \mathbb{E}\left[\sigma_{max}\left(\mathbf{J}^{\mathbf{h}^g\top}_{\mathbf{h}^1}\right)\right] \mathbb{E}\left[\sigma_{max}\left(\mathbf{H}_g\right)\right] \mathbb{E}\left[\sigma_{max}\left(\mathbf{J}^{\mathbf{h}^g}_{\mathbf{h}^\beta}\right)\right] \tag{16}$$

$$\mathbb{E}\left[\sigma_{max}\left(\mathbf{G}_{b^1,b^\beta}\right)\right] \leq \Sigma_{max}\left(\mathbf{G}_{b^1,b^\beta}\right) \triangleq \mathbb{E}\left[\sigma_{max}\left(\mathbf{J}^{\mathbf{h}^g\top}_{\mathbf{h}^1}\right)\right] \mathbb{E}\left[\sigma_{max}\left(\mathbf{H}_g\right)\right] \mathbb{E}\left[\sigma_{max}\left(\mathbf{J}^{\mathbf{h}^g}_{\mathbf{h}^\beta}\right)\right] \tag{17}$$

For proof see Appendix A.3

Note that the expectations for layers > 1 is over random networks realizations and averaged over data $\mathbf{x}^0$; i.e. they are taken with respect to the Gaussian measure, whereas the expectation for first layer weights is taken with respect to the empirical distribution of $\mathbf{x}^0$ (see equation 4).

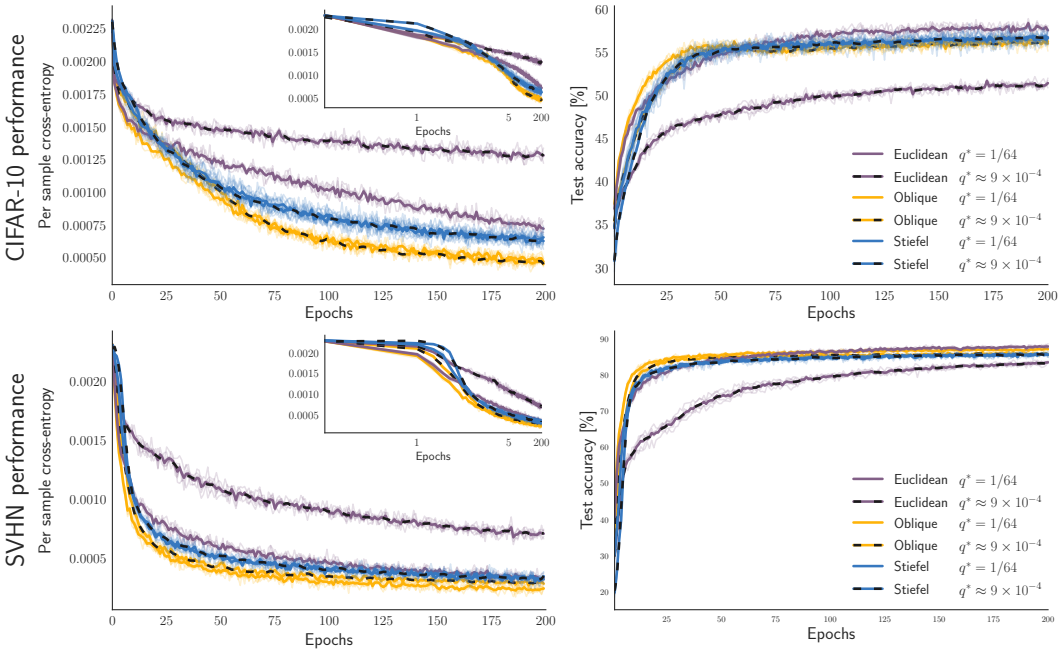


Figure 1: **Manifold constrained networks are insensitive to the choice of** q^*: Train loss and test accuracy for Euclidean, Stiefel and Oblique networks with two different values of q^*. The manifold constrained networks minimize the training loss at approximately the same rate, being faster than both Euclidean networks. Despite this, there is little difference between the test accuracy of the Stiefel and Oblique networks and the Euclidean networks initialized with $q^* = 9 \times 10^{-4}$. Notably, the latter attains a marginally higher test set accuracy towards the end of training.

Depending on the choice of q^* and therefore implicitly both the rescaling of $\mathbf{x}^0$ and the values of $\mathbb{E}[\phi(\mathbf{h})]$ the singular values of the weight blocks might dominate those associated with biases dominate — compare equation 14 and equation 17

Theorem (Bound on the Fisher Information Eigenvalues)**.** *If* $\left\|\mathbb{E}\left[\mathbf{x}^0\right]\right\|_2 \leq 1$ *then eigenvalue associated with* b^1 *will dominate, giving an upper bound on* $\lambda_{max}(\bar{\mathbf{G}})$

$$\mathbb{E}\left[\lambda_{max}(\bar{\mathbf{G}})\right] \leq \mathbb{E}\left[\sigma_{max}\left(\bar{\mathbf{G}}_{b^1,b^1}\right)\right] + \Sigma_{\max}\left(\mathbf{G}_{b^1,\mathrm{vec}(\mathbf{W}^1)}\right) + \sum_{\beta>1} \Sigma_{max}\left(\bar{\mathbf{G}}_{b^1,b^\beta}\right) + \Sigma_{max}\left(\bar{\mathbf{G}}_{\mathrm{vec}(b^1),\mathrm{vec}(\mathbf{W}^\beta)}\right)$$

otherwise the maximal eigenvalue of the FIM is bounded by

$$\mathbb{E}\left[\lambda_{max}(\bar{\mathbf{G}})\right] \leq \mathbb{E}\left[\sigma_{max}\left(\bar{\mathbf{G}}_{\mathrm{vec}(\mathbf{W}^1),\mathrm{vec}(\mathbf{W}^1)}\right)\right] + \Sigma_{\max}\left(\mathbf{G}_{b^1,\mathrm{vec}(\mathbf{W}^1)}\right) + \sum_{\beta>1} \Sigma_{max}\left(\mathbf{G}_{\mathrm{vec}(\mathbf{W}^1),b^\beta}\right) + \Sigma_{max}\left(\mathbf{G}_{\mathrm{vec}(\mathbf{W}^1),\mathrm{vec}(\mathbf{W}^\beta)}\right)$$

Moreover, it is interesting to note two things. Firstly, $\mathbb{E}\left[\sigma_{\max}\left(\mathbf{J}^{\mathbf{h}^g}_{\mathbf{h}^\alpha}\right)\right]$ *factor appear in all the above summands. Secondly, we can bound* $\sigma_{\max}$ *for the diagonal blocks with* $\mathbb{E}\left[\lambda_{\max}\left(\mathbf{H}_g\right)\right]\mathbb{E}\left[\sigma_{\max}\left(\mathbf{J}^{\mathbf{h}^g}_{\mathbf{h}^\alpha}\right)\right]^2$*. These two fact reveal that the FIM maximum eigenvalue is upper upperbounded by a quadratic function of the spectral radius of the input-output Jacobian.*

The functional form of the bound is essentially quadratic in $\mathbb{E}\left[\sigma_{max}(\mathbf{J}^{\mathbf{h}^g}_{\mathbf{h}^1})\right]$ since the term appears in the summand as with powers at most two. This result shows that the strong smoothness, given by the maximum eigenvalue of the FIM, is *proportional to* the squared maximum singular value of the input-output Jacobian. Moreover, the bound essentially depends on q^* via the expectation $\mathbb{E}[\phi(h)]$, through $\mathbf{J}^{\mathbf{h}^g}_{\mathbf{h}^1}$ and implicitly through $\mathbf{H}_g$. For regression problems this dependence is monotonically increasing in q^* (Pennington et al., 2018; 2017) since $\mathbf{H}_g$ is just the identity. However, this does not hold for all generalized linear models since $\lambda_{max}(\mathbf{H}_g)$ is not necessarily a monotonically increasing function of

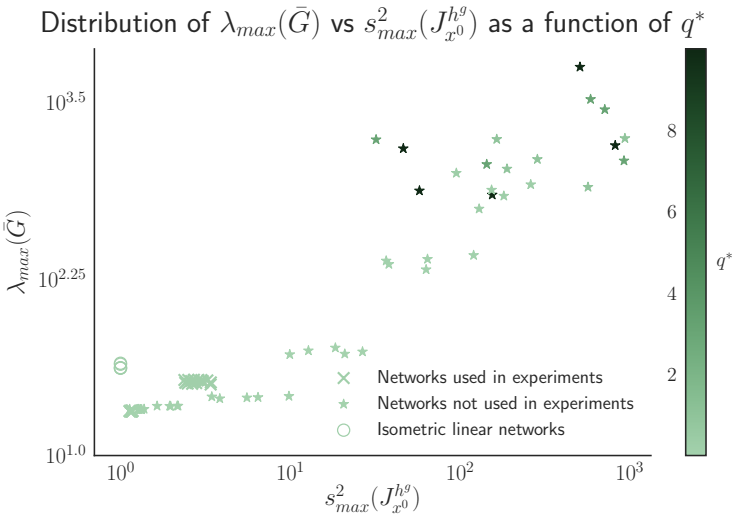


Figure 2: At initialization the maximum curvature of the loss landscape (measured by the $\lambda_{\max}$ of the Fisher correlates highly ($\rho = 0.65$) with the maximum squared singular value of the Jacobian $\mathbf{J}^{\mathbf{h}^g}_{\mathbf{x}^0}$. The choice of choice of the preactivation variance, q^* affects not only the conditioning of the gradients but also the gradient Lipschitz constant.

the pre-activation variance at layer $\mathbf{h}^g$. We demonstrate this in the case of softmax regression in the Appendix B.2. Finally, to obtain a specific bound on $\lambda_{max}(\bar{G})$ we might consider bounding each $\mathbb{E}\left[\sigma_{max}(\mathbf{J}^{\mathbf{h}^g}_{\mathbf{h}^\alpha})\right]$ appearing in theorem 3 in terms of its Frobenius norm. The corresponding result is the eigenvalue bound derived by (Karakida et al., 2018).

[t]

3.1 Numerical Experiments

To experimentally test the potential effect of maintaining orthogonality throughout training and compare it to the unconstrained optimization (Pennington et al., 2017), we trained a 200 layer $\tanh$ network on CIFAR-10 and SVHN [1]. Following (Pennington et al., 2017) we set the width of each layer to be $N = 400$ and chose the $\sigma_{\mathbf{W}}$, $\sigma_{\mathbf{b}}$ in such a way to ensure that χ concentrates on 1 but s^2_{max} varies as a function of q^* (see Fig. 2). We considered four different critical initializations with $q^* = \left[10^{-4}, \quad \frac{1}{64}, \quad \frac{1}{2}, 8\right]$, which differ both in spread of the singular values as well as in the resulting training speed and final test accuracy as reported by (Pennington et al., 2017). In the main text we focus on the smaller values since those networks should be closer to being isometric and therefore, by our theory, ought to train better. The remaining two networks with $q^* = \left[\frac{1}{2}, quad8\right]$ are presented in the Appendix A.1. To test how enforcing strict orthogonality or near orthogonality affects convergence speed and the maximum eigenvalues of the Fisher information matrix, we trained Stiefel and Oblique constrained networks and compared them to the unconstrained "Euclidean" network described in (Pennington et al., 2017). We used a Riemannian version of ADAM (Kingma & Ba, 2014). When performing gradient descent on non-Euclidean manifolds, we split the variables into three groups: (1) Euclidean variables (e.g. the weights of the classifier layer, biases), (2) non-negative scaling $\sigma_{\mathbf{W}}$ both optimized using the regular version of ADAM, and (3) manifold variables optimized using Riemannian ADAM. The initial learning rates for all the groups, as well as the non-orthogonality penalty (see **??**) for Oblique networks were chosen via Bayesian optimization, maximizing validation set accuracy after 50 epochs. All networks were trained with a minibatch size of 1000. We trained 5 networks of each kind, and collected eigenvalue and singular value statistics every 5 epochs, from the first to the fiftieth, and then after the hundredth and two hundredth epochs.

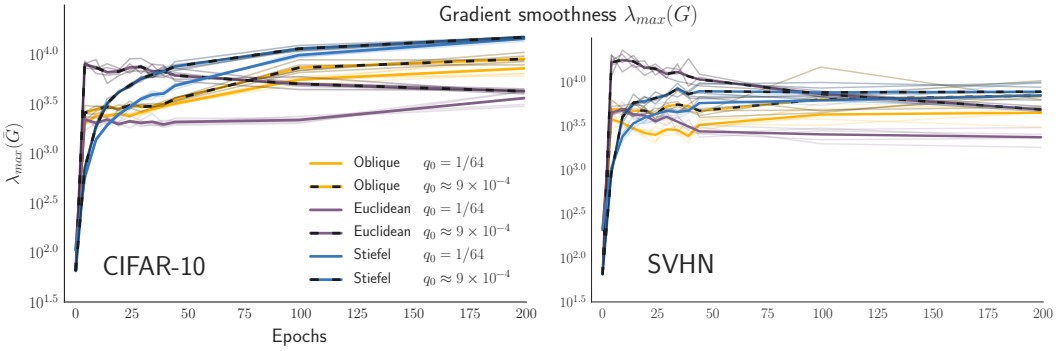
Based on the bound on the maximum eigenvalue of the Fisher information matrix derived in Section 3, we predicted that at initialization $\lambda_{max}(\bar{\mathbf{G}})$ should covary with $\sigma^2_{max}(\mathbf{J}^{\mathbf{h}^g}_{\mathbf{x}^0})$. We tested our prediction using the empirical Fisher information matrix (Kunstner et al., 2019) and we find a significant correlation between the two (Pearson coefficient $\rho = 0.64$). The numerical values are presented in

[1] `https://github.com/iclr2020info-geom/info-geom`

Fig. 2. Additionally we see that both the maximum singular value and maximum eigenvalue increase monotonically as a function of q^*. Motivated by the previous work by (Saxe et al., 2013) showing depth independent learning dynamics in linear orthogonal networks, we included 5 instantiations of this model in the comparison. The input to the linear network was normalized the same way as the critical, non-linear networks with $q^* = 1/64$. The deep linear networks had a substantially larger $\lambda_{max}(\bar{\mathbf{G}})$ than its non-linear counterparts initialized with identically scaled input (Fig. 2). Having established a connection between q^* the maximum singular value of the hidden layer input-output Jacobian and the maximum eigenvalue of the Fisher information, we investigate the effects of initialization on subsequent optimization. As reported by (Pennington et al., 2017), the learning speed and generalization peak at intermediate values of $q^* \approx 10^{-0.5}$. This result is counter intuitive given that the maximum eigenvalue of the Fisher information matrix, much like that of the Hessian in convex optimization, upper bounds the maximal learning rate (Boyd & Vandenberghe, 2004; Bottou et al., 2016). To gain insight into the effects of the choice of q^* on the convergence rate, we trained the Euclidean networks and estimated the local values of λ_{max} during optimization. At the same time we asked whether we can effectively control the two aforesaid quantities by constraining the weights of each layer to be orthogonal or near orthogonal. To this end we trained Stiefel and Oblique networks and recorded the same statistics.

We present training results in Fig. 1, where it can be seen that Euclidean networks with $q^* \approx 9 \times 10^{-4}$ perform worse with respect to training loss and test accuracy than those initialized with $q^* = 1/64$. On the other hand, manifold constrained networks are insensitive to the choice of q^*. Moreover, Stiefel and Oblique networks perform marginally worse on the test set compared to the Euclidean network with $q^* = 1/64$, despite attaining a lower training loss. This latter fact indicates that manifold constrained networks the are perhaps prone to overfitting.

We observe that reduced performance of Euclidean networks initialized with $q^* \approx 9 \times 10^{-4}$ may partially be explained by their rapid increase in $\lambda_{max}(\bar{\mathbf{G}})$ within the initial 5 epochs of optimization (see Fig. 3.1 in the Appendix). While all networks undergo this rapid increase, it is most pronounced for Euclidean networks with $q^* \approx 9 \times 10^{-4}$. The increase $\lambda_{max}(\bar{\mathbf{G}})$ correlates with the inflection point in the training loss curve that can be seen in the inset of Fig. 1. Interestingly, the manifold constrained networks optimize efficiently despite differences in $\lambda_{max}(\bar{\mathbf{G}})$ (Kohler et al., 2018).



Notably, the Euclidean network with $q^* = 1/64$ has almost an order of magnitude smaller $\lambda_{max}(\bar{\mathbf{G}})$ than the Stiefel and Oblique networks, but reduces training loss at a slower rate.

Notably, the Euclidean network with $q^* = 1/64$ has almost an order of magnitude smaller $\lambda_{max}(\bar{\mathbf{G}})$ than the Stiefel and Oblique networks, but reduces training loss at a slower rate.

Figure 3: For manifold constrained networks, gradient smoothness is not predictive of optimization rate. Euclidean networks with a low initial $\lambda_{max}(\bar{G})$ rapidly become less smooth, whereas Euclidean networks with a larger $\lambda_{max}(\bar{G})$ remain relatively smoother.
Notably, the Euclidean network with $q^* = 1/64$ has almost an order of magnitude smaller $\lambda_{max}(\bar{\mathbf{G}})$ than the Stiefel and Oblique networks, but reduces training loss at a slower rate.

4 Discussion

Critical orthogonal initializations have proven tremendously successful in rapidly training very deep neural networks (Pennington et al., 2017; Chen et al., 2018; Pennington et al., 2018; Xiao et al., 2018a). Despite their elegant derivation drawing on methods from free probability and mean

field theory, they did not offer a clear optimization perspective on the mechanisms driving their success. With this work we complement the understanding of critical orthogonal initializations by showing that the maximum eigenvalue of the Fisher information matrix, and consequentially the local gradient smoothness is proportional to the maximum singular value of the input-output Jacobian. This gives an information geometric account of why the step size and training speed depend on q^* via its effect on $\mathbb{E}\left[s_{max}(\mathbf{J}^{\mathbf{h}^L}_{\mathbf{h}^1})\right]$. We observed in numerical experiments that the paradoxical results reported in (Pennington et al., 2017) whereby training speed and generalization attains a maximum for $q^* = 10^{-0.5}$ can potentially be explained by a rapid increase of the maximum eigenvalue of the FIM during training for the networks initialized with Jacobians closer to being isometric (i.e., smaller q^*). This increase effectively limits the learning rate during the early phase of optimization and highlights the need to analyze the trajectories of training rather than just initializations. We relate that to the recently proposed Neural Tangent Kernel (Jacot et al., 2018; Lee et al., 2019). The NTK is defined as

$$\hat{\Theta}_{t,i,j} \triangleq \mathbf{J}^{\mathbf{h}^g}_{\mathbf{x}^0} \mathbf{J}^{\mathbf{h}^g \top}_{\mathbf{x}^0} \tag{18}$$

for $i, j \in N^g|\mathcal{D}|$ representing the block indices running over N^g outputs of the network and $|\mathcal{D}|$ data samples. The NTK is the derivative of a kernel defined by a random neural network. It prescribes the time evolution of the function and therefore offers a unique insight into the training dynamics. Importantly, the spectrum of the NTK coincides with that of the Fisher information *for regression problems* (see Appendix B.4). In other words,

It is therefore interesting to understand the predictiveness of the Neural Tangent Kernel at initialization given its spectrum. Such a result has been recently presented by (Lee et al., 2019), who show that the discrepancy between training with a NTK frozen at initialization ($f^{lin}_t(\mathbf{x}^0)$) and a continuously updated one ($f_t(, \mathbf{x}^0)$) can be bounded. Importantly the authors showed that rate at which discrepancy accrues depends exponentially on the smallest eigenvalue of the NTK. Given that the spectra of the Neural Tangent Kernel and the Fisher Information matrix coincide we can reason about this discrepancy over training time in terms of the smallest and largest eigenvalues of the Fisher Information matrix.

Lemma 4 ((Lee et al., 2019))**.** *The discrepancy between $g^{lin}(t) = f^{lin}_t(\mathbf{x}^0) - \mathbf{y}$ and $g(t) = f_t(\mathbf{x}^0) - \mathbf{y}$*

$$e^{\lambda_{min}(\bar{\mathbf{G}}_0)\eta t} \left\| g^{\text{lin}}(t) - g(t) \right\|_2 \leq \left(\eta \int_0^t e^{\lambda_{min}(\bar{\mathbf{G}})_0 \eta s} \left\| \left(\bar{\mathbf{G}}_s - \bar{\mathbf{G}}_0 \right) \right\| \left\| g^{\text{lin}}(s) \right\|_2 ds \right) e^{\int_0^t \left(\eta \left\| \left(\bar{\mathbf{G}}_s - \bar{\mathbf{G}}_0 \right) \right\| \right) ds} \tag{19}$$

where η is the learning rate.

Given the approximately inverse relation between the maximum and minimum eigenvalues of the Fisher information matrix (Karakida et al., 2018), decreasing $q*$ increases $\lambda_{min}(\bar{\mathbf{G}}_0)$ and the the solutions rapidly diverge. This implies that a low condition number $\frac{\lambda_{max}(\bar{\mathbf{G}}_0)}{\lambda_{max}(\bar{\mathbf{G}}_0)}$ may be undesirable, and a degree of anisotropy is necessary for the Fisher Information matrix to be predictive of training performance.

Finally, we compared manifold constrained networks with the Euclidean network, each evaluated with two initial values of q^*. From these experiments we draw the conclusion that manifold constrained networks are less sensitive to the initial strong smoothness, unlike their Euclidean counterparts. Furthermore, we observe that the rate at which Stiefel and Oblique networks decrease training loss is not dependent on their gradient smoothness, a result which is consistent with the recent analysis of (Kohler et al., 2018).

REFERENCES

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, N.J. ; Woodstock, December 2007. ISBN 978-0-691-13298-3.

Shun-ichi Amari, Ryo Karakida, and Masafumi Oizumi. Fisher Information and Natural Gradient Learning of Random Deep Networks. *arXiv:1808.07172 [cond-mat, stat]*, August 2018. URL `http://arxiv.org/abs/1808.07172`. arXiv: 1808.07172.

Nitin Bansal, Xiaohan Chen, and Zhangyang Wang. Can We Gain More from Orthogonality Regularizations in Training Deep Networks? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman,

N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 4261–4271. Curran Associates, Inc., 2018.

Aleksandar Botev, Hippolyt Ritter, and David Barber. Practical Gauss-Newton Optimisation for Deep Learning. In *International Conference on Machine Learning*, pp. 557–565, July 2017. URL http://proceedings.mlr.press/v70/botev17a.html.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization Methods for Large-Scale Machine Learning. *arXiv:1606.04838 [cs, math, stat]*, June 2016. URL http://arxiv.org/abs/1606.04838. arXiv: 1606.04838.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization, With Corrections 2008*. Cambridge University Press, Cambridge, UK ; New York, 1 edition edition, March 2004. ISBN 978-0-521-83378-3.

Minmin Chen, Jeffrey Pennington, and Samuel Schoenholz. Dynamical Isometry and a Mean Field Theory of RNNs: Gating Enables Signal Propagation in Recurrent Neural Networks. In *International Conference on Machine Learning*, pp. 873–882, July 2018. URL http://proceedings.mlr.press/v80/chen18i.html.

Minhyung Cho and Jaehyung Lee. Riemannian approach to batch normalization. In *Advances in Neural Information Processing Systems*, pp. 5229–5239, 2017.

Alan Edelman, Tomás A. Arias, and Steven T. Smith. The Geometry of Algorithms with Orthogonality Constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, January 1998. ISSN 0895-4798, 1095-7162. doi: 10.1137/S0895479895290954. URL http://epubs.siam.org/doi/10.1137/S0895479895290954.

Mehrtash Harandi and Basura Fernando. Generalized BackPropagation, \'{E}tude De Cas: Orthogonality. *arXiv:1611.05927 [cs]*, November 2016. URL http://arxiv.org/abs/1611.05927. 00004 arXiv: 1611.05927.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]*, December 2015. URL http://arxiv.org/abs/1512.03385. 01528 arXiv: 1512.03385.

Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167 [cs]*, February 2015. URL http://arxiv.org/abs/1502.03167. 00385 arXiv: 1502.03167.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. *arXiv:1806.07572 [cs, math, stat]*, June 2018. URL http://arxiv.org/abs/1806.07572. arXiv: 1806.07572.

Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal Statistics of Fisher Information in Deep Neural Networks: Mean Field Approach. *arXiv:1806.01316 [cond-mat, stat]*, June 2018. URL http://arxiv.org/abs/1806.01316. arXiv: 1806.01316.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, December 2014. URL http://arxiv.org/abs/1412.6980. 01869 arXiv: 1412.6980.

Jonas Kohler, Hadi Daneshmand, Aurelien Lucchi, Ming Zhou, Klaus Neymeyr, and Thomas Hofmann. Exponential convergence rates for Batch Normalization: The power of length-direction decoupling in non-convex optimization. *arXiv:1805.10694 [cs, stat]*, May 2018. URL http://arxiv.org/abs/1805.10694. arXiv: 1805.10694.

Frederik Kunstner, Lukas Balles, and Philipp Hennig. Limitations of the Empirical Fisher Approximation. *arXiv:1905.12558 [cs, stat]*, May 2019.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep Neural Networks as Gaussian Processes. *arXiv:1711.00165 [cs, stat]*, October 2017. URL http://arxiv.org/abs/1711.00165. arXiv: 1711.00165.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent. *arXiv:1902.06720 [cs, stat]*, February 2019.

James Martens and Roger Grosse. Optimizing Neural Networks with Kronecker-factored Approximate Curvature. In *International Conference on Machine Learning*, pp. 2408–2417, June 2015. URL http://proceedings.mlr.press/v37/martens15.html.

Alexander G. de G. Matthews, Mark Rowland, Jiri Hron, Richard E. Turner, and Zoubin Ghahramani. Gaussian Process Behaviour in Wide Deep Neural Networks. *arXiv:1804.11271 [cs, stat]*, April 2018. URL http://arxiv.org/abs/1804.11271. arXiv: 1804.11271.

Radford M. Neal. *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. Springer New York, New York, NY, 1996. ISBN 978-0-387-94724-2 978-1-4612-0745-0. doi: 10.1007/978-1-4612-0745-0. URL http://link.springer.com/10.1007/978-1-4612-0745-0.

Jeffrey Pennington and Yasaman Bahri. Geometry of neural network loss surfaces via random matrix theory. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 2798–2806, Sydney, NSW, Australia, 2017. JMLR.org.

Jeffrey Pennington, Sam Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. *Advances in neural information processing systems*, 2017.

Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. The Emergence of Spectral Universality in Deep Networks. *arXiv:1802.09979 [cs, stat]*, February 2018. URL http://arxiv.org/abs/1802.09979. arXiv: 1802.09979.

Tim Salimans and Diederik P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *arXiv:1602.07868 [cs]*, February 2016. URL http://arxiv.org/abs/1602.07868. 00003 arXiv: 1602.07868.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013. URL http://arxiv.org/abs/1312.6120. 00083.

Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep Information Propagation. *arXiv:1611.01232 [cs, stat]*, November 2016. URL http://arxiv.org/abs/1611.01232. arXiv: 1611.01232.

Christiane Tretter. *Spectral Theory of Block Operator Matrices and Applications*. IMPERIAL COLLEGE PRESS, October 2008. ISBN 978-1-86094-768-1 978-1-84816-112-2. doi: 10.1142/p493. URL http://www.worldscientific.com/worldscibooks/10.1142/p493.

Eugene Vorontsov, Chiheb Trabelsi, Samuel Kadoury, and Chris Pal. On orthogonality and learning recurrent networks with long term dependencies. *arXiv:1702.00071 [cs]*, January 2017. URL http://arxiv.org/abs/1702.00071. arXiv: 1702.00071.

Scott Wisdom, Thomas Powers, John R. Hershey, Jonathan Le Roux, and Les Atlas. Full-capacity unitary recurrent neural networks. *arXiv:1611.00035 [cs, stat]*, October 2016. URL http://arxiv.org/abs/1611.00035. arXiv: 1611.00035.

Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel Schoenholz, and Jeffrey Pennington. Dynamical Isometry and a Mean Field Theory of CNNs: How to Train 10,000-Layer Vanilla Convolutional Neural Networks. In *International Conference on Machine Learning*, pp. 5393–5402, July 2018a. URL http://proceedings.mlr.press/v80/xiao18a.html.

Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel S. Schoenholz, and Jeffrey Pennington. Dynamical Isometry and a Mean Field Theory of CNNs: How to Train 10,000-Layer Vanilla Convolutional Neural Networks. *arXiv:1806.05393 [cs, stat]*, June 2018b. URL http://arxiv.org/abs/1806.05393. arXiv: 1806.05393.

A APPENDIX

A.1 ADDITIONAL FIGURES

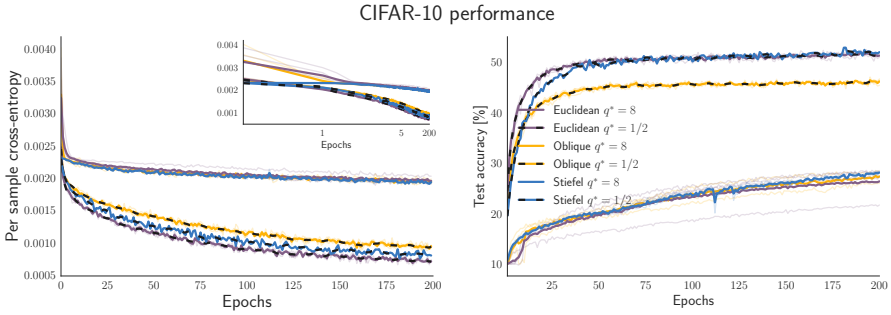


Figure 4: Training performance for networks with $q^* = \frac{1}{2}$ and $q^* = 8$. The behavior of the training loss as well as the validation accuracy is consistent with the observations that for a large range of parameters q^* the manifold constrained networks are insensitive to initialization and gradient smoothness.
The manifold constrained networks deal more favorably given a less smooth loss landscape.

A.2 BLOCK GERSHGORIN THEOREM

In Section 3, we considered a block diagonal approximation to the Fisher information matrix and derived an upper bound on the spectral norm for all the blocks. Using the properties of the off-diagonal blocks, we can get a more accurate estimate of the maximal eigenvalue of the Fisher information might be. First, let us consider an arbitrarily partitioned matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, with spectrum $\lambda(\mathbf{A})$ The partitioning is done with respect to the set

$$\pi = \{p_j\}_{j=0}^{L} \tag{20}$$

with the elements of the set satisfying $0 < p_1 < p_2 < \ldots < p_L = N$. Then each block of the matrix $A_{i,j}$ is a potentially rectangular matrix in $\mathbb{R}^{(p_i - p_{i-1}) \times (p_j - p_{j-1})}$. We assume that $\mathbf{A}_{i,i}$ is self-adjoint for all i.
Let us define a disk as

$$C(c, r) \triangleq \{\lambda : \|c - \lambda\| \leq r\}. \tag{21}$$

The theorem as presented in (Tretter, 2008) shows that the eigenvalues of $\lambda(\mathbf{A})$ are contained in a union of Gershgorin disks defined as follows

$$\lambda(\mathbf{A}) \subset \bigcup_{i=1}^{L} \left\{ \bigcup_{k=1}^{p_i - p_{i-1}} C\left(\lambda_k(\mathbf{A}_{ii}), \sum_{j=1, j \neq i}^{L} s_{max}(\mathbf{A}_{i,j}) \right) \right\} \tag{22}$$

where the inner union is over a set disks for each eigenvalue of the block diagonal $\mathbf{A}_{i,i}$ while the outer union is over the L blocks in $\mathbf{A}$. The radius of the disk is constant for every eigenvalue in the ith diagonal block $\mathbf{A}_{i,i}$ and is given by the sum of singular values of the off diagonal blocks. Therefore, the largest eigenvalue of $\mathbf{A}$ lies in

$$\lambda_{max}(\mathbf{A}) \subset \bigcup_{i=1}^{L} C\left(\lambda_{max}(\mathbf{A}_{ii}), \sum_{j=1, j \neq i}^{L} s_{max}(\mathbf{A}_{i,j}) \right) \tag{23}$$

A.3 DERIVATION OF THE EXPECTED SINGULAR VALUES

$$\sigma_{max}\left(\mathbf{G}_{\text{vec}(\mathbf{W}^1), \text{vec}(\mathbf{W}^\beta)}\right) = \mathbb{E}\left[\sigma_{max}\left(\phi(h) \mathbb{1} \mathbf{x}^{0\top}\right)\right] \otimes \mathbb{E}\left[\sigma_{max}\left(\mathbf{J}_{\mathbf{h}^1}^{\mathbf{h}^g \top} \mathbf{H}_g \mathbf{J}_{\mathbf{h}^\beta}^{\mathbf{h}^g}\right)\right] \tag{24}$$

$$= \sqrt{N^\beta} \left|\mathbb{E}\left[\phi(h)\right]\right| \left\|\mathbb{E}\left[\mathbf{x}^0\right]\right\|_2 \mathbb{E}\left[\sigma_{max}\left(\mathbf{J}_{\mathbf{h}^1}^{\mathbf{h}^g \top} \mathbf{H}_g \mathbf{J}_{\mathbf{h}^\beta}^{\mathbf{h}^g}\right)\right] \tag{25}$$

$$\leq \sqrt{N^\beta} \left|\mathbb{E}\left[\phi(h)\right]\right| \left\|\mathbb{E}\left[\mathbf{x}^0\right]\right\|_2 \mathbb{E}\left[\sigma_{max}\left(\mathbf{J}_{\mathbf{h}^1}^{\mathbf{h}^g \top}\right)\right] \mathbb{E}\left[\sigma_{max}\left(\mathbf{H}_g\right)\right] \mathbb{E}\left[\sigma_{max}\left(\mathbf{J}_{\mathbf{h}^\beta}^{\mathbf{h}^g}\right)\right] \tag{26}$$

$$\sigma_{max}\left(\mathbf{G}_{\mathrm{vec}(\mathbf{W}^1),b^\beta}\right) \leq \mathbb{E}\left[\sigma_{max}\left(\mathbf{x}^{0\top}\otimes\mathbf{I}\right)\right]\mathbb{E}\left[\sigma_{max}\left(\mathbf{J}_{\mathbf{h}^\alpha}^{\mathbf{h}^g\top}\mathbf{H}_g\mathbf{J}_{\mathbf{h}^\beta}^{\mathbf{h}^g}\right)\right] \tag{27}$$

$$= \left\|\mathbb{E}\left[\mathbf{x}^0\right]\right\|_2\left(\mathbb{E}\left[\sigma_{max}\ \mathbf{J}_{\mathbf{h}^1}^{\mathbf{h}^g\top}\mathbf{H}_g\mathbf{J}_{\mathbf{h}^\beta}^{\mathbf{h}^g}\right)\right] \tag{28}$$

$$\leq \left\|\mathbb{E}\left[\mathbf{x}^0\right]\right\|_2\mathbb{E}\left[\sigma_{max}\left(\mathbf{J}_{\mathbf{h}^1}^{\mathbf{h}^g\top}\right)\right]\mathbb{E}\left[\sigma_{max}\left(\mathbf{H}_g\right)\right]\mathbb{E}\left[\sigma_{max}\left(\mathbf{J}_{\mathbf{h}^\beta}^{\mathbf{h}^g}\right)\right] \tag{29}$$

$$\sigma_{max}\left(\mathbf{G}_{b^1,\mathrm{vec}(\mathbf{W}^\beta)}\right) \leq \mathbb{E}\left[\sigma_{max}\left(\mathbf{x}^{\beta-1\top}\otimes\mathbf{I}\right)\right] \tag{30}$$

$$= \sqrt{N^\beta}\left|\mathbb{E}\left[\phi(h)\right]\right|\mathbb{E}\left[\sigma_{max}\left(\mathbf{J}_{\mathbf{h}^1}^{\mathbf{h}^g\top}\mathbf{H}_g\mathbf{J}_{\mathbf{h}^\beta}^{\mathbf{h}^g}\right)\right] \tag{31}$$

$$\leq \left|\mathbb{E}\left[\phi(h)\right]\right|\mathbb{E}\left[\sigma_{max}\left(\mathbf{J}_{\mathbf{h}^1}^{\mathbf{h}^g\top}\right)\right]\mathbb{E}\left[\sigma_{max}\left(\mathbf{H}_g\right)\right]\mathbb{E}\left[\sigma_{max}\left(\mathbf{J}_{\mathbf{h}^\beta}^{\mathbf{h}^g}\right)\right] \tag{32}$$

$$\sigma_{max}\left(\mathbf{G}_{b^1,b^\beta}\right) = \mathbb{E}\left[\sigma_{max}\left(\mathbf{J}_{\mathbf{h}^1}^{\mathbf{h}^g\top}\mathbf{H}_g\mathbf{J}_{\mathbf{h}^\beta}^{\mathbf{h}^g}\right)\right] \tag{33}$$

$$\leq \mathbb{E}\left[\sigma_{max}\left(\mathbf{J}_{\mathbf{h}^1}^{\mathbf{h}^g\top}\right)\right]\mathbb{E}\left[\sigma_{max}\left(\mathbf{H}_g\right)\right]\mathbb{E}\left[\sigma_{max}\left(\mathbf{J}_{\mathbf{h}^\beta}^{\mathbf{h}^g}\right)\right] \tag{34}$$

B APPENDIX

You may include other additional sections here.

B.1 BOUND THE DIAGONAL ELEMENT

$$\sigma_{max}\left(\mathbf{G}_{b^1,b^1}\right) \leq \Sigma_{max}\left(\mathbf{G}_{b^1,b^1}\right) \tag{35}$$

$$\triangleq \mathbb{E}\left[\sigma_{max}\left(\mathbf{H}_g\right)\right]\mathbb{E}\left[\sigma_{max}\left(\mathbf{J}_{\mathbf{h}^\beta}^{\mathbf{h}^g}\right)\right]^2 \tag{36}$$

$$\sigma_{max}\left(\mathbf{G}_{\mathrm{vec}(\mathbf{W}^1),\mathrm{vec}(\mathbf{W}^1)}\right) \leq \Sigma_{max}\left(\mathbf{G}_{\mathrm{vec}(\mathbf{W}^1),\mathrm{vec}(\mathbf{W}^1)}\right) \tag{37}$$

$$\triangleq \sigma_{max}\left(\mathrm{Cov}[\mathbf{x}^0,\mathbf{x}^0]\right)\mathbb{E}\left[\sigma_{max}\left(\mathbf{H}_g\right)\right]\mathbb{E}\left[\sigma_{max}\left(\mathbf{J}_{\mathbf{h}^\beta}^{\mathbf{h}^g}\right)\right]^2 \tag{38}$$

B.2 MONTECARLO ESTIMATE OF SPECTRAL RADIUS OF $\mathbf{H}_g$ FOR 10 WAY SOFTMAX CLASSIFICATION

B.3 MANIFOLD OPTIMIZATION

The potentially non-convex constraint set constitutes a Riemannian manifold, when it is locally isomorphic to $\mathbb{R}^n$, differentiable and endowed with a suitable (Riemannian) metric, which allows us to measure distances in the tangent space and consequentially also define distances on the manifold. There is considerable freedom in choosing a Riemannian metric; here we consider the metric inherited from the Euclidean embedding space which is defined as $\langle\mathbf{W},\mathbf{W}'\rangle \triangleq \mathrm{Tr}(\mathbf{W}'^\top\mathbf{W})$. To optimize a cost function with respect to parameters lying in a non-Euclidean manifold we must define a descent direction. This is done by defining a manifold equivalent of the directional derivative. An intuitive approach replaces the movement along a vector $\mathbf{t}$ with movement along a geodesic curve $\gamma(t)$, which lies in the manifold and connects two points $\mathbf{W},\mathbf{W}'\in\mathcal{M}$ such that $\gamma(0)=\mathbf{W}$, $\gamma(1)=\mathbf{W}'$. The derivative of an arbitrary smooth function $f(\gamma(t))$ with respect to t then defines a tangent vector for each t.

Tangent vector $\xi_\mathbf{W}$ is a tangent vector at $\mathbf{W}$ if $\xi_\mathbf{W}$ satisfies $\gamma(0)=\mathbf{W}$ and

$$\xi_\mathbf{W} \triangleq \left.\frac{\mathrm{d}f(\gamma(t))}{\mathrm{d}t}\right|_{t=0} \triangleq \gamma'(0)f \tag{39}$$

The set of all tangents to $\mathcal{M}$ at $\mathbf{W}$ is referred to as the tangent space to $\mathcal{M}$ at $\mathbf{W}$ and is denoted by $T_\mathbf{W}\mathcal{M}$. The geodesic importantly is then specified by a constant velocity curve $\gamma''(t)=0$ with initial velocity $\xi_\mathbf{W}$. To perform a gradient step, we must then move along $\xi_\mathbf{W}$ while respecting the manifold constraint. This is achieved by applying the exponential map defined as $\mathrm{Exp}_\mathbf{W}(\xi_\mathbf{W}) \triangleq \gamma(1)$, which moves $\mathbf{W}$ to another point $\mathbf{W}'$ along the geodesic. While certain manifolds, such as the Oblique manifold, have efficient closed-form exponential maps, for general Riemannian

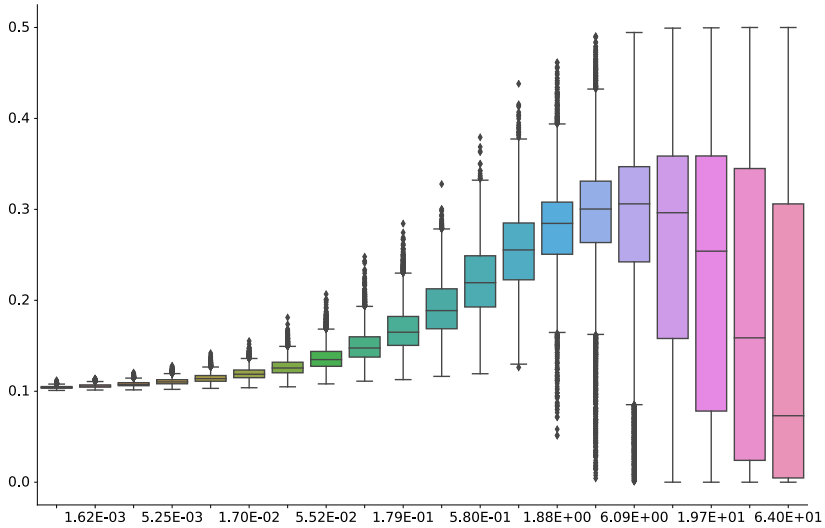


Figure 5: **Distribution of $\lambda_{max}(\mathbf{H}_g)$ as a function of q^***: In general, increasing the variance of the distribution of h^g does not result in a monotonic increase in the spectral radius of the Hessian of the GLM layer. We plot the distribution of the maximum eigenvalues as a function of the variance of the softmax layer obtained from factorizing $10,000$ random matrices.

manifolds, the computation of the exponential map involves numerical solution to a non-linear ordinary differential equation (Absil et al., 2007). An efficient alternative to numerical integration is given by an orthogonal projection onto the manifold. This projection is formally referred to as a retraction $\mathrm{Rt}_{\mathbf{W}} : T_{\mathbf{W}}\mathcal{M} \to \mathcal{M}$.

Finally, gradient methods using Polyak (heavy ball) momentum (e.g. ADAM (Kingma & Ba, 2014)) require the iterative updating of terms which naturally lie in the tangent space. The parallel translation $\mathcal{T}_\zeta(\xi) : T\mathcal{M} \bigoplus T\mathcal{M} \to T\mathcal{M}$ generalizes vector composition from Euclidean to non-Euclidean manifolds, by moving the tangent ξ along the geodesic with initial velocity $\zeta \in \mathcal{T}$ and endpoint $\mathbf{W}'$, and then projecting the resulting vector onto the tangent space $T_{\mathbf{W}'}\mathcal{M}$. As with the exponential map, parallel transport $\mathcal{T}$ may require the solution of non-linear ordinary differential equation. To alleviate the computational burden, we consider *vector transport* as an effective, projection-like solution to the parallel translation problem. We overload the notation and also denote it as $\mathcal{T}$, highlighting the similar role that the two mappings share. Technically, the geodesics and consequentially the exponential map, retraction as well as transport $\mathcal{T}$ depend on the choice of the Riemannian metric. Putting the equations together the updating scheme for Riemannian stochastic gradient descent on the manifold is

$$\mathbf{W}_{t+1} = \Pi_{\mathbf{W}_t}(-\eta_t \, \mathrm{grad} f) \tag{40}$$

where Π is either the exponential map Exp or the retraction Rt and $\mathrm{grad} f$ is the gradient of the function $f(\mathbf{W})$ lying in the tangent space $T_{\mathbf{W}}\mathcal{M}$.

B.3.1 Optimizing over the Oblique manifold

(Cho & Lee, 2017) proposed an updating scheme for optimizing neural networks where the weights of each layer are constrained to lie in the oblique manifold $\mathrm{Ob}(p, n)$. Using the fact that the manifold itself is a product of p unit-norm spherical manifolds, they derived an efficient, closed-form Riemannian gradient descent updating scheme. In particular the optimization simplifies to the optimization over $\mathrm{Ob}(1, n)$ for each column $\mathbf{w}_{i\in\{1,\dots,p\}}$ of $\mathbf{W}$.

Oblique gradient The gradient $\mathrm{grad} f$ of the cost function f with respect to the weights lying in $\mathrm{Ob}(1, n)$ is given as a projection of the Euclidean gradient $\mathrm{Grad} f$onto the tangent at $\mathbf{w}$

$$\mathrm{grad} f = \mathrm{Grad} f - (\mathbf{w}^\top \mathrm{Grad} f)\mathbf{w} \tag{41}$$

Oblique exponential map The exponential map $\mathrm{Exp}_{\mathbf{w}}$ moving $\mathbf{w}$ to $\mathbf{w}'$ along a geodesic with initial velocity $\xi_{\mathbf{w}}$

$$\mathrm{Exp}_{\mathbf{w}} = \xi_{\mathbf{w}} \cos(\|\mathbf{w}\|) + \frac{\mathbf{w}}{\|\mathbf{w}\|} \sin(\|\mathbf{w}\|) \tag{42}$$

Oblique parallel translation The parallel translation $\mathcal{T}$ moves the tangent vector $\xi_{\mathbf{w}}$ along the geodesic with initial velocity $\zeta_{\mathbf{w}}$

$$\begin{aligned} \mathcal{T}_{\zeta_{\mathbf{w}}}(\xi_{\mathbf{w}}) = \xi_{\mathbf{w}} - \\ \frac{\zeta_{\mathbf{w}}}{\|\zeta_{\mathbf{w}}\|}((1-\cos(\|\zeta_{\mathbf{w}}\|)) + \mathbf{w}\sin(\|\zeta_{\mathbf{w}}\|))\frac{\zeta_{\mathbf{w}}}{\|\zeta_{\mathbf{w}}\|}^{\top}\xi_{\mathbf{w}} \end{aligned} \tag{43}$$

B.3.2 Optimizing over the Stiefel manifold

Optimization over Stiefel manifolds in the context of neural networks has been studied by (Harandi & Fernando, 2016; Wisdom et al., 2016; Vorontsov et al., 2017). Unlike (Wisdom et al., 2016; Vorontsov et al., 2017) we propose the parametrization using the Euclidean metric, which results in a different definition of vector transport.

Stiefel gradient The gradient $\mathrm{grad} f$ of the cost function f with respect to the weights lying in $\mathrm{St}(p, n)$ is given as a projection of the Euclidean gradient $\mathrm{Grad} f$onto the tangent at $\mathbf{W}$ (Edelman et al., 1998; Absil et al., 2007)

$$\begin{aligned} \mathrm{grad} f = (\mathbf{I} - \mathbf{W}\mathbf{W}^{\top})\mathrm{Grad} f \\ + \frac{1}{2}\mathbf{W}\left(\mathbf{W}^{\top}\mathrm{Grad} f - \mathrm{Grad} f^{\top}\mathbf{W}\right) \end{aligned} \tag{44}$$

Stiefel retraction The retraction $\mathrm{Rt}_{\mathbf{W}}(\xi_{\mathbf{W}})$ for the Stiefel manifold is given by the Q factor of the QR decomposition (Absil et al., 2007).

$$\mathrm{Rt}_{\mathbf{W}}(\xi_{\mathbf{W}}) = \mathrm{qf}(\mathbf{W} + \xi_{\mathbf{W}}) \tag{45}$$

Stiefel vector transport The vector transport $\mathcal{T}$ moves the tangent vector $\xi_{\mathbf{w}}$ along the geodesic with initial velocity $\zeta_{\mathbf{w}}$ for $\mathbf{W} \in \mathrm{St}(p, n)$ endowed with the Euclidean metric.

$$\mathcal{T}_{\zeta_{\mathbf{w}}}(\xi_{\mathbf{w}}) = \left(\mathbf{I} - \mathbf{Y}\mathbf{Y}^{\top}\right)\xi_{\mathbf{W}} + \frac{1}{2}\mathbf{Y}\left(\mathbf{Y}^{\top}\xi_{\mathbf{W}} - \xi_{\mathbf{W}}{}^{\top}\mathbf{Y}\right) \tag{46}$$

where $\mathbf{Y} \triangleq \mathrm{Rt}_{\mathbf{W}}(\zeta_{\mathbf{W}})$. It is easy to see that the transport $\mathcal{T}$ consists of a retraction of tangent $\zeta_{\mathbf{W}}$ followed by the orthogonal projection of $\eta_{\mathbf{W}}$ at $\mathrm{Rt}_{\mathbf{W}}(\zeta_{\mathbf{W}})$. The projection is the same as the one mapping $\mathrm{P} : \mathrm{Grad} f \to \mathrm{grad} f$ in equation 44.

B.3.3 Optimizing over non-compact manifolds

The critical weight initialization yielding a singular spectrum of the Jacobian tightly concentrating on 1 implies that a substantial fraction of the pre-activations lie in expectation in the linear regime of the squashing nonlinearity and as a consequence the network acts quasi-linearly. To relax this constraint during training we allow the scales of the manifold constrained weights to vary. We chose to represent the weights as a product of a scaling diagonal matrix and a matrix belonging to the manifold. Then the optimization of each layer consists in the optimization of the two variables in the product. In this work we only consider isotropic scalings, but the method generalizes easily to the use of any invertible square matrix.

B.4 FIM and NTK have the same spectrum

The empirical Neural Tangent Kernel (NTK) Recall the definition in equation 18:

$$\hat{\Theta}_{t,i,j} \triangleq \mathbf{J}_{\mathbf{x}^0}^{\mathbf{h}^g}\mathbf{J}_{\mathbf{x}^0}^{\mathbf{h}^g\top} \tag{47}$$

which gives a $N^g|\mathcal{D}|$ by $N^g|\mathcal{D}|$ kernel matrix. By comparison the empirical Fisher Information matrix with a Gaussian likelihood is

$$\sum_{i=1^{|\mathcal{D}|}} \mathbf{J}_{\theta}^{h^g\top}\nabla^2_{h^g}\mathcal{L}\,\mathbf{J}_{\theta}^{h^g} \tag{48}$$

To see that the spectra of these two coincide consider the third order tensor underlying both $\mathbf{J}^{\mathbf{h}^g}_{\mathbf{h}^1 i}$ for $i \in 1 \dots |\mathcal{D}|$, additionally consider and unfolding $\mathbf{A}$ with dimensions $|\theta|$ by $N^g|\mathcal{D}|$; i.e. we construct a matrix with dimension of number of parameters by number of outputs times number of data points. Then

$$\bar{\mathbf{G}} = \mathbf{A}^\top \mathbf{A} \tag{49}$$

$$\hat{\Theta} = \mathbf{A}\mathbf{A}^\top \tag{50}$$

$$\tag{51}$$

and their spectra trivially coincide.

Remark. *It is interesting to note that when the Fisher information metric and NTK are applied to a regression problem with Gaussian noise then the relation between admits the following interpretation. For $\mathcal{L} = \frac{1}{2}\|\hat{y} - y\|$ the Fisher information matrix $\bar{\mathbf{G}}$ is the Riemannian metric on the tangent bundle and $\hat{\Theta}$ is the Riemannian metric on the co-tangent bundle.*