The Natural Neural Tangent Kernel: Neural Network Training Dynamics under Natural Gradient Descent

Tim G. J. Rudner[*]
University of Oxford

Florian Wenzel
TU Kaiserslautern

Yee Whye Teh
University of Oxford

Yarin Gal
University of Oxford

1 Introduction

Gradient-based optimization methods have proven successful in learning complex, overparameterized neural networks from non-convex objectives. Yet, the precise theoretical relationship between gradient-based optimization methods, the resulting training dynamics, and generalization in deep neural networks (DNNs) remains unclear. Recent work has investigated the training dynamics of DNNs via a function-space perspective through the lens of the neural tangent kernel (NTK) [Jacot et al., 2018]. Follow-up work explored avenues to compute the solution to the differential equation described by the NTK approximately and analytically [Arora et al., 2019, Lee et al., 2019] at great computational cost and at the cost of exactness, and Lee et al. [2019] has highlighted the relationship between Gaussian processes with an NTK covariance function and deep ensemble uncertainty estimates [Lakshminarayanan et al., 2017], suggesting possible applications of the NTK as a tool to assess model uncertainty.

In this work, we investigate the training dynamics of overparameterized neural networks under natural gradient descent [Amari, 1998, Martens, 2014]. Taking a function-space view of the training dynamics, we give exact analytic solutions to the training dynamics on training points as well as to the training dynamics linearized around the parameters at initialization evaluated on any arbitrary input. Furthermore, we derive a bound on the discrepancy between the distributions over functions at the optimum of natural gradient descent and the distribution over functions under the analytic solution to the linearized natural gradient descent training dynamics.

2 Preliminaries

In this section, we will introduce the NTK of a deep neural network, describe natural gradient descent, and show how to estimate the Fisher information matrix of a DNN.

2.1 Training Dynamics under Gradient Descent and the Neural Tangent Kernel

We define the NTK for n data points as in Lee et al. [2019] by stacking data points on top of each other. Let $\mathcal{X} \in \mathbb{R}^{dn}$ be the concatenation of training points, $\mathcal{Y} \in \mathbb{R}^{n}$ the concatenation of training targets. We define the concatenated output of the DNN $f_\theta(\cdot)$ for all points as $f_\theta(\mathcal{X}) = \text{vec}\left(f_\theta(x)\right)_{x \in \mathcal{X}} \in \mathbb{R}^{kn}$, where θ are the network parameters and k is the number of output dimensions. We further define the concatenated loss $\mathcal{L}(f_\theta(\cdot)) : \mathbb{R}^{dn} \to \mathbb{R}$ as $\mathcal{L}(f_\theta(\mathcal{X})) = \sum_i \ell(f(x_i)$ and the concatenated likelihood as $\tilde{p}(\mathcal{Y}|f_\theta(\mathcal{X})) = \prod_i p(y_i|f(x_i))$. The gradient descent training dynamics of a DNN $f_\theta(\cdot)$ are then

[*] Corresponding author: tim.rudner@cs.ox.ac.uk.

4th Workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada.

given by

$$\dot{f}_{\theta_t}(\mathcal{X}) = \frac{\partial}{\partial t} f_{\theta_t}(\mathcal{X}) \overset{\text{def}}{=} -\eta \nabla_\theta f_{\theta_t}(\mathcal{X}) \nabla_\theta f_{\theta_t}(\mathcal{X})^\top \nabla_f \mathcal{L}(f_\theta(\mathcal{X})), \tag{1}$$

where η is the learning rate. The corresponding NTK under gradient descent is

$$\Theta_{\theta_t}(\mathcal{X}, \mathcal{X}) \overset{\text{def}}{=} \nabla_\theta f_{\theta_t}(\mathcal{X}) \nabla_\theta f_{\theta_t}(\mathcal{X})^\top. \tag{2}$$

The NTK allows us to study generalization of DNNs from a function-space perspective [Jacot et al., 2018] and has been applied to a range of prediction tasks [Arora et al., 2019, Lee et al., 2019]. Crucially, to achieve tractability, prior work has mostly considered infinitely-wide DNNs for which the NTK is initialized randomly but stays constant during training.

2.2 Natural Gradient Descent

Natural gradients have been successfully used to improve optimization speed in many applications [Galy-Fajou et al., 2019, Pascanu and Bengio, 2014, Wenzel et al., 2019] and have been shown to have desirable theoretical properties [Amari, 1998, Amari et al., 2019, Karakida et al., 2019, Martens, 2014, Yang and Amari, 1997, Zhang et al., 2019]. Consider a model with an output (target) distribution, $p(y|f_\theta(x))$ which is parameterized by a DNN, $f_\theta(x)$. Let $\mathcal{L}$ be a loss function which depends on $p(y|f_\theta(x))$. The natural gradient of $\mathcal{L}$ with respect to the set of model parameters θ is then given by

$$\hat{\nabla}_\theta \mathcal{L}(\theta) = -F(\theta)^{-1} \nabla_\theta \mathcal{L}(\theta), \tag{3}$$

where $F(\theta)$ is the Fisher information matrix, and $\hat{\nabla}_\theta$ denotes the natural gradient operator. In the following, we will assume the log-likelihood loss $\mathcal{L}(p(y|f_\theta(x))) = \sum_i \log p(y_i|f_\theta(x_i))$.

Using these definitions, we can now state the DNN training dynamics as a function of time under *natural gradient descent*. Following the natural parameter space training dynamics from Equation (3), the training dynamics for a DNN for training data points $\mathcal{X}$ and output dimension k is given by the differential equation

$$\frac{\partial}{\partial t}\theta_t = -\eta \hat{\nabla}_\theta \mathcal{L}(\theta_t) = -\eta F(\theta_t)^{-1} \nabla_\theta f_{\theta_t}(\mathcal{X})^\top \nabla_f \mathcal{L}(f_\theta(\mathcal{X})),$$

where we note that the DNN parameters are a function of time and define $\theta(t) \overset{\text{def}}{=} \theta_t$ for ease of notation. The last equality follows by the chain rule, since $\mathcal{L}(\theta_t)$ is a function of f_{θ_t}.

2.3 Estimating the Fisher Information Matrix

Following Pascanu and Bengio [2014], the *Fisher information matrix* of a DNN is given by $F(\theta_t) = \mathbb{E}_{p(x)}[F(\theta_t|x)]$, where

$$F(\theta_t|x) = \mathbb{E}_{p(y|f_{\theta_t}(x))}[\nabla_\theta \log p(y|f_{\theta_t}(x))^\top \nabla_\theta \log p(y|f_{\theta_t}(x))] \tag{4}$$

is the Fisher information matrix conditioned on a single data point x. The *empirical Fisher information matrix* $\widetilde{F}(\theta_t)$ is obtained by assuming the empirical distribution of the data $p(x) = p(\mathcal{X})$ leading to

$$\widetilde{F}(\theta_t) = \frac{1}{n} \nabla_\theta f_{\theta_t}(\mathcal{X})^\top \mathbb{E}_{\tilde{p}(y|f(\mathcal{X}))}[\nabla_f \log \tilde{p}(y|f(\mathcal{X}))^\top \nabla_f \log \tilde{p}(y|f(\mathcal{X}))] \nabla_\theta f_\theta(\mathcal{X}),$$

where we use the matrix identity $A^\top A = \sum_i A_{(i.)}^\top A_{(i.)}$. For a Gaussian likelihood with variance σ^2, the empirical Fisher information matrix simplifies to

$$\widetilde{F}(\theta_t) = \frac{1}{n} \nabla_\theta f_{\theta_t}(\mathcal{X})^\top \frac{1}{\sigma^2} \nabla_\theta f_{\theta_t}(\mathcal{X}) = \frac{1}{n\sigma^2} \nabla_\theta f_{\theta_t}(\mathcal{X})^\top \nabla_\theta f_{\theta_t}(\mathcal{X}). \tag{5}$$

A detailed derivation of the concatenated Fisher information matrix can be found in Appendix B.

3 Training Dynamics under Natural Gradient Descent

In this section, we will take a function-space view of natural gradient descent, introduce the *natural neural tangent kernel*, give exact analytic solutions to the training dynamics evaluated on training points and to the training dynamics linearized around the parameters at initialization evaluated on any arbitrary input, and derive a bound on the discrepancy between the distributions over functions at the optimum of natural gradient descent and the analytic solution to the natural gradient descent training dynamics.

3.1 The Natural Neural Tangent Kernel

As for the training dynamics under gradient descent, since $f_{\theta_t}(\mathcal{X})$ is a function of θ_t, which in turn is a function of time, we can express the function-space training dynamics under natural gradient descent in terms of the parameter-space training dynamics as follows:

$$\frac{\partial}{\partial t} f_{\theta_t}(\mathcal{X}) = -\eta \nabla_\theta f_{\theta_t}(\mathcal{X}) F(\theta_t)^{-1} \nabla_\theta f_{\theta_t}(\mathcal{X})^\top \nabla_f \mathcal{L}(f_\theta(\mathcal{X})). \tag{6}$$

Inspecting these dynamics more closely, we see that the concatenated tangent kernel of a DNN f_θ with p parameters evaluated on training points $\mathcal{X}$ is given by the $nk \times nk$-matrix

$$\Theta^{\text{nat}}_{\theta_t}(\mathcal{X}, \mathcal{X}) = \nabla_\theta f_{\theta_t}(\mathcal{X}) F(\theta_t)^{-1} \nabla_\theta f_{\theta_t}(\mathcal{X})^\top, \tag{7}$$

where $\nabla_\theta f_{\theta_t}(\mathcal{X})$ is a $nk \times p$-matrix. We will refer to this kernel as the *Natural Neural Tangent Kernel* (natural NTK). Following Jacot et al. [2018], we can thus express the function-space training dynamics under natural gradient descent on an arbitrary test data point x in terms of the natural NTK evaluated on x and the training points $\mathcal{X}$,

$$\Theta^{\text{nat}}_{\theta_t}(x, \mathcal{X}) = \nabla_\theta f_{\theta_t}(x) F(\theta_t)^{-1} \nabla_\theta f_{\theta_t}(\mathcal{X})^\top, \tag{8}$$

which yields the function-space training dynamics

$$\frac{\partial}{\partial t} f_{\theta_t}(\mathcal{X}) = -\eta \Theta^{\text{nat}}_{\theta_t}(x, \mathcal{X}) \nabla_f \mathcal{L}(f_\theta(\mathcal{X})). \tag{9}$$

Using the expression of the empirical Fisher information given in Equation (5) for a Gaussian likelihood with variance σ^2, the empirical natural NTK evaluated on an an arbitrary data point x and n training points $\mathcal{X}$ is given by

$$\hat{\Theta}^{\text{nat}}_{\theta_t}(x, \mathcal{X}) = n\sigma^2 \nabla_\theta f_{\theta_t}(x) \left(\nabla_\theta f_{\theta_t}(\mathcal{X})^\top \nabla_\theta f_{\theta_t}(\mathcal{X}) \right)^{-1} \nabla_\theta f_{\theta_t}(\mathcal{X})^\top. \tag{10}$$

A detailed derivation of the concatenated natural NTK can be found in Appendix C.[2]

3.2 Exact Solution to the Non-linearized Dynamics on Training Points

Gradient-based optimization of DNNs is scalable and can, under certain conditions, converge to a global optimum [Chizat and Bach, 2018, Du et al., 2019, Oymak and Soltanolkotabi, 2019, Zhang et al., 2019] but in practice requires significant fine-tuning of hyperparameters and is often slow to converge. In this section, we consider the function-space training dynamics under natural gradient descent on training as well on *test* points. In particular, we solve the training dynamics under natural gradient descent on training points *exactly* and follow prior work [Lee et al., 2019] in linearizing the training dynamics around the DNN parameters at initialization to make predictions on *test points*. In other words, our solution allows us to make predictions from a DNN trained until convergence via natural gradient descent at arbitrary points in the input space without actually performing natural gradient descent.

Assumption 1 (Network Overparameterization)**.** *Let $f_{\theta_t}(\cdot)$ be a* DNN *with $|\theta_t| = p \, \forall t$, the number of network parameters. Assume that the* DNN *is overparameterized, that is, $nk \leq p$.*

Assumption 2 (Positive Definiteness of the Gram Matrix)**.** *Let $nk \leq p$, and define the Jacobian at initialization and at time step t during training as $J_0(x) \overset{\text{def}}{=} \nabla_\theta f_{\theta_t}(x)\big|_{\theta_t=\theta_0} \in \mathbb{R}^{nk \times p}$ and $J_t(x) \overset{\text{def}}{=} \nabla_\theta f_{\theta_t}(x) \in \mathbb{R}^{k \times p}$, respectively, and let $G_0(x) = J_0(x) J_0(x)^\top \in \mathbb{R}^{k \times k}$ and $G_t(x) = J_t(x) J_t(x)^\top \in \mathbb{R}^{k \times k}$ be the corresponding Gram matrices. Assume that $G_t(x) = J_t(x) J_t(x)^\top$ is positive definite for all $t \geq 0$.*

[2] We note that the natural NTK, $\hat{\Theta}^{\text{nat}}_{\theta_t}$, is a valid kernel function. To see that it is, note that (i) Jacot et al. [2018] showed that the NTK induced by gradient descent is a valid kernel function and that (ii) the inverse Gram matrix $\left(\nabla_\theta f_{\theta_t}(\mathcal{X})^\top \nabla_\theta f_{\theta_t}(\mathcal{X}) \right)^{-1}$ in Equation (10) can be evaluated on *any* set of points from the empirical distribution of the data and does not necessarily have to be evaluated on the full dataset $\mathcal{X}$ and, more specifically, does not have to be evaluated on the data points of the second argument of the natural NTK. As a result, the natural NTK is symmetric in its arguments, and all arguments for the validity of the NTK under gradient descent given in [Jacot et al., 2018] carry over to the natural NTK.

In the setting $nk = p$, the natural NTK evaluated on the training points $\mathcal{X}$ simplifies to the (scaled) identity matrix

$$\Theta^{\text{nat}} \overset{\text{def}}{=} \hat{\Theta}_t^{\text{nat}}(\mathcal{X}, \mathcal{X}) = n\sigma^2 I_{nk}. \tag{11}$$

For $nk < p$, the Fisher matrix is almost surely singular but can be computed (non-uniquely) via the generalized inverse [Bernacchia et al., 2018],

$$\widetilde{F}^{\dagger}(\theta_t) = n\sigma^2 \nabla_\theta f_{\theta_t}(\mathcal{X})^\top G_t^{-1}(\mathcal{X}) G_t^{-1}(\mathcal{X}) \nabla_\theta f_{\theta_t}(\mathcal{X}), \tag{12}$$

where $G_t(\mathcal{X}) \overset{\text{def}}{=} \nabla_\theta f_{\theta_t}(\mathcal{X}) \nabla_\theta f_{\theta_t}(\mathcal{X})^\top \in \mathbb{R}^{nk \times nk}$, and thus all terms involving $\mathcal{X}$ cancel out and we again get

$$\Theta^{\text{nat}} \overset{\text{def}}{=} \hat{\Theta}_t^{\text{nat}}(\mathcal{X}, \mathcal{X}) = n\sigma^2 I_{nk}. \tag{13}$$

For the remainder of this paper, we assume the natural NTK in the overparameterized setting is computed using the generalized inverse so that

$$\hat{\Theta}_{\theta_t}^{\text{nat}}(x, \mathcal{X}) = n\sigma^2 \nabla_\theta f_{\theta_t}(x) \nabla_\theta f_{\theta_t}(\mathcal{X})^\top \left(\nabla_\theta f_{\theta_t}(\mathcal{X}) \nabla_\theta f_{\theta_t}(\mathcal{X})^\top \right)^{-1}. \tag{14}$$

Proposition 1 (Solution to Natural Gradient Descent Dynamics on Training Points)**.** *Under Assumption 1 and under MSE loss, the function-space training dynamics* (6) *under natural gradient descent on the training points $\mathcal{X}$, $\dot{f}_{\theta_t}(\mathcal{X})$, are linear in f_{θ_t} and can be solved analytically*

$$f_{\theta_t}(\mathcal{X}) = \left(I - e^{-\frac{\eta}{n}\Theta^{\text{nat}} \cdot t} \right) \mathcal{Y} + e^{-\frac{\eta}{n}\Theta^{\text{nat}} \cdot t} f_{\theta_0}(\mathcal{X}). \tag{15}$$

Proof. See Appendix D. □

We note that Proposition 1 deals with the stylized case of infinitesimally small step sizes. Zhang et al. [2019] consider regular step sizes and show that, under certain regularity and Lipschitzness assumptions, natural gradient descent can be shown to exhibit fast convergence to a global optimum. For the remainder of this paper, we will focus on the function space defined by the solution of the DNN training dynamics under natural gradient descent instead of the training dynamics on the training points only.

3.3 Exact Solution to the Linearized Dynamics on Training and Test Points

In the previous section, we showed that for overparameterized DNNs, the natural gradient descent training dynamics on the training points can be solved analytically. Unfortunately, on a test point x, the training dynamics under natural gradient descent,

$$\dot{f}_{\theta_t}(\mathcal{X}) = -\eta \hat{\Theta}_t^{\text{nat}}(x, \mathcal{X}) \nabla_f \mathcal{L}(f_{\theta_t}(\mathcal{X})), \tag{16}$$

$$\text{with} \quad \hat{\Theta}_t^{\text{nat}}(x, \mathcal{X}) = n\sigma^2 \nabla_\theta f_{\theta_t}(x) \nabla_\theta f_{\theta_t}(\mathcal{X})^\top \left(\nabla_\theta f_{\theta_t}(\mathcal{X}) \nabla_\theta f_{\theta_t}(\mathcal{X})^\top \right)^{-1}$$

are not linear in f_{θ_t} anymore, since the NTK is now evaluated on training and test points and does thus not cancel out with the inverse. As a result, the natural NTK does not reduce to the scaled identity matrix and continues to depend on f_{θ_t} non-linearly.

In order to solve for the function-space defined by the DNN on any input point at any time during training, we follow Lee et al. [2019] and assume a linear evolution of the DNN parameters:

Assumption 3 (Linearization)**.** *Assume that $f_{\theta_t}^{\text{lin}}(x) \overset{\text{def}}{=} f_{\theta_0}(x) + \nabla_\theta f_{\theta_t}(x)\big|_{\theta_t = \theta_0}(\theta_t - \theta_0)$.*

Corollary 1. *Throughout training via natural gradient descent, the solution to the linearized function-space training dynamics evaluated on the training points, $\dot{f}_{\theta_t}^{\text{lin}}(\mathcal{X})$, is identical to the solution to the non-linearized function-space training dynamics on the training points, $\dot{f}_{\theta_t}(\mathcal{X})$, that is, $f_{\theta_t}^{\text{lin}}(\mathcal{X}) = f_{\theta_t}(\mathcal{X})$.*

Proof. See Appendix D. □

Proposition 2 (Solution to Linearized Natural Gradient Descent Dynamics on a Test Point)**.** *Under Assumption 1 and under MSE loss, the linearized function-space training dynamics under natural gradient descent on a test point x, $\dot{f}_{\theta_t}^{\text{lin}}(x)$, can be solved as*

$$f_{\theta_t}^{\text{lin}}(x) = f_{\theta_0}(x) - \frac{1}{n} \hat{\Theta}_0^{\text{nat}}(x, \mathcal{X}) \left(I - e^{-\eta\sigma^2 t} \right) f_{\theta_0}(\mathcal{X}) + \frac{1}{n} \hat{\Theta}_0^{\text{nat}}(x, \mathcal{X}) \left(I - e^{-\eta\sigma^2 t} \right) \mathcal{Y}, \tag{17}$$

$$\textit{with} \quad \hat{\Theta}_0^{\text{nat}}(x, \mathcal{X}) = n\sigma^2 \nabla_\theta f_{\theta_t}(x)\big|_{\theta_t = \theta_0} \nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t = \theta_0}^\top \left(\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t = \theta_0} \nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t = \theta_0}^\top \right)^{-1}.$$

Proof. See Appendix D. □

Corollary 2. *In the limit of training time, as $t \to \infty$, the solution to the linearized training dynamics under natural gradient descent tends to*

$$\lim_{t\to\infty} f^{\text{lin}}_{\theta_t}(x) = f_{\theta_0}(x) + \frac{1}{n}\hat{\Theta}^{\text{nat}}_0(x,\mathcal{X})\left(\mathcal{Y} - f_{\theta_0}(\mathcal{X})\right),$$

$$\textit{with} \quad \hat{\Theta}^{\text{nat}}_0(x,\mathcal{X}) = n\sigma^2 \nabla_\theta f_{\theta_t}(x)\big|_{\theta_t=\theta_0} \nabla_\theta f_{\theta_t}(\mathcal{X})\big|^\top_{\theta_t=\theta_0}\left(\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}\nabla_\theta f_{\theta_t}(\mathcal{X})\big|^\top_{\theta_t=\theta_0}\right)^{-1}.$$

In the limit of training time, as $t \to \infty$, $f^{\text{lin}}_{\theta_t}(x)$ can be thought of as a perturbation of the DNN at initialization (evaluated on x). Corollary 2 states that this perturbation is given by a linear transformation defined by the natural neural tangent kernel at initialization evaluated on a test point and all training points applied to the difference between the function values of the DNN at initialization and the true targets.[3]

From Corollary 2, we see that, at convergence the solution to the linearized training dynamics under natural gradient descent is identical to the solution to the linearized training dynamics of gradient descent derived in Lee et al. [2019]. Consequently, if the linearized training dynamics under gradient descent and natural gradient descent result in the same predictive distribution at convergence, this suggests that if we can relate the predictive distribution under natural gradient descent and linearized natural gradient descent at convergence, we may be able to shed light on the DNN predictions under converged gradient descent.

4 Discrepancy between Predictions under Linearized and True Dynamics

While prior work (e.g., Lee et al. [2019]) justified linearizing the DNN training dynamics around the parameters at initialization by showing that, in the limit of infinitely wide hidden layers, the NTK stays constant during training and the discrepancy between predictions from a linearized DNN match those of a non-linearized DNN trained with gradient descent, we make no infinite width assumptions. Instead, we derive an upper bound on the discrepancy between predictions on arbitrary points in input space from a DNN trained via natural gradient descent and the analytical solution to the linearized dynamics which varies with the the DNN architecture, initialization, its activation functions and hyperparameters, the resulting training dynamics (via spectral norms dependent on s), and the data.

Lemma 1 (Natural Neural Tangent Kernel Bound). *Let $\hat{\Theta}^{\text{nat}}_0(x,\mathcal{X})$ and $\hat{\Theta}^{\text{nat}}_t(x,\mathcal{X})$ be the natural* NTK *at initialization and at some time step t during training, respectively, and let $\lambda_{\max}(G_t(x))$ be the largest eigenvalue of the Gram matrix for $t \geq 0$. Under Assumption 1, for any random initialization,*

$$\frac{1}{n\sigma^2}||\hat{\Theta}^{\text{nat}}_0(x,\mathcal{X}) - \hat{\Theta}^{\text{nat}}_t(x,\mathcal{X})||_2 \leq \sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_0(\mathcal{X}))}} + \sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_t(\mathcal{X}))}} + \frac{||J_0(x) - J_s(x)||_2}{\sqrt{\lambda_{\min}(G_t(\mathcal{X}))}} \tag{18}$$

Proof. See Appendix E. □

In order to get guarantees for the quality of the predictions from the solution to the linearized training dynamics, we consider the discrepancy between such predictions and predictions from a DNN trained via natural gradient descent. Theorem 1 establishes an upper bound on this discrepancy and guarantees that, for any random initialization, the predictions from the linearized model are in some neighborhood of the predictions of the true optimum of the trained DNN.

Theorem 1 (Prediction Error under Linearized Training Dynamics). *Let (x, y) be an input-output test set pair, $\mathcal{X}$ and $\mathcal{Y}$ the training input and output sets, respectively, η the learning rate, σ^2 the variance of the Gaussian likelihood, $f^{\text{lin}}_{\theta_t}(x)$ the function predictions from the analytical solution to the linearized training dynamics under natural gradient descent (a random variable), and $f_{\theta_t}(x)$ the function predictions obtained from running natural gradient descent (also a random variable).*

[3] Following Lee et al. [2019], we note that for discrete training dynamics, $\frac{\partial}{\partial t}\theta_t$ and $\frac{\partial}{\partial t}f_{\theta_t}(\cdot)$ above would be replaced by $\theta_{t+1} - \theta_t$ and $f_{\theta_{t+1}}(\mathcal{X}) - f_{\theta_t}(\mathcal{X})$, respectively, and every $e^{-\frac{\eta}{n}\Theta^{\text{nat}}\cdot t}$ term would be replaced by $I - \left(I - \frac{\eta}{n}\Theta^{\text{nat}}\right)^t$.

Under Assumption 1 and under MSE loss, for $g_{\theta_t}(x) = f_{\theta_t}(x) - y$, the spectral norm $|| \cdot ||_2$, and the natural NTK *at initialization, $\hat{\Theta}_0^{\text{nat}}(x, \mathcal{X})$, and at time t, $\hat{\Theta}_t^{\text{nat}}(x, \mathcal{X})$, for any random initialization,*

$$
\begin{aligned}
||g_{\theta_t}^{\text{lin}}(x) - g_{\theta_t}(x)||_2 \leq ||f_{\theta_0}(\mathcal{X}) - \mathcal{Y}||_2 \Bigg(& \sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_0(\mathcal{X}))}} \left(I - e^{-\eta\sigma^2 t} \right) \\
& + \eta\sigma^2 \int_0^t \left(\sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_s(\mathcal{X}))}} + \frac{||J_0(x) - J_s(x)||_2}{\sqrt{\lambda_{\min}(G_s(\mathcal{X}))}} \right) e^{-\eta\sigma^2 s} \mathrm{d}s \Bigg),
\end{aligned}
\tag{19}
$$

Proof. See Appendix E. □

Corollary 3 (Prediction Error under Linearized Training Dynamics at Convergence)**.** *In the limit of training time, as $t \to \infty$, where $\lim_{t\to\infty} g_{\theta_t}(x) = g_\theta^\star(x)$, is a function prediction when natural gradient descent has reached an optimum, for any random initialization,*

$$
\begin{aligned}
\lim_{t\to\infty} ||g_{\theta_t}^{\text{lin}}(x) - g_\theta^\star(x)||_2 \leq ||f_{\theta_0}(\mathcal{X}) - \mathcal{Y}||_2 \Bigg(& \sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_0(\mathcal{X}))}} \\
& + \eta\sigma^2 \lim_{t\to\infty} \int_0^t \left(\sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_s(\mathcal{X}))}} + \frac{||J_0(x) - J_s(x)||_2}{\sqrt{\lambda_{\min}(G_s(\mathcal{X}))}} \right) e^{-\eta\sigma^2 s} \mathrm{d}s \Bigg).
\end{aligned}
\tag{20}
$$

As can be seen from Corollary 3, the discrepancy between predictions under the linearized and the true dynamics $||g_{\theta_t}^{\text{lin}}(x) - g_{\theta_t}^\star(x)||_2$ will be smallest when $\sqrt{\lambda_{\max}(G_0(x))/\lambda_{\min}(G_0(\mathcal{X}))}$ is close to unity and the ratio $||J_0(x) - J_s(x)||_2/\sqrt{\lambda_{\min}(G_s(\mathcal{X}))}$ is minimal throughout training.

These bounds are functions of the DNN architecture and hyperparameters, the initialization scheme, and the data and will be loose for most DNNs used in practice. While we know that the bounds are tight in the infinite-width limit, future research may be able to elucidate for which finite-width DNN architectures and initialization schemes linearized and non-linearized training dynamics converge to similar distributions over functions.

5 Conclusion

We presented upper bounds on the discrepancy between predictions obtained from the solution to the linearized training dynamics under natural gradient descent and predictions under non-linearized training dynamics as a function of the DNN architecture and hyperparameters, the initialization scheme, and the data.

References

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural Comput.*, 10(2):251–276, February 1998. ISSN 0899-7667. doi: 10.1162/089976698300017746. URL http://dx.doi.org/10.1162/089976698300017746.

Shun-ichi Amari, Ryo Karakida, and Masafumi Oizumi. Fisher information and natural gradient learning in random deep networks. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 694–702. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/amari19a.html.

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net, 2019.

Alberto Bernacchia, Mate Lengyel, and Guillaume Hennequin. Exact natural gradient in deep linear networks and its application to the nonlinear case. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 5941–5950. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7834-exact-natural-gradient-in-deep-linear-networks-and-its-application-to-the-nonlinear-case.pdf.

Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 3040–3050, USA, 2018. Curran Associates Inc. URL http://dl.acm.org/citation.cfm?id=3327144.3327226.

Simon S. Du, Xiyu Zhai, Barnabás Póczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. URL https://openreview.net/forum?id=S1eK3i09YQ.

Théo Galy-Fajou, Florian Wenzel, Christian Donner, and Manfred Opper. Multi-class gaussian process classification made conjugate: Efficient inference via data augmentation. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial*, 2019.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 8571–8580. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/8076-neural-tangent-kernel-convergence-and-generalization-in-neural-networks.pdf.

Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal statistics of fisher information in deep neural networks: Mean field approach. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1032–1041. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/karakida19a.html.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent. *arXiv e-prints*, art. arXiv:1902.06720, Feb 2019.

James Martens. New insights and perspectives on the natural gradient method, 2014.

Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4951–4960, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/oymak19a.html.

Razvan Pascanu and Yoshua Bengio. Revisiting natural gradient for deep networks. In *International Conference on Learning Representations*, 2014.

Florian Wenzel, Théo Galy-Fajou, Christian Donner, Marius Kloft, and Manfred Opper. Efficient gaussian process classification using pólya-gamma data augmentation. In *AAAI Conference on Artificial Intelligence*, 2019.

Howard Hua Yang and Shun-Ichi Amari. The efficiency and the robustness of natural gradient descent learning rule. In *Proceedings of the 10th International Conference on Neural Information Processing Systems*, NIPS'97, pages 385–391, Cambridge, MA, USA, 1997. MIT Press. URL http://dl.acm.org/citation.cfm?id=3008904.3008959.

Guodong Zhang, James Martens, and Roger Grosse. Fast convergence of natural gradient descent for overparameterized neural networks. In *arXiv*, 2019.

Supplementary Material

Appendix A Natural Gradient Descent

Consider a model of an output (target) distribution, $p(y|f_\theta(x))$, parameterized by a deep neural network (DNN), $f_\theta(x)$, and let $\mathcal{L}$ be a loss function which depends on $p(y|f_\theta(x))$. The natural gradient of $\mathcal{L}$ with respect to the set of parameters θ is then given by

$$\hat{\nabla}_\theta \mathcal{L}(\theta) = -\eta F(\theta)^{-1} \nabla_\theta \mathcal{L}(\theta), \tag{21}$$

where η is the learning rate, $F(\theta)$ is the Fisher information matrix, and $\hat{\nabla}_\theta$ is the natural gradient operator. In the following, we will assume the log-likelihood loss $\mathcal{L}(p(y|f_\theta(x))) = \sum_i \log p(y_i|f_\theta(x_i))$.

To obtain the Fisher information matrix $F(\theta)$, we first consider the Fisher information matrix conditioned on a data point x, which, by the chain rule, is given by

$$\begin{aligned} F(\theta|x) &= \mathbb{E}_{p(y|f_\theta(x))}[\nabla_\theta \log p(y|f_\theta(x))^\top \nabla_\theta \log p(y|f_\theta(x))] \\ &= \nabla_\theta f_\theta(x)^\top \hat{F}(f_\theta(x)) \nabla_\theta f_\theta(x), \end{aligned} \tag{22}$$

where

$$\hat{F}(f_\theta(x)) \overset{\text{def}}{=} \mathbb{E}_{p(y|f_\theta(x))}\left[\nabla_f \log p(y|f_\theta(x))^\top \nabla_f \log p(y|f_\theta(x))\right] \tag{23}$$

is the Fisher information matrix of $f_\theta(x)$ under the likelihood $p(y|f_\theta(x))$. We note that for a Gaussian likelihood $p(y|f_\theta(x)) = \mathcal{N}(y|f_\theta(x), \sigma^2)$, we have $\hat{F}(f_\theta(x)) = \frac{1}{\sigma^2}$.

To obtain a Fisher information matrix that is independent of x, we follow Pascanu and Bengio [2014] and introduce a distribution over the inputs, $p(x)$, which from hereon we will assume to be the empirical distribution of the data. This way, we are able to express the empirical Fisher information of the DNN parameters as

$$\widetilde{F}(\theta) = \mathbb{E}_{p(x)}[F(\theta|x)] = \frac{1}{n}\sum_i \nabla_\theta f_\theta(x_i)^\top \hat{F}(f_\theta(x_i)) \nabla_\theta f_\theta(x_i). \tag{24}$$

Using a Gaussian likelihood with variance σ^2 (which, for maximum-likelihood estimation, is equivalent to minimization under MSE loss), this expression simplifies to

$$\widetilde{F}(\theta) = \mathbb{E}_{p(x)}[F(\theta|x)] = \frac{1}{n\sigma^2}\sum_i \nabla_\theta f_\theta(x_i)^\top \nabla_\theta f_\theta(x_i). \tag{25}$$

Note that, in practice, $\hat{F}(f_\theta(x))$ can also be estimated using the empirical distribution, yielding a biased approximation of the Fisher information matrix given by

$$\widetilde{\widetilde{F}}(\theta) = \frac{1}{n}\sum_i \nabla_\theta f_\theta(x_i)^\top \nabla_f \log p(y_i|f_\theta(x_i))^\top \nabla_f \log p(y_i|f_\theta(x_i)) \nabla_\theta f_\theta(x_i). \tag{26}$$

Appendix B Derivation of the concatenated Fisher information matrix

First, consider the Fisher information matrix conditioned on a single data point, x, as before:

$$F(\theta|x) = \mathbb{E}_{p(y|f_\theta(x))}[\nabla_\theta \log p(y|f_\theta(x))^\top \nabla_\theta \log p(y|f_\theta(x))].$$

The empirical Fisher information matrix with the concatenated likelihood, $\tilde{p}(\mathcal{Y}|f_\theta(\mathcal{X})) = \prod_i p(y_i|f(x_i))$, is

$$\begin{aligned} \widetilde{F}(\theta) &= \frac{1}{n}\sum_i F(\theta|x_i) && (27) \\ &= \frac{1}{n}\sum_i \mathbb{E}_{p(y_i|f_\theta(x_i))}[\nabla_\theta \log p(y_i|f_\theta(x_i))^\top \nabla_\theta \log p(y_i|f_\theta(x_i))] && (28) \\ &= \frac{1}{n}\mathbb{E}_{\tilde{p}(\mathcal{Y}|f(\mathcal{X}))}[\sum_i \nabla_\theta \log p(y_i|f_\theta(x_i))^\top \nabla_\theta \log p(y_i|f_\theta(x_i))] && (29) \\ &= \frac{1}{n}\mathbb{E}_{\tilde{p}(\mathcal{Y}|f(\mathcal{X}))}[\nabla_\theta \log \tilde{p}(\mathcal{Y}|f(\mathcal{X}))^\top \nabla_\theta \log \tilde{p}(\mathcal{Y}|f(\mathcal{X}))] && (30) \\ &= \frac{1}{n}\nabla_\theta f_\theta(\mathcal{X})^\top \underbrace{\mathbb{E}_{\tilde{p}(\mathcal{Y}|f(\mathcal{X}))}\left[\nabla_f \log \tilde{p}(\mathcal{Y}|f(\mathcal{X}))^\top \nabla_f \log \tilde{p}(\mathcal{Y}|f(\mathcal{X}))\right]}_{\overset{\text{def}}{=}\hat{F}(f_\theta(\mathcal{X}))} \nabla_\theta f_\theta(\mathcal{X}) && (31) \\ &= \frac{1}{n}\nabla_\theta f_\theta(\mathcal{X})^\top \hat{F}(f_\theta(\mathcal{X})) \nabla_\theta f_\theta(\mathcal{X}), && (32) \end{aligned}$$

where we used the matrix identity $A^\top A = \sum_i A_{(i.)}^\top A_{(i.)}$. As before, using a Gaussian likelihood with variance σ^2, $\hat{F}(f_\theta(x)) = \frac{1}{\sigma^2}$, leads to the empirical Fisher information matrix

$$\begin{aligned}\widetilde{F}(\theta) &= \frac{1}{n}\nabla_\theta f_\theta(\mathcal{X})^\top \frac{1}{\sigma^2}\nabla_\theta f_\theta(\mathcal{X}) && (33)\\ &= \frac{1}{n\sigma^2}\nabla_\theta f_\theta(\mathcal{X})^\top \nabla_\theta f_\theta(\mathcal{X}). && (34)\end{aligned}$$

Appendix C Functional Gradient Descent and the Natural Neural Tangent Kernel

We will now derive the natural NTK from natural gradient descent on the loss $\mathcal{L}$ for the case of a single data point. Following Equation (3), the evolution of the parameters θ and the corresponding DNN output $f_\theta(x)$ under continuous-time natural gradient descent with learning rate η,

$$\begin{aligned}\dot{\theta}_t &= -\eta\hat{\nabla}_\theta \mathcal{L}(\theta) && (35)\\ \dot{f}_{\theta_t}(x) &= \frac{\partial}{\partial\theta} f_{\theta_t}(x)\frac{\partial}{\partial t}\theta_t, && (36)\end{aligned}$$

then become

$$\begin{aligned}\dot{\theta}_t = \frac{\partial\theta_t}{\partial t} &= -\eta F^{-1}(\theta)\nabla_\theta\mathcal{L}(\theta) && (37)\\ &= -\eta F^{-1}(\theta)\nabla_\theta f_{\theta_t}(x')^\top \nabla_f \mathcal{L}(f_{\theta_t}(x')) && (38)\\ \dot{f}_{\theta_t}(x) = \frac{\partial f_{\theta_t}}{\partial t} &= \frac{\partial}{\partial\theta} f_{\theta_t}(x)\frac{\partial}{\partial t}\theta_t && (39)\\ &= -\eta\nabla_\theta f_{\theta_t}(x) F^{-1}(\theta)\nabla_\theta f_{\theta_t}(x')^\top \nabla_f\mathcal{L}(f_{\theta_t}(x')). && (40)\end{aligned}$$

The natural NTK at time t is then given by

$$\Theta_{\theta_t}^{\text{nat}}(x, x') = \nabla_\theta f_{\theta_t}(x) F^{-1}(\theta_t)\nabla_\theta f_{\theta_t}(x')^\top, \qquad (41)$$

which describes the function-space training dynamics induced by optimizing the DNN parameters via natural gradient descent on a single data point x'.

For a single training point x from the empirical distribution, using Equation (24), we obtain

$$\begin{aligned}\Theta_{\theta_t}^{\text{nat}}(x, x') &= \nabla_\theta f_{\theta_t}(x) F^{-1}(\theta_t)\nabla_\theta f_{\theta_t}(x')^\top && (42)\\ &= \nabla_\theta f_{\theta_t}(x)\left(\nabla_\theta f_{\theta_t}(x')^\top \hat{F}(f_{\theta_t}(x'))\nabla_\theta f_{\theta_t}(x')\right)^{-1}\nabla_\theta f_{\theta_t}(x')^\top, && (43)\end{aligned}$$

and noting that, as before, under a Gaussian likelihood $\hat{F}(f_\theta(x)) = \frac{1}{\sigma^2}$, the expression simplifies to

$$\Theta_{\theta_t}^{\text{nat}}(x, x') = \sigma^2\nabla_\theta f_{\theta_t}(x)\left(\nabla_\theta f_{\theta_t}(x')^\top \nabla_\theta f_{\theta_t}(x')\right)^{-1}\nabla_\theta f_{\theta_t}(x')^\top. \qquad (44)$$

Next, for a set of n training points $\mathcal{X}$, we can express the training dynamics above as

$$\begin{aligned}\dot{\theta}_t = \frac{\partial\theta_t}{\partial t} &= -\eta F^{-1}(\theta)\nabla_\theta\mathcal{L}(\theta) && (45)\\ &= -\eta F^{-1}(\theta)\nabla_\theta f_{\theta_t}(\mathcal{X})^\top \nabla_f \mathcal{L}(f_{\theta_t}(\mathcal{X})) && (46)\\ \dot{f}_{\theta_t}(x) = \frac{\partial f_{\theta_t}}{\partial t} &= \frac{\partial}{\partial\theta} f_{\theta_t}(x)\frac{\partial}{\partial t}\theta_t && (47)\\ &= -\eta\nabla_\theta f_{\theta_t}(x) F^{-1}(\theta)\nabla_\theta f_{\theta_t}(\mathcal{X})^\top \nabla_f\mathcal{L}(f_{\theta_t}(\mathcal{X})). && (48)\end{aligned}$$

The natural NTK at time t is then given by

$$\Theta_{\theta_t}^{\text{nat}}(x, \mathcal{X}) = \nabla_\theta f_{\theta_t}(x) F^{-1}(\theta_t)\nabla_\theta f_{\theta_t}(\mathcal{X})^\top, \qquad (49)$$

which describes the function-space training dynamics induced by optimizing the DNN parameters via natural gradient descent on a set of training points $\mathcal{X}$.

For a set of training points $\mathcal{X}$ from the empirical distribution, using Equation (27), we obtain

$$\begin{aligned}\Theta_{\theta_t}^{\text{nat}}(x, \mathcal{X}) &= \nabla_\theta f_{\theta_t}(x) F^{-1}(\theta_t)\nabla_\theta f_{\theta_t}(\mathcal{X})^\top && (50)\\ &= \nabla_\theta f_{\theta_t}(x)\left(\nabla_\theta f_{\theta_t}(x')^\top \hat{F}(f_{\theta_t}(\mathcal{X}))\nabla_\theta f_{\theta_t}(\mathcal{X})\right)^{-1}\nabla_\theta f_{\theta_t}(\mathcal{X})^\top, && (51)\end{aligned}$$

and noting that, as before, under a Gaussian likelihood $\hat{F}(f_\theta(x)) = \frac{1}{\sigma^2}$, the expression simplifies to

$$\Theta_{\theta_t}^{\text{nat}}(x, \mathcal{X}) = \sigma^2\nabla_\theta f_{\theta_t}(x)\left(\nabla_\theta f_{\theta_t}(\mathcal{X})^\top \nabla_\theta f_{\theta_t}(\mathcal{X})\right)^{-1}\nabla_\theta f_{\theta_t}(\mathcal{X})^\top. \qquad (52)$$

Appendix D An Analytic Solution to Natural Gradient Descent

Assumption 1 (Network Overparameterization)**.** *Let $f_{\theta_t}(\cdot)$ be a* DNN *with $|\theta_t| = p$, the number of network parameters. Assume that $nk \leq p$, that is, the* DNN *is overparameterized.*

The empirical natural NTK is given by

$$\hat{\Theta}_t^{\text{nat}}(x, \mathcal{X}) = n\sigma^2 \nabla_\theta f_{\theta_t}(x) \left(\nabla_\theta f_{\theta_t}(\mathcal{X})^\top \nabla_\theta f_{\theta_t}(\mathcal{X}) \right)^{-1} \nabla_\theta f_{\theta_t}(\mathcal{X})^\top. \tag{53}$$

If the Jacobian is square and the natural NTK is evaluated on the training points $\mathcal{X}$, the natural NTK is equal to the (scaled) identity, that is,

$$\Theta^{\text{nat}} \overset{\text{def}}{=} \hat{\Theta}_t^{\text{nat}}(\mathcal{X}, \mathcal{X}) = n\sigma^2 I_{nk}, \tag{54}$$

For $nk < p$, the Fisher matrix is almost surely singular but can be computed via the generalized inverse [Bernacchia et al., 2018],

$$\widetilde{F}^\dagger(\theta_t) = n\sigma^2 \nabla_\theta f_{\theta_t}(\mathcal{X})^\top G_t^{-1}(\mathcal{X}) G_t^{-1}(\mathcal{X}) \nabla_\theta f_{\theta_t}(\mathcal{X}), \tag{55}$$

where $G_t(\mathcal{X}) \overset{\text{def}}{=} \nabla_\theta f_{\theta_t}(\mathcal{X}) \nabla_\theta f_{\theta_t}(\mathcal{X})^\top \in \mathbb{R}^{nk \times nk}$, and thus

$$\begin{aligned}
\Theta^{\text{nat}} \overset{\text{def}}{=} \hat{\Theta}_t^{\text{nat}}(\mathcal{X}, \mathcal{X}) &= \nabla_\theta f_{\theta_t}(\mathcal{X}) \left(n\sigma^2 \nabla_\theta f_{\theta_t}(\mathcal{X})^\top G_t^{-1}(\mathcal{X}) G_t^{-1}(\mathcal{X}) \nabla_\theta f_{\theta_t}(\mathcal{X}) \right) \nabla_\theta f_{\theta_t}(\mathcal{X})^\top && (56) \\
&= n\sigma^2 G_t(\mathcal{X}) G_t^{-1}(\mathcal{X}) G_t^{-1}(\mathcal{X}) G_t(\mathcal{X}) && (57) \\
&= n\sigma^2 I_{nk}. && (58)
\end{aligned}$$

Proposition 3 (Solution to Natural Gradient Descent Dynamics on Training Points)**.** *Under Assumption 1 and under MSE loss, the function-space training dynamics under natural gradient descent on the training points $\mathcal{X}$, $\dot{f}_{\theta_t}(\mathcal{X})$, are linear in f_{θ_t} and can be solved as*

$$f_{\theta_t}(\mathcal{X}) = \left(I - e^{-\frac{\eta}{n} \Theta^{\text{nat}} \cdot t} \right) \mathcal{Y} + e^{-\frac{\eta}{n} \Theta^{nat} \cdot t} f_{\theta_0}(\mathcal{X}). \tag{59}$$

Proof. Under Assumption 1, MSE loss, and $nk \leq p$, the function-space training dynamics,

$$\begin{aligned}
\frac{\partial f_{\theta_t}(\mathcal{X})}{\partial t} = \dot{f}_{\theta_t}(\mathcal{X}) &= -\eta \underbrace{\left(n\sigma^2 \nabla_\theta f_{\theta_t}(x) \nabla_\theta f_{\theta_t}(\mathcal{X})^\top \left(\nabla_\theta f_{\theta_t}(\mathcal{X}) \nabla_\theta f_{\theta_t}(\mathcal{X})^\top \right)^{-1} \right)}_{=\Theta^{\text{nat}}} \nabla_f \mathcal{L}(f_{\theta_t}(\mathcal{X})) && (60) \\
&= -\eta n\sigma^2 \nabla_f \mathcal{L}(f_{\theta_t}(\mathcal{X})), && (61)
\end{aligned}$$

become linear in $f_{\theta_t}(\cdot)$ under MSE loss and can therefore be solved analytically without making any linearization assumptions.

To solve the training dynamics, consider $\dot{f}_{\theta_t}(\mathcal{X})$. For MSE loss, we get

$$\begin{aligned}
\dot{f}_{\theta_t}(\mathcal{X}) &= -\eta \, \Theta^{\text{nat}} \nabla_f \frac{1}{2n} ||f_{\theta_t}(\mathcal{X}) - \mathcal{Y}||_2^2 && (62) \\
&= -\eta \, \Theta^{\text{nat}} \nabla_f \frac{1}{2n} (f_{\theta_t}(\mathcal{X})^\top f_{\theta_t}(\mathcal{X}) - 2 f_{\theta_t}(\mathcal{X})^\top \mathcal{Y} - \mathcal{Y}^\top \mathcal{Y}) && (63) \\
&= -\frac{\eta}{n} \, \Theta^{\text{nat}} (f_{\theta_t}(\mathcal{X}) - \mathcal{Y}) && (64) \\
&= -\frac{\eta}{n} \, \Theta^{\text{nat}} f_{\theta_t}(\mathcal{X}) + \frac{\eta}{n} \, \Theta^{\text{nat}} \mathcal{Y}. && (65)
\end{aligned}$$

Rearranging,

$$\frac{\partial f_{\theta_t}(\mathcal{X})}{\partial t} + \frac{\eta}{n} \, \Theta^{\text{nat}} f_{\theta_t}(\mathcal{X}) = \frac{\eta}{n} \, \Theta^{\text{nat}} \mathcal{Y}, \tag{66}$$

we see that this is a first-order linear ordinary differential equation and can be solved analytically and has solution

$$f_{\theta_t}(\mathcal{X}) = \mathcal{Y} + e^{-\frac{\eta}{n} \, \Theta^{\text{nat}} \cdot t} \, c. \tag{67}$$

For $t = 0$, the DNN at initialization, we get

$$f_{\theta_0}(\mathcal{X}) = \mathcal{Y} + e^{-\frac{\eta}{n} \, \Theta^{\text{nat}} \cdot 0} \, c = \mathcal{Y} + c \quad \Longleftrightarrow \quad c = f_{\theta_0}(\mathcal{X}) - \mathcal{Y}, \tag{68}$$

and thus

$$
\begin{aligned}
f_{\theta_t}(\mathcal{X}) &= \mathcal{Y} + e^{-\frac{\eta}{n}\,\Theta^{\text{nat}}\cdot t}(f_{\theta_0}(\mathcal{X}) - \mathcal{Y}) \qquad (69)\\
&= \left(I - e^{-\frac{\eta}{n}\Theta^{\text{nat}}\cdot t}\right)\mathcal{Y} + e^{-\frac{\eta}{n}\Theta^{\text{nat}}\cdot t} f_{\theta_0}(\mathcal{X}) \qquad (70)\\
&= \left(I - e^{-\frac{\eta}{n}\Theta^{\text{nat}}\cdot t}\right)\mathcal{Y} + e^{-\frac{\eta}{n}\Theta^{\text{nat}}\cdot t} f_{\theta_0}(\mathcal{X}). \qquad (71)
\end{aligned}
$$

□

To solve the training dynamics on an arbitrary point x, we can follow previous work [Lee et al., 2019] and make a linearization assumption in conjunction with our solution to the training dynamics on the training points derived above. In particular, we assume a linear evolution of the DNN.

Assumption 3 (Linearization)**.** *Assume that*

$$
f^{\text{lin}}_{\theta_t}(x) \overset{\text{def}}{=} f_{\theta_0}(x) + \nabla_\theta f_{\theta_t}(x)\big|_{\theta_t=\theta_0}(\theta_t - \theta_0). \tag{72}
$$

We then we get the linearized training dynamics on the training points $\mathcal{X}$,

$$
\frac{\partial(\theta_t - \theta_0)}{\partial t} = -\eta n\sigma^2 \nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}^\top \left(\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0} \nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}^\top\right)^{-1} \nabla_{f^{\text{lin}}}\mathcal{L}(f^{\text{lin}}_{\theta_t}(\mathcal{X})) \tag{73}
$$

$$
\begin{aligned}
\frac{\partial f_{\theta_t}(x)}{\partial t} &= -\eta n\sigma^2 \nabla_\theta f_{\theta_t}(x)\big|_{\theta_t=\theta_0} \nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}^\top \left(\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0} \nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}^\top\right)^{-1} \nabla_{f^{\text{lin}}}\mathcal{L}(f^{\text{lin}}_{\theta_t}(\mathcal{X})) \qquad (74)\\
&= -\eta\, \hat{\Theta}^{\text{nat}}_0(x, \mathcal{X}) \nabla_{f^{\text{lin}}}\mathcal{L}(f^{\text{lin}}_{\theta_t}(\mathcal{X}), \qquad (75)
\end{aligned}
$$

where

$$
\hat{\Theta}^{\text{nat}}_0(x, \mathcal{X}) \overset{\text{def}}{=} n\sigma^2 \nabla_\theta f_{\theta_t}(x)\big|_{\theta_t=\theta_0} \nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}^\top \left(\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0} \nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}^\top\right)^{-1} \tag{76}
$$

is the natural NTK at initialization.

Proposition 4 (Solution to Linearized Natural Gradient Descent Dynamics on Training Points)**.** *Under Assumption 1 and under MSE loss, the linearized function-space training dynamics under natural gradient descent on the training points $\mathcal{X}$, $\dot{f}^{\text{lin}}_{\theta_t}(\mathcal{X})$, can be solved as*

$$
f^{\text{lin}}_{\theta_t}(\mathcal{X}) = \left(I - e^{-\frac{\eta}{n}\Theta^{\text{nat}}\cdot t}\right)\mathcal{Y} + e^{-\frac{\eta}{n}\Theta^{\text{nat}}\cdot t} f_{\theta_0}(\mathcal{X}). \tag{77}
$$

Proof. Consider the linearized training dynamics under natural gradient descent, given by

$$
\begin{aligned}
\frac{\partial f_{\theta_t}(x)}{\partial t} &= -\eta n\sigma^2 \nabla_\theta f_{\theta_t}(x)\big|_{\theta_t=\theta_0} \nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}^\top \left(\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0} \nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}^\top\right)^{-1} \nabla_{f^{\text{lin}}}\mathcal{L}(f^{\text{lin}}_{\theta_t}(\mathcal{X})) \qquad (78)\\
&= -\eta\, \hat{\Theta}^{\text{nat}}_0(x, \mathcal{X}) \nabla_{f^{\text{lin}}}\mathcal{L}(f^{\text{lin}}_{\theta_t}(\mathcal{X}), \qquad (79)
\end{aligned}
$$

where

$$
\hat{\Theta}^{\text{nat}}_0(x, \mathcal{X}) \overset{\text{def}}{=} n\sigma^2 \nabla_\theta f_{\theta_t}(x)\big|_{\theta_t=\theta_0} \nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}^\top \left(\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0} \nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}^\top\right)^{-1} \tag{80}
$$

is the natural NTK at initialization.

The linearized function-space training dynamics on the training points then define a first-order ordinary differential equation, which can be solved analytically as

$$
f^{\text{lin}}_{\theta_t}(\mathcal{X}) = \left(I - e^{-\frac{\eta}{n}\Theta^{\text{nat}}\cdot t}\right)\mathcal{Y} + e^{-\frac{\eta}{n}\Theta^{\text{nat}}\cdot t} f^{\text{lin}}_{\theta_0}(\mathcal{X}), \tag{81}
$$

where we used the fact that the natural NTK is the scaled identity.

Using the fact that $f^{\text{lin}}_{\theta_0}(\mathcal{X}) = f_{\theta_0}(\mathcal{X})$, we can write the function-space solution as

$$
f^{\text{lin}}_{\theta_t}(\mathcal{X}) = \left(I - e^{-\frac{\eta}{n}\Theta^{\text{nat}}\cdot t}\right)\mathcal{Y} + e^{-\frac{\eta}{n}\Theta^{\text{nat}}\cdot t} f_{\theta_0}(\mathcal{X}). \tag{82}
$$

□

Corollary 4. *Throughout training via natural gradient descent, the solution to the linearized function-space training dynamics evaluated on the training points, $\dot{f}^{\text{lin}}_{\theta_t}(\mathcal{X})$, is identical to the solution to the non-linearized function-space training dynamics on the training points, $\dot{f}_{\theta_t}(\mathcal{X})$, that is,*

$$
f^{\text{lin}}_{\theta_t}(\mathcal{X}) = f_{\theta_t}(\mathcal{X}). \tag{83}
$$

Proof. The result follows immediately from Proposition 1 and Proposition 4. □

Proposition 5 (Parameter-space Solution to Linearized Natural Gradient Descent Dynamics)**.** *Under Assumption 1 and under MSE loss, the linearized training dynamics under natural gradient descent on the* DNN *parameters can be solved analytically as*

$$\theta_t = \left(I - e^{-\frac{\eta}{n}\Theta^{\text{nat}}\cdot t}\right)\left[\left(\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}^{\top}\left(\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}^{\top}\right)^{-1}\right)(f_{\theta_0}(\mathcal{X}) - \mathcal{Y}) + \theta_0\right], \tag{84}$$

where we used the fact that the natural NTK *is constant in f_{θ_t} and θ_t.*

Proof. Consider the linearized training dynamics on the DNN parameters, $\dot{\theta}_t$,

$$\frac{\partial(\theta_t - \theta_0)}{\partial t} = -\eta n\sigma^2\nabla_\theta f^{\text{lin}}_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}^{\top}\left(\nabla_\theta f^{\text{lin}}_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}\nabla_\theta f^{\text{lin}}_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}^{\top}\right)^{-1}\nabla_{f^{\text{lin}}}\mathcal{L}(f^{\text{lin}}_{\theta_t}(\mathcal{X})), \tag{85}$$

which defines a first-order ordinary differential equation and can be solved analytically as

$$\theta_t = \left(I - e^{-\frac{\eta}{n}\Theta^{\text{nat}}\cdot t}\right)\left[\left(\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}^{\top}\left(\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}^{\top}\right)^{-1}\right)(f_{\theta_0}(\mathcal{X}) - \mathcal{Y}) + \theta_0\right]. \tag{86}$$

□

Proposition 6 (Solution to Linearized Natural Gradient Descent Dynamics on Test Points)**.** *Under Assumption 1 and under MSE loss, the linearized function-space training dynamics under natural gradient descent on a test point x, $\dot{f}^{\text{lin}}_{\theta_t}(x)$, can be solved as*

$$\begin{aligned} f^{\text{lin}}_{\theta_t}(x) = {} & f_{\theta_0}(x) \\ & - \sigma^2\nabla_\theta f_{\theta_t}(x)\big|_{\theta_t=\theta_0}\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}^{\top}\left(\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}^{\top}\right)^{-1}\left(I - e^{-\eta\sigma^2 t}\right)f_{\theta_0}(\mathcal{X}) \\ & + \sigma^2\nabla_\theta f_{\theta_t}(x)\big|_{\theta_t=\theta_0}\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}^{\top}\left(\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}^{\top}\right)^{-1}\left(I - e^{-\eta\sigma^2 t}\right)\mathcal{Y}. \end{aligned} \tag{87}$$

Proof. Consider the linearized training dynamics *on a test point*, x, under MSE loss,

$$\begin{aligned} \frac{\partial f_{\theta_t}(x)}{\partial t} &= -\eta n\sigma^2\nabla_\theta f_{\theta_t}(x)\big|_{\theta_t=\theta_0}\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}^{\top}\left(\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}^{\top}\right)^{-1}\nabla_{f^{\text{lin}}}\mathcal{L}(f^{\text{lin}}_{\theta_t}(\mathcal{X})) && (88) \\ &= -\eta\hat{\Theta}^{\text{nat}}_0(x,\mathcal{X})\nabla_{f^{\text{lin}}}\mathcal{L}(f^{\text{lin}}_{\theta_t}(\mathcal{X})) && (89) \\ &= -\frac{\eta}{n}\hat{\Theta}^{\text{nat}}_0(x,\mathcal{X})(f^{\text{lin}}_{\theta_t}(\mathcal{X}) - \mathcal{Y}), && (90) \end{aligned}$$

which defines a first-order ordinary differential equation and can be solved analytically.

Using the fact that $f^{\text{lin}}_{\theta_0}(x) = f_{\theta_0}(x)$ and that, by Corollary 1, $f^{\text{lin}}_{\theta_t}(\mathcal{X}) = f_{\theta_t}(\mathcal{X})$, the solution to the linearized training dynamics of natural gradient descent is given by

$$\begin{aligned} f^{\text{lin}}_{\theta_t}(x) &= f_{\theta_0}(x) + \frac{1}{n}\hat{\Theta}^{\text{nat}}_0(x,\mathcal{X})\left(I - e^{-\frac{\eta}{n}\Theta^{\text{nat}}\cdot t}\right)(\mathcal{Y} - f_{\theta_0}(\mathcal{X})) && (91) \\ &= f_{\theta_0}(x) + \frac{1}{n}\hat{\Theta}^{\text{nat}}_0(x,\mathcal{X})\left(I - e^{-\frac{\eta}{n}\Theta^{\text{nat}}\cdot t}\right)(\mathcal{Y} - f_{\theta_0}(\mathcal{X})) && (92) \\ &= f_{\theta_0}(x) - \frac{1}{n}\hat{\Theta}^{\text{nat}}_0(x,\mathcal{X})\left(I - e^{-\frac{\eta}{n}\Theta^{\text{nat}}\cdot t}\right)f_{\theta_0}(\mathcal{X}) + \frac{1}{n}\hat{\Theta}^{\text{nat}}_0(x,\mathcal{X})\left(I - e^{-\frac{\eta}{n}\Theta^{\text{nat}}\cdot t}\right)\mathcal{Y} && (93) \\ &= f_{\theta_0}(x) \\ &\quad - \sigma^2\nabla_\theta f_{\theta_t}(x)\big|_{\theta_t=\theta_0}\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}^{\top}\left(\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}^{\top}\right)^{-1}\left(I - e^{-\eta\sigma^2 t}\right)f_{\theta_0}(\mathcal{X}) \\ &\quad + \sigma^2\nabla_\theta f_{\theta_t}(x)\big|_{\theta_t=\theta_0}\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}^{\top}\left(\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0}^{\top}\right)^{-1}\left(I - e^{-\eta\sigma^2 t}\right)\mathcal{Y}. && (94) \end{aligned}$$

□

Corollary 5. *In the limit of training time, as $t \to \infty$, the solution to the linearized training dynamics under natural gradient descent tends to*

$$\lim_{t \to \infty} f^{\text{lin}}_{\theta_t}(x) = f_{\theta_0}(x) + \frac{1}{n} \hat{\Theta}^{\text{nat}}_0(x, \mathcal{X}) \left(\mathcal{Y} - f_{\theta_0}(\mathcal{X}) \right), \tag{95}$$

where

$$\hat{\Theta}^{\text{nat}}_0(x, \mathcal{X}) \overset{\text{def}}{=} n\sigma^2 \nabla_\theta f_{\theta_t}(x)\big|_{\theta_t=\theta_0} \nabla_\theta f_{\theta_t}(\mathcal{X})\big|^\top_{\theta_t=\theta_0} \left(\nabla_\theta f_{\theta_t}(\mathcal{X})\big|_{\theta_t=\theta_0} \nabla_\theta f_{\theta_t}(\mathcal{X})\big|^\top_{\theta_t=\theta_0} \right)^{-1}. \tag{96}$$

Appendix E Discrepancy between function-space predictions under linearized and non-linearized training dynamics

In Appendix D, we made an explicit linearization assumption about the training dynamics under natural gradient descent, the analytical solution of which may be different from the (intractable) solution to the training dynamics under non-linearized natural gradient descent. Below, we derive an upper bound on the function-space discrepancy *on a test point* between the analytic solution to the linearized training dynamics under natural gradient descent and the function obtained by performing natural gradient descent on the training data until convergence.

Assumption 4 (Positive Definiteness of the Gram Matrix)**.** *Let $nk \leq p$, and define the Jacobian at initialization and at time step t during training as $J_0(x) \overset{\text{def}}{=} \nabla_\theta f_{\theta_t}(x)\big|_{\theta_t=\theta_0} \in \mathbb{R}^{nk \times p}$ and $J_t(x) \overset{\text{def}}{=} \nabla_\theta f_{\theta_t}(x) \in \mathbb{R}^{nk \times p}$, respectively, and let $G_0(x) = J_0(x)J_0(x)^\top \in \mathbb{R}^{nk \times nk}$ and $G_t(x) = J_t(x)J_t(x)^\top \in \mathbb{R}^{nk \times nk}$ be the corresponding Gram matrices. Assume that $G_t(x) = J_t(x)J_t(x)^\top$ is positive definite for all $t \geq 0$.*

Lemma 2 (Natural Neural Tangent Kernel Bound)**.** *Let $\hat{\Theta}^{\text{nat}}_0(x, \mathcal{X})$ and $\hat{\Theta}^{\text{nat}}_t(x, \mathcal{X})$ be the natural* NTK *at initialization and at some time step t during training, respectively, and let $\lambda_{\max}(G_t(x))$ be the largest eigenvalue of the Gram matrix for $t \geq 0$. Under Assumption 1, for any random initialization,*

$$\frac{1}{n\sigma^2} ||\hat{\Theta}^{\text{nat}}_0(x, \mathcal{X}) - \hat{\Theta}^{\text{nat}}_t(x, \mathcal{X})||_2 \leq \sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\max}(G_0(\mathcal{X}))}} + \sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_t(\mathcal{X}))}} + \frac{||J_0(x) - J_s(x)||_2}{\sqrt{\lambda_{\min}(G_t(\mathcal{X}))}} \tag{97}$$

Proof. For ease of notation, we define $J_t(x) \overset{\text{def}}{=} \nabla_\theta f_{\theta_t}(x)$ and $J_0(x) \overset{\text{def}}{=} \nabla_\theta f_{\theta_t}(x)\big|_{\theta_t=\theta_0}$. We then have

$$\hat{\Theta}^{\text{nat}}_t(x, \mathcal{X}) = n\sigma^2 J_t(x) J_t(\mathcal{X})^\top \left(J_t(\mathcal{X}) J_t(\mathcal{X})^\top \right)^{-1} \tag{98}$$

and

$$\hat{\Theta}^{\text{nat}}_0(x, \mathcal{X}) = n\sigma^2 J_0(x) J_0(\mathcal{X})^\top \left(J_0(\mathcal{X}) J_0(\mathcal{X})^\top \right)^{-1}, \tag{99}$$

and thus

$$\frac{1}{n\sigma^2} ||\hat{\Theta}^{\text{nat}}_0(x, \mathcal{X}) - \hat{\Theta}^{\text{nat}}_t(x, \mathcal{X})||_2 \tag{100}$$

$$= \left|\left| J_0(x) J_0(\mathcal{X})^\top \left(J_0(\mathcal{X}) J_0(\mathcal{X})^\top \right)^{-1} - J_t(x) J_t(\mathcal{X})^\top \left(J_t(\mathcal{X}) J_t(\mathcal{X})^\top \right)^{-1} \right|\right|_2. \tag{101}$$

Adding and subtracting $J_0(x) J_t(\mathcal{X})^\top \left(J_t(\mathcal{X}) J_t(\mathcal{X})^\top \right)^{-1}$, we get

$$\begin{aligned} &\frac{1}{n\sigma^2} ||\hat{\Theta}^{\text{nat}}_0(x, \mathcal{X}) - \hat{\Theta}^{\text{nat}}_t(x, \mathcal{X})||_2 \\ &\quad = \Big|\Big| J_0(x) J_0(\mathcal{X})^\top \left(J_0(\mathcal{X}) J_0(\mathcal{X})^\top \right)^{-1} - J_0(x) J_t(\mathcal{X})^\top \left(J_t(\mathcal{X}) J_t(\mathcal{X})^\top \right)^{-1} \\ &\qquad + J_0(x) J_t(\mathcal{X})^\top \left(J_t(\mathcal{X}) J_t(\mathcal{X})^\top \right)^{-1} - J_t(x) J_t(\mathcal{X})^\top \left(J_t(\mathcal{X}) J_t(\mathcal{X})^\top \right)^{-1} \Big|\Big|_2, \end{aligned} \tag{102}$$

which allows us to separate Jacobians evaluated on training and on test points, that is,

$$\begin{aligned} \frac{1}{n\sigma^2} ||\hat{\Theta}^{\text{nat}}_0(x, \mathcal{X}) - \hat{\Theta}^{\text{nat}}_t(x, \mathcal{X})||_2 = \Big|\Big| J_0(x) \left(J_0(\mathcal{X})^\top \left(J_0(\mathcal{X}) J_0(\mathcal{X})^\top \right)^{-1} - J_t(\mathcal{X})^\top \left(J_t(\mathcal{X}) J_t(\mathcal{X})^\top \right)^{-1} \right) \\ - \left(J_t(x) - J_0(x) \right) J_t(\mathcal{X})^\top \left(J_t(\mathcal{X}) J_t(\mathcal{X})^\top \right)^{-1} \Big|\Big|_2. \end{aligned} \tag{103}$$

By applying the triangle inequality, we can now establish an upper bound,

$$
\begin{aligned}
&\frac{1}{n\sigma^2}||\hat{\Theta}_0^{\text{nat}}(x,\mathcal{X}) - \hat{\Theta}_t^{\text{nat}}(x,\mathcal{X})||_2 \\
&\quad \leq \Big|\Big| J_0(x)\left(J_0(\mathcal{X})^\top \left(J_0(\mathcal{X})J_0(\mathcal{X})^\top\right)^{-1} - J_t(\mathcal{X})^\top \left(J_t(\mathcal{X})J_t(\mathcal{X})^\top\right)^{-1}\right)\Big|\Big|_2 \\
&\qquad + \Big|\Big| \left(J_t(x) - J_0(x)\right)\left(J_t(\mathcal{X})^\top \left(J_t(\mathcal{X})J_t(\mathcal{X})^\top\right)^{-1}\right)\Big|\Big|_2 \qquad (104)\\
&\quad \leq ||J_0(x)||_2 \Big|\Big| J_0(\mathcal{X})^\top \left(J_0(\mathcal{X})J_0(\mathcal{X})^\top\right)^{-1} - J_t(\mathcal{X})^\top \left(J_t(\mathcal{X})J_t(\mathcal{X})^\top\right)^{-1}\Big|\Big|_2 \\
&\qquad + ||J_0(x) - J_t(x)||_2 \Big|\Big| J_t(\mathcal{X})^\top \left(J_t(\mathcal{X})J_t(\mathcal{X})^\top\right)^{-1}\Big|\Big|_2 \qquad (105)\\
&\quad \leq ||J_0(x)||_2 \left(\Big|\Big| J_0(\mathcal{X})^\top \left(J_0(\mathcal{X})J_0(\mathcal{X})^\top\right)^{-1}\Big|\Big|_2 + \Big|\Big| J_t(\mathcal{X})^\top \left(J_t(\mathcal{X})J_t(\mathcal{X})^\top\right)^{-1}\Big|\Big|_2\right) \\
&\qquad + ||J_0(x) - J_t(x)||_2 \Big|\Big| J_t(\mathcal{X})^\top \left(J_t(\mathcal{X})J_t(\mathcal{X})^\top\right)^{-1}\Big|\Big|_2, \qquad (106)
\end{aligned}
$$

and noting that the spectral norm $||\cdot||_2$ of a matrix A is given by its largest singular value, $\sigma_{\max}(A)$, or, equivalently, by the square root of the largest eigenvalue of $A^\top A$, $\lambda_{\max}(A^\top A)$, we define $G_t(x) = J_t(x)J_t(x)^\top \in \mathbb{R}^{nk\times nk}$ for a general x and write

$$
\begin{aligned}
\Big|\Big| J_t(x)^\top \left(J_t(x)J_t(x)^\top\right)^{-1}\Big|\Big|_2 &= \sigma_{\max}\left(J_t(x)^\top \left(J_t(x)J_t(x)^\top\right)^{-1}\right) && (107)\\
&= \sqrt{\lambda_{\max}\left(\left(J_t(x)^\top \left(J_t(x)J_t(x)^\top\right)^{-1}\right)^\top \left(J_t(x)^\top \left(J_t(x)J_t(x)^\top\right)^{-1}\right)\right)} && (108)\\
&= \sqrt{\lambda_{\max}\left(\left(J_t(x)J_t(x)^\top\right)^{-1}\right)} && (109)\\
&= \sqrt{\lambda_{\max}(G_t(x)^{-1})} && (110)\\
&= \frac{1}{\sqrt{\lambda_{\min}(G_t(x))}} && (111)
\end{aligned}
$$

and, by definition,

$$
||J_t(x)||_2 = \sqrt{\lambda_{\max}(G_t(x))} \quad \forall t. \tag{112}
$$

We can now express the bound above as

$$
\begin{aligned}
\frac{1}{n\sigma^2}||\hat{\Theta}_0^{\text{nat}}(x,\mathcal{X}) - \hat{\Theta}_t^{\text{nat}}(x,\mathcal{X})||_2 &\leq \sqrt{\lambda_{\max}(G_0(x))}\left(\frac{1}{\sqrt{\lambda_{\min}(G_0(\mathcal{X}))}} + \frac{1}{\sqrt{\lambda_{\min}(G_t(\mathcal{X}))}}\right) \\
&\quad + ||J_0(x) - J_t(x)||_2 \frac{1}{\sqrt{\lambda_{\min}(G_t(\mathcal{X}))}} \qquad (113)\\
&= \sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_0(\mathcal{X}))}} + \sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_t(\mathcal{X}))}} + \frac{||J_0(x) - J_s(x)||_2}{\sqrt{\lambda_{\min}(G_t(\mathcal{X}))}}
\end{aligned}
$$

concluding the proof. □

Theorem 2 (Prediction Error under Linearized Training Dynamics)**.** *Let (x, y) be an input-output test set pair, $\mathcal{X}$ and $\mathcal{Y}$ the training input and output sets, respectively, η the learning rate, σ^2 the variance of the Gaussian likelihood, $f^{\text{lin}}_{\theta_t}(x)$ the function predictions from the analytical solution to the linearized training dynamics under natural gradient descent (a random variable), and $f_{\theta_t}(x)$ the function predictions obtained from running natural gradient descent (also a random variable). Under Assumption 1, and under MSE loss, for $g_{\theta_t}(x) = f_{\theta_t}(x) - y$, the spectral norm $||\cdot||_2$, and the natural* NTK *at initialization, $\hat{\Theta}_0^{\text{nat}}(x,\mathcal{X})$, and at time t, $\hat{\Theta}_t^{\text{nat}}(x,\mathcal{X})$, for any random initialization,*

$$
\begin{aligned}
||g^{\text{lin}}_{\theta_t}(x) - g_{\theta_t}(x)||_2 &\leq ||f_{\theta_0}(\mathcal{X}) - \mathcal{Y}||_2 \Bigg(\sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_0(\mathcal{X}))}} \left(I - e^{-\eta\sigma^2 t}\right) \\
&\quad + \eta\sigma^2 \int_0^t \left(\sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_s(\mathcal{X}))}} + \frac{||J_0(x) - J_s(x)||_2}{\sqrt{\lambda_{\min}(G_s(\mathcal{X}))}}\right) e^{-\eta\sigma^2 s} ds\Bigg), \qquad (114)
\end{aligned}
$$

Proof. Consider

$$\frac{\partial}{\partial t}\left(g^{\text{lin}}_{\theta_t}(x) - g_{\theta_t}(x)\right) = -\frac{\eta}{n}\hat{\Theta}^{\text{nat}}_0(x,\mathcal{X})g^{\text{lin}}_{\theta_t}(\mathcal{X}) - \left(-\frac{\eta}{n}\hat{\Theta}^{\text{nat}}_t(x,\mathcal{X})g_{\theta_t}(\mathcal{X})\right) \tag{115}$$

$$\begin{aligned}= &-\frac{\eta}{n}\hat{\Theta}^{\text{nat}}_0(x,\mathcal{X})g^{\text{lin}}_{\theta_t}(\mathcal{X}) + \frac{\eta}{n}\hat{\Theta}^{\text{nat}}_t(x,\mathcal{X})g^{\text{lin}}_{\theta_t}(\mathcal{X}) \\ &- \frac{\eta}{n}\hat{\Theta}^{\text{nat}}_t(x,\mathcal{X})g^{\text{lin}}_{\theta_t}(\mathcal{X}) + \frac{\eta}{n}\hat{\Theta}^{\text{nat}}_t(x,\mathcal{X})g_{\theta_t}(\mathcal{X})\end{aligned} \tag{116}$$

$$= -\frac{\eta}{n}\left(\hat{\Theta}^{\text{nat}}_0(x,\mathcal{X}) - \hat{\Theta}^{\text{nat}}_t(x,\mathcal{X})\right)g^{\text{lin}}_{\theta_t}(\mathcal{X}) + \frac{\eta}{n}\hat{\Theta}^{\text{nat}}_{\theta_t}(x,\mathcal{X})\underbrace{\left(g_{\theta_t}(\mathcal{X}) - g^{\text{lin}}_{\theta_t}(\mathcal{X}))\right)}_{=0} \tag{117}$$

$$= -\frac{\eta}{n}\left(\hat{\Theta}^{\text{nat}}_0(x,\mathcal{X}) - \hat{\Theta}^{\text{nat}}_t(x,\mathcal{X})\right)g^{\text{lin}}_{\theta_t}(\mathcal{X}), \tag{118}$$

where the last line follows from the fact that, by Corollary 1, on the training data, $f^{\text{lin}}_{\theta_s}(\mathcal{X}) = f_{\theta_s}(\mathcal{X})$. Integrating with respect to t, taking the norm, and repeatedly applying the triangle inequality then yields

$$||g^{\text{lin}}_{\theta_t}(x) - g_{\theta_t}(x)||_2 = \frac{\eta}{n}\left|\left|\int_0^t \left(\hat{\Theta}^{\text{nat}}_0(x,\mathcal{X}) - \hat{\Theta}^{\text{nat}}_s(x,\mathcal{X})\right)g^{\text{lin}}_{\theta_s}(\mathcal{X})\partial s\right|\right|_2 \tag{119}$$

$$\leq \frac{\eta}{n}\int_0^t \left|\left|\left(\hat{\Theta}^{\text{nat}}_0(x,\mathcal{X}) - \hat{\Theta}^{\text{nat}}_s(x,\mathcal{X})\right)g^{\text{lin}}_{\theta_s}(\mathcal{X})\right|\right|_2 \mathrm{d}s \tag{120}$$

$$\leq \frac{\eta}{n}\int_0^t ||\hat{\Theta}^{\text{nat}}_0(x,\mathcal{X}) - \hat{\Theta}^{\text{nat}}_s(x,\mathcal{X})||_2\,||g^{\text{lin}}_{\theta_s}(\mathcal{X})||_2 \mathrm{d}s. \tag{121}$$

Noting that

$$||g^{\text{lin}}_{\theta_s}(\mathcal{X})||_2 = \left|\left|\left(I - e^{-\frac{\eta}{n}\Theta^{\text{nat}}\cdot s}\right)\mathcal{Y} + e^{-\frac{\eta}{n}\Theta^{\text{nat}}\cdot s}f_{\theta_0}(\mathcal{X}) - \mathcal{Y}\right|\right|_2 \tag{122}$$

$$= \left|\left|e^{-\frac{\eta}{n}\Theta^{\text{nat}}\cdot s}\left(f_{\theta_0}(\mathcal{X}) - \mathcal{Y}\right)\right|\right|_2 \tag{123}$$

$$= e^{-\frac{\eta}{n}\Theta^{\text{nat}}\cdot s}||f_{\theta_0}(\mathcal{X}) - \mathcal{Y}||_2, \tag{124}$$

we can rewrite the above bound as

$$||g^{\text{lin}}_{\theta_t}(x) - g_{\theta_t}(x)||_2 \leq \frac{\eta}{n}||f_{\theta_0}(\mathcal{X}) - \mathcal{Y}||_2 \int_0^t ||\hat{\Theta}^{\text{nat}}_0(x,\mathcal{X}) - \hat{\Theta}^{\text{nat}}_s(x,\mathcal{X})||_2 e^{-\frac{\eta}{n}\Theta^{\text{nat}}\cdot s}\partial s. \tag{125}$$

By Lemma 1, we then have

$$||g^{\text{lin}}_{\theta_t}(x) - g_{\theta_t}(x)||_2 \leq \frac{\eta}{n}||f_{\theta_0}(\mathcal{X}) - \mathcal{Y}||_2 \int_0^t ||\hat{\Theta}^{\text{nat}}_0(x,\mathcal{X}) - \hat{\Theta}^{\text{nat}}_s(x,\mathcal{X})||_2 e^{-\frac{\eta}{n}\Theta^{\text{nat}}\cdot s}\mathrm{d}s \tag{126}$$

$$\begin{aligned}\leq \frac{\eta}{n}||f_{\theta_0}(\mathcal{X}) - \mathcal{Y}||_2 \int_0^t n\sigma^2 \Bigg(&\sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_0(\mathcal{X}))}} \\ &+ \sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_s(\mathcal{X}))}} + \frac{||J_0(x) - J_s(x)||_2}{\sqrt{\lambda_{\min}(G_s(\mathcal{X}))}}\Bigg)e^{-\frac{\eta}{n}\Theta^{\text{nat}}\cdot s}\mathrm{d}s\end{aligned} \tag{127}$$

$$\begin{aligned}= \eta\,||f_{\theta_0}(\mathcal{X}) - \mathcal{Y}||_2 \Bigg(&\sigma^2 \int_0^t \sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_0(\mathcal{X}))}}e^{-\frac{\eta}{n}\Theta^{\text{nat}}\cdot s}\mathrm{d}s \\ &+ \sigma^2 \int_0^t \sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_s(\mathcal{X}))}} + \frac{||J_0(x) - J_s(x)||_2}{\sqrt{\lambda_{\min}(G_s(\mathcal{X}))}}e^{-\frac{\eta}{n}\Theta^{\text{nat}}\cdot s}\mathrm{d}s\Bigg).\end{aligned} \tag{128}$$

Using the fact that $\Theta^{\text{nat}} \approx n\sigma^2 I_{nk}$, we then have

$$\begin{aligned}
||g^{\text{lin}}_{\theta_t}(x) - g_{\theta_t}(x)||_2 \leq \eta\, ||f_{\theta_0}(\mathcal{X}) - \mathcal{Y}||_2 \Bigg(& \sigma^2 \sqrt{\frac{\lambda_{\min}(G_0(x))}{\lambda_{\min}(G_0(\mathcal{X}))}} \int_0^t e^{-\eta\sigma^2 s} \mathrm{d}s \\
& + \sigma^2 \int_0^t \left(\sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_s(\mathcal{X}))}} + \frac{||J_0(x) - J_s(x)||_2}{\sqrt{\lambda_{\min}(G_s(\mathcal{X}))}} \right) e^{-\eta\sigma^2 s} \mathrm{d}s \Bigg)
\end{aligned} \tag{129}$$

$$\begin{aligned}
= \eta ||f_{\theta_0}(\mathcal{X}) - \mathcal{Y}||_2 \Bigg(& \frac{1}{\eta} \sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_0(\mathcal{X}))}} \left(I - e^{-\eta\sigma^2 t} \right) \\
& + \sigma^2 \int_0^t \left(\sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_s(\mathcal{X}))}} + \frac{||J_0(x) - J_s(x)||_2}{\sqrt{\lambda_{\min}(G_s(\mathcal{X}))}} \right) e^{-\eta\sigma^2 s} \mathrm{d}s \Bigg)
\end{aligned} \tag{130}$$

$$\begin{aligned}
= ||f_{\theta_0}(\mathcal{X}) - \mathcal{Y}||_2 \Bigg(& \sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_0(\mathcal{X}))}} \left(I - e^{-\eta\sigma^2 t} \right) \\
& + \eta\sigma^2 \int_0^t \left(\sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_s(\mathcal{X}))}} + \frac{||J_0(x) - J_s(x)||_2}{\sqrt{\lambda_{\min}(G_s(\mathcal{X}))}} \right) e^{-\eta\sigma^2 s} \mathrm{d}s \Bigg).
\end{aligned} \tag{131}$$

This concludes the proof. □

Corollary 6 (Prediction Error under Linearized Training Dynamics at Convergence). *In the limit of training time, as $t \to \infty$, where $\lim_{t\to\infty} g_{\theta_t}(x) = g^\star_\theta(x)$, is a function prediction when natural gradient descent has reached an optimum, for any random initialization,*

$$\begin{aligned}
\lim_{t\to\infty} ||g^{\text{lin}}_{\theta_t}(x) - g^\star_\theta(x)||_2 \leq ||f_{\theta_0}(\mathcal{X}) - \mathcal{Y}||_2 \Bigg(& \sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_0(\mathcal{X}))}} \\
& + \eta\sigma^2 \lim_{t\to\infty} \int_0^t \frac{\sqrt{\lambda_{\max}(G_0(x))} + ||J_0(x) - J_s(x)||_2}{\sqrt{\lambda_{\min}(G_s(\mathcal{X}))}} e^{-\eta\sigma^2 s} ds \Bigg).
\end{aligned} \tag{132}$$

Assumption 5 (Local Lipschitzness). *Assume that*

$$||J_0(x) - J_t(x)||_2 \leq ||J_0(x)||_2, \tag{133}$$

which can be viewed as assuming a type of local Lipschitzness

Assumption 6 (Increasing Maximum Eigenvalue of the Jacobian). *Assume that*

$$||J_0(\mathcal{X})||_2 \leq ||J_t(\mathcal{X})||_2 \quad \forall t \geq 0, \tag{134}$$

which corresponds to the largest eigenvalues of the Jacobian increasing during training.

Corollary 7. *Under Assumption 5 and Assumption 6, in the limit of training time, as $t \to \infty$, where $\lim_{t\to\infty} g_{\theta_t}(x) = g^\star_\theta(x)$, is a function prediction when natural gradient descent has reached an optimum, for any random initialization,*

$$\lim_{t\to\infty} ||g^{\text{lin}}_{\theta_t}(x) - g^\star_\theta(x)||_2 \leq 3\, ||f_{\theta_0}(\mathcal{X}) - \mathcal{Y}||_2 \sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_0(\mathcal{X}))}}. \tag{135}$$

Proof. Under Assumption 5 and Assumption 6, we have

$$\lim_{t\to\infty} ||g^{\text{lin}}_{\theta_t}(x) - g^\star_\theta(x)||_2 \tag{136}$$

$$\leq ||f_{\theta_0}(\mathcal{X}) - \mathcal{Y}||_2 \left(\sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_0(\mathcal{X}))}} + \eta\sigma^2 \lim_{t\to\infty} \int_0^t 2 \sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_0(\mathcal{X}))}} e^{-\eta\sigma^2 s} \mathrm{d}s \right) \tag{137}$$

$$= ||f_{\theta_0}(\mathcal{X}) - \mathcal{Y}||_2 \left(\sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_0(\mathcal{X}))}} + 2\eta\sigma^2 \sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_0(\mathcal{X}))}} \lim_{t\to\infty} \int_0^t e^{-\eta\sigma^2 s} \mathrm{d}s \right) \tag{138}$$

$$= ||f_{\theta_0}(\mathcal{X}) - \mathcal{Y}||_2 \left(\sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_0(\mathcal{X}))}} + 2 \sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_0(\mathcal{X}))}} \lim_{t\to\infty} \left(1 - e^{-\eta\sigma^2 t} \right) \right) \tag{139}$$

$$= 3\, ||f_{\theta_0}(\mathcal{X}) - \mathcal{Y}||_2 \sqrt{\frac{\lambda_{\max}(G_0(x))}{\lambda_{\min}(G_0(\mathcal{X}))}}. \tag{140}$$

□