Neural Tangent Kernels Under Stochastic Data Augmentation

Joshua DeOliveira, Sajal Chakroborty, Walter Gerych, Elke Rundensteiner

Worcester Polytechnic Institute, Worcester, MA
{jcdeoliveira, schakroborty, wgerych, rundenst} @wpi.edu

Abstract

The learning dynamics of modern neural networks remain an open problem in deep learning. The *Neural Tangent Kernel* (NTK) offers an elegant description of training dynamics in the infinite-width limit, yet its classical formulation assumes a static data set. Modern model training practice departs from this strong assumption through the use of on-the-fly data augmentations (e.g. additive noise). In this work, we conduct an NTK-driven analysis of how data transformations affect a neural net's evolution in the function space. Our theoretical contributions characterize how repeated Gaussian perturbations from NTK-derived covariances can steer neural-net optimizations toward user-specified behavior. These theoretical insights are empirically validated by controlled experiments. Taken together, our results lay the foundation for a promising future research direction that transforms the NTK from a descriptive to a *prescriptive* tool, enabling control of neural net training trajectories and behavior of inference generalization with grounded interventions.

1 Introduction

Background. Neural networks have demonstrated remarkable efficacy in solving complex learning tasks across diverse data modalities and application domains (Miotto et al. 2018; Halbouni et al. 2022). Despite their empirical success, a rigorous theoretical understanding of their training dynamics remains elusive. One avenue towards tackling this fundamental problem in deep learning has been through the Neural Tangent Kernel (NTK) (Jacot, Gabriel, and Hongler 2018), which provides a kernel-based perspective on the dynamics of neural networks, particularly in the infinite-width limit.

State-of-the-Art. For a neural net $f : \mathbb{R}^n \to \mathbb{R}^m$ parameterized by θ, Jacot, Gabriel, and Hongler (2018) proved in their seminal paper that the time evolution[1] of f in function space is determined by the negative gradient of the objective function $\mathcal{C}$ with respect to the NTK Θ imposed by f:

$$d_t f(\cdot;\theta_t) = -\nabla_{\Theta_f}\mathcal{C}|_{f(\cdot;\theta_t)} \tag{1}$$

$$\Theta_f(\mathbf{z},\mathbf{x};\theta_t) := \langle \nabla_{\boldsymbol{\theta}} f(\mathbf{z};\theta_t), \nabla_{\boldsymbol{\theta}} f(\mathbf{x};\theta_t)\rangle \tag{2}$$

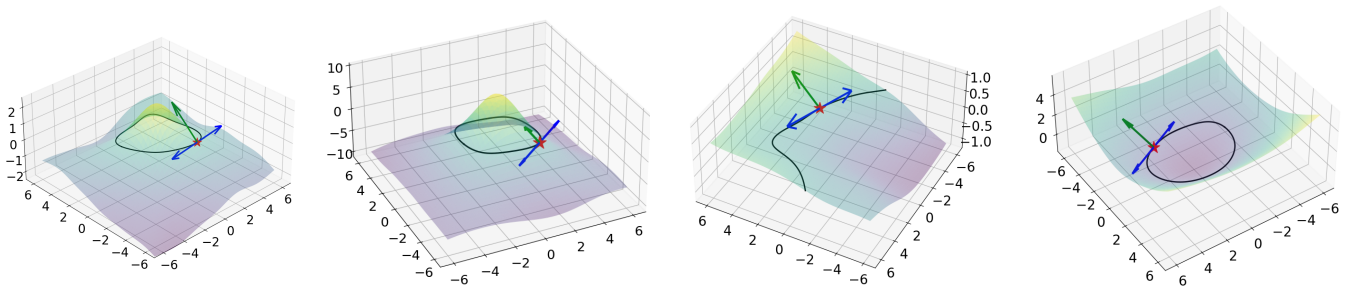
[1] continuous-time training under an infinitesimal learning rate

Notably, as the number of neurons in the hidden layers or the number of filters in convolutional layers of f approaches infinitely many, the infinitely-wide NTK (Θ^∞) remains constant during infinite-width training. It has also been proven by Jacot, Gabriel, and Hongler (2018) and Arora et al. (2019b) that for a finite-width architecture with width w, under the limit $\lim_{t\to\infty}$, Θ converges to its infinite-counterpart Θ^∞ within error $|\Theta - \Theta^\infty| = \mathcal{O}\left(1/\sqrt{w}\right)$. With closed-form analytical solutions of Θ^∞ having been derived for MLPs, CNNs (Arora et al. 2019b) and attention variants (Hron et al. 2020; Yang 2020), the NTK offers the promise of a mathematically-tractable approach to be able to analyze finite-width neural net learning processes in the future.

Problem. However, while NTK theory has provided valuable insights into the training behavior of neural networks, its foundational analysis in Jacot, Gabriel, and Hongler (2018) assumes a fixed training dataset and full-batch gradient descent. Yet crucial aspects of real-world neural network training do not match these assumptions. Real-world training instead involve numerous advances such as dynamic data augmentation techniques and multi-staged training like fine-tuning (Hu et al. 2022). There is a clear need to extend the analysis of NTK to incorporate these real-world training advances, thereby improving our understanding of practical deep learning strategies. We thus explore how the NTK can be leveraged to control neural net learning trajectories under training with dynamic Gaussian data augmentations.

Contributions. In this paper, we provide the first analysis for infinite-width networks trained under dynamic augmentations. Namely, we theoretically prove that due to the geometric symmetries of a neural net's NTK (see Figure 1), solely injecting well-placed Gaussian noise onto training instances allows for intentional control of training trajectories (Definition 3.3). Driven by our analysis, we empirically validate that our theoretical claims are observable in a controlled case study, and further demonstrate the promise of a new line of research into steering a neural network's learning through designed perturbations. Our contributions are as follows:

- We demonstrate that datasets repeatedly augmented with noise sampled from precisely defined Gaussians have the remarkable ability to steer or freeze neural-net dynamics in function space for chosen regions of the input space.
- We demonstrate that when perturbing training data with Gaussian augmentations, there locally exists only 1 local



(a) tanh, width = 20 (b) tanh width = 100 (c) GELU, width = 20 (d) GELU, width = 100

Figure 1: NTK Symmetries across neural nets. For a fixed anchor input z and reference point x (red star), each panel visualizes the scalar field $u \mapsto \Theta(z,u) - \Theta(z,x)$ of a 2-layer neural net and the **level set** $\{u \in \mathbb{R}^2 \mid \Theta(z,u) = \Theta(z,x)\}$. The blue arrows show directions that lie on the level-set, whereas the green arrow shows the gradient's path off the level set. Panels differ by activation (tanh vs. GELU) and hidden width (20 vs. 100). This illustrates how different architectures deform the symmetries of the NTK. This paper's analysis shows that perturbing training data to lie either on or off these level sets can control training.

principal direction in which function-space is affected.

- We show in that for the infinite-width 2-layer ReLU network that (i) its NTK coincides exactly with inner-product level sets, and (ii) for each training instance, there exists globally only 2 principal directions that affect function dynamics *independent of input space dimension.*

All proofs not in our paper are in supplementary materials.

2 Related Work

To the best of our knowledge, we fill a gap in NTK research by analyzing input-space NTK equivalence classes, showing that Gaussian noise can steer or freeze function-space dynamics[2].

Neural Tangent Kernels. Prior work beyond the landmark investigation (Jacot, Gabriel, and Hongler 2018) has studied how the NTK relates to Hessians (Bordelon, Canatar, and Pehlevan 2020) and to Fisher information matrices (Martens 2020). Others (Fort, Hu, and Lakshminarayanan 2019) have explored neural net ensembles under the NTK. Work by Shan and Bordelon (2021) has investigated the impact of NTK's Gram matrix alignment on learning in finite-width models. Additional foundational work in NTK theory has found analytical solutions for Θ^{∞} in simple networks (Arora et al. 2019b), CNNs (Novak et al. 2019), residual networks (Yang 2020), and simple attention (Hron et al. 2020). These extensions have been impactful, but *retain* the original assumptions of a *static* data set, which do not match modern training.

NTK Under Noise. A growing body of work investigates how stochastic elements in the training regime affect the NTK. (Mei and Montanari 2019) established precise asymptotics for NTK regression under label noise. Similarly, (Tsilivis and Kempe 2022) show that dropout rescales the NTK spectrum, while (Refinetti, Ingrosso, and Goldt 2023) trace its evolution under SGD noise. While prior studies focus on fixed label-noise or gradient-level noise, our paper investigates noise to perturb training instances in the *input* space that purposefully aim to steer dynamics.

[2]For greater detail of why we conduct this analysis in function-space over instead of parameter-space dynamics, see Appendix A.

Symmetries Under Neural Nets. Geometric deep learning (Bronstein et al. 2017) is a subfield that studies neural nets and their learning through the perspective of their geometric properties. While there has been non-NTK work in this space (Godfrey et al. 2022), tackling this analysis from an NTK lens has become a burgeoning line of work such as NTK symmetries under infinite-width ensembles (Gerken and Kessel 2024), Group-CNNs (Misof, Kessel, and Gerken 2025), or multi-class settings (Perin and Deny 2024). Our paper's focus is distinct in that it studies how Gaussian augmentation impact symmetries of the NTK, and consequently how these symmetries affect neural net training.

3 Data Stochasticity on NTK Dynamics

In this section, we demonstrate we can derive an expanded version of NTK theory that can handle the more flexible assumptions about the training dataset. Thus, all existing NTK theory shown in (Jacot, Gabriel, and Hongler 2018) still holds, and acts as a special case of what is shown here.

Introduction and Setup

When we consider a dataset of N training instances $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$, we can imagine a *stochastic dataset* $\tilde{\mathcal{X}} = \{X_t^{(1)}, X_t^{(2)}, \ldots, X_t^{(N)}\}$ that perturbs elements of $\mathcal{X}$ following a Gaussian distribution $\epsilon_t^{(i)} \sim \mathcal{N}(\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$ according to some instance-dependent mean $\boldsymbol{\mu_i}$ and covariance $\boldsymbol{\Sigma_i}$:

$$X_t^{(i)} = x_i + \epsilon_t^{(i)}, \quad i = 1, 2, \ldots, N. \tag{3}$$

Re-evaluating the integral of the kernel gradient (Equation 1)

$$\nabla_{\Theta_f} C|_{f_t}(z) = \int \Theta_f(z, x') \frac{\partial C}{\partial f_t(x')} d\omega(x') \tag{4}$$

where x' is defined under a novel set of dirac measures determined by the sampling of $\hat{x}_i = X_t^{(i)}$ at every time t

$$\omega(x') = \frac{1}{N} \sum_{i=1}^{N} \delta_{\hat{x}_i}, \quad \delta_{\hat{x}_i}(x) = \begin{cases} 1 & x = \hat{x}_i \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Thus, we then reconsider the function evolution:

$$d_t f(z;\theta_t) = -\frac{1}{N}\sum_{i=1}^{N}\Theta_f(z, X_t^{(i)};\theta_t)\widehat{C}(X_t^{(i)}, y_i^*;\theta_t) \quad (6)$$

where $\widehat{C}(x, y^*;\theta_t) = \frac{\partial C}{\partial \hat{y}} C(\hat{y}, y^*)\Big|_{\hat{y}=f(x;\theta_t)}$

Stochastic Differential Equation (SDE) Formulation

Thus, the stochastic differential equation governing the function evolution of f, $\mathbf{W}_t \in \mathbb{R}^n$, is an n-dimensional Wiener process with drift and diffusion terms μ and σ (Equation 7).

$$d_t f(z;\theta_t) = \mu(z;\theta_t)dt + \sigma(z;\theta_t)d\mathbf{W}_t, \quad (7)$$

A first-order Taylor expansion written in terms of $x_i \in \mathcal{X}$,

$$\mu(z;\theta_t) = -\mathbb{E}[\Theta_f(z, x_i;\theta_t)\widehat{C}(x_i, y_i^*;\theta_t) + \nabla_{\mathbf{x}}\Theta_f(z, x_i;\theta_t)^\top \Sigma_i \nabla_{\mathbf{x}}\widehat{C}(x_i, y_i^*;\theta_t)] \quad (8)$$

$$\sigma(z;\theta_t) = -\mathbb{E}[\Theta_f(z, x_i;\theta_t)\nabla_{\mathbf{x}}\widehat{C}(x_i, y_i^*;\theta_t)\epsilon_t^{(i)} + \nabla_{\mathbf{x}}\Theta_f(z, x_i;\theta_t)\epsilon_t^{(i)}\widehat{C}(x_i, y_i^*;\theta_t)] \quad (9)$$

shows the drift term is linear in the per-instance covariance. With scaling or canceling kernel-gradient components, one can change in the drift of the function evolution of f. In Definitions 3.1 and 3.2, we define these two behaviors as *steerable* and *freezable* dynamics, which are investigated further in Section 4 to demonstrate their construction.

Definition 3.1 (Steerable Dynamics)**.** Training is *steerable* at a reference point $z \in \mathbb{R}^n$ if there exists a perturbation covariance $\Sigma(x) \in \mathbb{R}^{n\times n}$ such that the expected functional evolution of its drift $\mu(z;\theta_t)$ is intentionally modulated, $\mu(z;\theta_t) := -\mathbb{E}[\Theta_f(z, x_i;\theta_t)\widehat{C}(x_i, y_i^*;\theta_t)] + \delta$, by $\delta \neq 0$.

Definition 3.2 (Freezable Dynamics)**.** Let $z \in \mathbb{R}^n$ be a reference point and $x \in \mathcal{X}$ be a training element. The training dynamics of $f(z)$ are said to be *frozen* under a perturbation scheme if the expected drift under perturbations matches the unperturbed case: $\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\Sigma(x))}[\Theta(z, x+\epsilon)] = \Theta(z, x)$.

In Definition 3.3, we define a stronger form of steerable dynamics, in which training is *controllable* with precisely prescribed dynamics throughout training. In Section 5, we investigate some simple strategies for controllable training.

Definition 3.3 (Controlling Training Trajectories)**.** Let $f_t(z) := f(z;\theta_t)$. We say that the training trajectory is *controllable* if there exists a family of perturbation covariances $\{\Sigma(x;t)\}_{x\in\mathcal{X}}$ such that for a prescribed target trajectory $\gamma_z : [0,T] \to \mathbb{R}$, we have $|f_t(z) - \gamma_z(t)| < \varepsilon \quad \forall t \in [0,T]$, for some small $\varepsilon > 0$.

Remark 3.4 (Recovery of the classical NTK flow). Setting every noise covariance to zero, $\Sigma_i = \mathbf{0}$ and $\mu_i = \mathbf{0}$, collapses the stochastic dataset to the original deterministic one:

$$\mu(z;\theta_t) \;=\; -\frac{1}{n}\sum_{i=1}^{n}\Theta_f\big(z, x_i;\theta_t\big)\,\widehat{C}(x_i, y_i^*;\theta_t\big),$$

With the diffusion term also vanishing, the SDE degenerates exactly to the original deterministic kernel gradient flow derived in Section 4 of Jacot, Gabriel, and Hongler (2018).

4 Result 1: Training Dynamic Symmetries via NTK Fiber Bundles

In this section, we show that the training dynamics shown in Section 3 under Gaussian perturbations can influence the training dynamics of a neural network to either steer or freeze dynamics. In particular, we demonstrate how freezon dynamics can be obtained in augmentation directions that lie on the same kernel fiber of the NTK.

Architecture-Induced Fiber Bundles

To properly investigate steerable perturbation schemes, we first have to define which instances will act identically under the kernel. Thus, we define a level-set function $g(u; z, x, f)$:

$$g : \mathbb{R}^n \to \mathbb{R} \quad g(u; z, x, f) := \Theta_f(z, u) - \Theta_f(z, x)$$

which maps the deviation by u from x under $\Theta_f(z;\cdot)$. If we assume $g(x; z, x, f) = 0$ and $\nabla_{\mathbf{u}} g(x; z, x, f) \neq 0$, the implicit function theorem then guarantees that the preimage $g^{-1}(0)$ is a locally smooth submanifold near x.

Definition 4.1. Let $z, x \in \mathbb{R}^n$. A *kernel fiber* under the NTK Θ_f is the pullback of the level set $\{0\} \subset \mathbb{R}$ via the map g, defined as $\mathcal{F}_{z|x} = g^{-1}(0; z, x, f)$.

The kernel fiber $\mathcal{F}_{z|x}$ in Definition 4.1 encapsulates a continuous equivalence class of points in input space that are indistinguishable under the NTK Θ_f from the perspective of z (Figure 1). Because $\Theta_f(z,\cdot)$ is smooth and its gradient at $u = x \neq z$ is typically non-zero, the implicit–function theorem guarantees that the level set $\Theta_f(z, u) = \Theta_f(z, x)$ forms a smooth $(n-1)$-dimensional hypersurface through x. However, when $z = x$ the kernel attains its local maximum $\Theta_f(z, z)$ (Arora et al. 2019b); for a generic architecture any small displacement changes that value, so the level set collapses to the single point $\mathcal{F}_{x|x} = \{x\}$. Proposition 4.2 conjectures that *self fibers* like $\mathcal{F}_{x|x}$ are nontrivial only for networks that possess explicit symmetries, or perturbation invariances; this, we argue however, is a niche set of non-standard architectures. For example, circularly padded CNNs (Li et al. 2019) are structured to be invariant to cyclic shifts, and self-attention without positional embeddings is explicitly permutation invariant in token space (Vaswani et al. 2017).

Proposition 4.2 (Fiber degeneracy at the self–pair)**.** *Fix an input $x \in \mathbb{R}^n$, then the fiber collapses to the singleton $\{x\}$* iff *$\Theta(x, x)$ attains a strict, non-degenerate local maximum.*

Fiber-Aligned Noise Freezes Dynamics

The local tangent space $T_x\mathcal{F}_{z|x}$ (Definition 4.3) of $\mathcal{F}_{z|x}$ forms the orthogonal complement to the gradient vector $\nabla_{\mathbf{x}}\Theta_f(z, x)$, with $\dim(T_x\mathcal{F}_{z|x}) = n - 1$.

Definition 4.3. The *local kernel fiber tangent space* is the pullback of the tangent distribution under Θ_f, given by $T_x\mathcal{F}_{z|x} = \{v \in \mathbb{R}^n : v^\top \nabla_x \Theta_f(z, x) = 0\}$ which defines local perturbation directions that preserve the kernel.

Lemma 4.4 shows that perturbations sampled within $T_x\mathcal{F}_{z|x}$ preserve the function dynamics of the network at z. Perturbations $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ can be sampled freely within

the subspace formed by covariance $\mathbf{\Sigma}$ in Equation 10 for **any** covariance matrix M in the $\mathbb{R}^{n-1}$ subspace of $T_x\mathcal{F}_{z|x}$,

$$\mathbf{\Sigma} = PMP^\top, \quad P = I - \frac{vv^\top}{v^\top v}\bigg|_{v=\nabla_{\mathbf{x}}\Theta_f(z,x)}. \tag{10}$$

Lemma 4.4 (Freeze Dynamics under Fiber-Aligned Noise)**.** *Let $x \in \mathbb{R}^n$ and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ be a Gaussian perturbation, with $\mathrm{Im}(\mathbf{\Sigma}) \subseteq T_x\mathcal{F}_{z|x}$. Then, under first-order NTK dynamics, $\mathbb{E}[d_t f(z;\theta(t))] = d_t f(z;\theta(t))|_{\epsilon=0}$.*

Our result establishes that noise lying within the local kernel fiber tangent spaces freezes dynamics in function space at point z *regardless of whether it is in the training set.*

Noise Off the Fiber Steers Dynamics

Next, we characterize the impact of noise when perturbed points lying *off* of the local kernel fiber, revealing tunable drift effects. If we reconsider the level-set function g, any elements x in the pullback of any strictly positive level set $\{a|a > 0\} \subset \mathbb{R}$ via map g will result in the NTK increasing; likewise, any elements x in the pullback of any strictly negative level set $\{a|a < 0\} \subset \mathbb{R}$ via map g will result in the NTK decreasing. As a result, manipulation of the kernel's input allows us to *steer* the evolution of f at a singular point.

Lemma 4.5 shows that noise can be designed in order to influence the kernel to drive the function evolution at a point.

Lemma 4.5 (Tunable Drift from Orthogonal Noise)**.** *Let $\epsilon \sim \mathcal{N}(0, \mathbf{\Sigma})$, with $\mathrm{Im}(\mathbf{\Sigma}) \not\subseteq T_x\mathcal{F}_{z|x}$. Then the induced drift at point z during training is given explicitly by:*

$$\mu(z) = -\Theta_f(z,x)\widehat{C}(x,y) - \nabla_x\Theta_f(z,x)^\top \mathbf{\Sigma} \nabla_x \widehat{C}(x,y).$$

Perturbing in a direction orthogonal to $T_x F_{z|x}$ therefore moves the point (z, x) to a level-set with larger or smaller kernel value, as visualized by the blue arrow in Figure 1. The kernel gradient $\nabla_x\Theta_f(z,x)$ provides the outward normal; stepping along it strengthens the interaction between z and x, stepping against it weakens it. Definition 4.6 extends this to handle steering of dynamics across multiple points.

Definition 4.6 (Multi-Point Kernel Fiber)**.** Given sets $Z, X \subseteq \mathbb{R}^{n_0}$, define the multi-point kernel fiber as:

$$\mathcal{F}_{Z|X} = \bigcap_{z\in Z, x\in X} T_x\mathcal{F}_{z|x}$$

By considering multiple points, we elevate the power of this analysis to show that the structure of the noise can be partitioned: a block that projects onto the common tangent space keeps the predictions on Z invariant, whereas its orthogonal complement steers the network elsewhere.

Theorem 4.7 (Simultaneous Multi-Point Steering)**.** *Partition the input space into invariant and adaptive subsets relative to a reference set Z. Define the field $\mathbf{\Sigma}(x)$ by:*

$$\mathbf{\Sigma}(x) = \begin{cases} P_{x,Z}\mathbf{\Sigma}_W P_{x,Z}^\top, & x \in \mathcal{I}_{Z,\text{invariant}}, \\ Q_{x,Z}\mathbf{\Sigma}_{\text{out}} Q_{x,Z}^\top, & x \in \mathcal{I}_{Z,\text{adaptive}}, \end{cases}$$

where $P_{x,Z}$ and $Q_{x,Z}$ are projections onto $T_x\mathcal{F}_{Z|X}$ and its orthogonal complement, respectively. Then:

- *Dynamics at all points in Z remain simultaneously invariant under noise from $\mathcal{I}_{Z,\text{invariant}}$.*
- *Dynamics at points outside Z become explicitly tunable through noise from $\mathcal{I}_{Z,\text{adaptive}}$*

Intuitively, Theorem 4.7 lets us "freeze" the learned representation at a reference set Z by injecting only noise directions that stay inside every kernel fiber that passes through Z. In practice this means we can protect critical calibration points while still accelerating learning elsewhere. Thus, we re-frame stochastic data augmentation as a geometric mechanism for control over the function evolution of dynamics.

5 Result 2: Controlling Learning with Noise

Using our findings from Section 4, which provide a geometric foundation for noise to steer training, we propose 3 protocols for controlling training. Namely, in the form boundary sharpening, boundary flattening, and Gram recovery with novel datasets. While this analysis is preliminary, it demonstrates a promising new direction in which further work can be done.

Boundary Sharpening Noise

One of our proposed controllable training regimes is to sharpen the decision boundaries of the network's inference surface. In a classification setting, for example, this conceptually drives the model to produce inferences with greater confidence in a particular class. To enhance the neural net's discrimination near decision boundaries, we propose

$$\mathrm{Im}(\mathbf{\Sigma}(x)) = \frac{\nabla_x\Theta_f(z,x)\nabla_x\widehat{C}(x,y)^\top}{\|\nabla_x\Theta_f(z,x)\|^2}. \tag{11}$$

We show more rigorously how this noise sampling sharpens the decision boundaries in Theorem 5.1. Intuitively, we amplify the rate of the neural net's function dynamics in regions of high variance of the kernel than the rest of the space. This leads to the slope of the inference between the boundary of differing inference classes to become steeper. Ergo, the boundary becomes sharpened through this behavior.

Theorem 5.1 (Boundary Sharpening Dynamics)**.** *With the noise choice*

$$\mathrm{Im}(\mathbf{\Sigma}(x)) = \frac{\nabla_x\Theta_f(z,x)\nabla_x\widehat{C}(x,y)^\top}{\|\nabla_x\Theta_f(z,x)\|^2}$$

the drift at boundary points explicitly pushes model predictions towards class separability.

Boundary Flattening Noise

Opposite of boundary sharpening, here we aim to flatten the inference surface. We propose that the noise covariance in Equation 12 flattens the local gradient magnitude,

$$\mathrm{Im}(\mathbf{\Sigma}(x)) = \frac{\nabla_x f(x)\nabla_x f(x)^\top}{\|\nabla_x f(x)\|^2} \tag{12}$$

Theorem 5.2 (Local Inference Surface Flattening)**.** *The above covariance explicitly reduces the drift magnitude, effectively flattening the local inference surface.*

Novel Datasets That Produce Identical Dynamics

Lastly, we show that one can construct a completely novel dataset that yields identical, or *near-identical* dynamics. We show this by constructing a near-identical Gram matrix using an original dataset where every element is perturbed within its local kernel fiber tangent space $T_x\mathcal{F}_{z|x}$,

$$\Sigma_i = P_{V_i^\perp}\Sigma_W P_{V_i^\perp}^\top, \quad P_{V_i^\perp} = I - V_i(V_i^\top V_i)^{-1}V_i^\top, \tag{13}$$

where the matrix of kernel gradients is defined as:

$$V_i = [\nabla_{x_i}\Theta_f(x_1, x_i), \ldots, \nabla_{x_i}\Theta_f(x_N, x_i)].$$

This, in first-order, constrains that every perturbation lives in the orthogonal complement of the subspace spanned by V_i. Intuitively, this means we only move x_i in directions that are *orthogonal* to first-order variations of the NTK, so that the Gram matrix will remain unchanged to leading order. To make this claim precise, in Lemmas 5.3 and 5.4 we show the upper and lower bound of the Frobenius norm of a perturbed Gram matrix with bounded additive noise on the data set.

Lemma 5.3. *Let $H \in \mathbb{R}^{n\times n}$ be the Gram matrix defined by $H_{ij} = \Theta(x_i, x_j)$, and let $\hat{H} \in \mathbb{R}^{n\times n}$ be the perturbed Gram matrix with $\hat{H}_{ij} = \Theta(x_i + \delta_i, x_j + \delta_j)$, where $\|\delta_i\|_2 = \|\delta_j\|_2 = \ell$ for all i, j. Then the Frobenius norm difference satisfies the following upper bound $\|\hat{H} - H\|_F \le \left(\ell^2 D + O(\ell^4)\right)^{1/2}$ where*

$$D = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\|\nabla_x\Theta(x_i, x_j)\|_2^2 + \|\nabla_y\Theta(x_i, x_j)\|_2^2\right).$$

Lemma 5.4. *Let $H \in \mathbb{R}^{n\times n}$ be the NTK Gram matrix defined by $H_{ij} = \Theta(x_i, x_j)$, and let $\hat{H} \in \mathbb{R}^{n\times n}$ be the perturbed Gram matrix where each input x_i is replaced by $x_i + \delta_i$, with $\delta_i \sim \mathcal{N}(0, \Sigma_i)$. Define*

$$V_i := [\nabla_{x_i}\Theta(x_1, x_i), \ldots, \nabla_{x_i}\Theta(x_n, x_i)] \in \mathbb{R}^{d\times n}$$
$$P_{V_i^\perp} := I - V_i(V_i^\top V_i)^{-1}V_i^\top,$$

and set the noise covariance as

$$\Sigma_i := P_{V_i^\perp}\Sigma_W P_{V_i^\perp}^\top,$$

where $\Sigma_W \in \mathbb{R}^{d\times d}$ is any positive semi-definite matrix. Then the perturbed Gram matrix satisfies:

$$\mathbb{E}[\hat{H}_{ij}] = H_{ij} + O(\ell^2), \quad \textit{and} \quad \mathbb{E}\left[\|\hat{H} - H\|_F^2\right] = O(\ell^4).$$

Remark 5.5. The upper bound in Lemma 5.3 implies that for small perturbations ($\ell \to 0$), the change in the Gram matrix is approximately linear in ℓ. However, for larger perturbations, higher-order terms in the Taylor expansion become dominant, resulting in an approximately quadratic dependence on ℓ. In the case of the 2-layer ReLU case study, Lemma 6.1 shows an example of a network with no higher order terms.

The fourth-order error rate suggests that, at least in principle, one can fabricate an entirely new dataset whose induced kernel is indistinguishable—up to $\mathcal{O}(\ell^2)$ fluctuations—from that of the original data. However, this optimistic picture overlooks an important geometric constraint: if the sample size exceeds the ambient dimension, the perturbations cannot explore enough independent directions to match every entry of the $n \times n$ Gram matrix simultaneously. The final lemma, Lemma 5.7, formalizes this limitation.

Lemma 5.6. *Let $H, \hat{H} \in \mathbb{R}^{n\times n}$ be NTK Gram matrices constructed from inputs x_i and perturbed inputs $x_i + \delta_i$, with $\|\delta_i\|_2 = \ell$. Assume:*

(i) For all i, j, $\|\nabla_x\Theta(x_i, x_j)\|^2 + \|\nabla_y\Theta(x_i, x_j)\|^2 \ge c > 0$.

(ii) The perturbations δ_i are random, isotropic unit vectors scaled by ℓ.

Then

$$\mathbb{E}\left[\|\hat{H} - H\|_F\right] \ge \left(\frac{\ell^2 D}{d} - O(n^2\ell^4)\right)^{1/2}.$$

These observations naturally raise the question of whether our fibre-aligned construction in (13) enjoys the best of both worlds—namely, small *worst-case* bias *and* vanishing *average-case* distortion when the perturbations are themselves random. The next lemma confirms precisely this point: when the noise covariance is projected onto the orthogonal complement of the gradient fibres, the first-order term in the Taylor expansion of every Gram entry cancels *in expectation*, and the mean-squared error shrinks quadratically in ℓ.

Lemma 5.7 (Irreducible Frobenius Error when $n > d$)**.** *Let $H \in \mathbb{R}^{n\times n}$ be the NTK Gram matrix defined by $H_{ij} = \Theta(x_i, x_j)$, where Θ is a smooth, symmetric, positive semi-definite kernel and $x_i \in \mathbb{R}^d$ with $n > d$. Let each perturbed input be $\hat{x}_i = x_i + \delta_i$, where $\delta_i \sim \mathcal{N}(0, \Sigma_i)$ with $\operatorname{rank}(\Sigma_i) \le d$. Define the expected Gram matrix under perturbation:*

$$\mathcal{G}(\{\Sigma_i\}) := \mathbb{E}\left[(\Theta(x_i + \delta_i, x_j + \delta_j))_{i,j=1}^n\right],$$

and the manifold of achievable Gram matrices:

$$\mathcal{M} := \left\{\mathcal{G}(\{\Sigma_i\}) \;\middle|\; \Sigma_i \in \mathbb{R}^{d\times d},\ \Sigma_i \succeq 0,\ \operatorname{rank}(\Sigma_i) \le d\right\}.$$

Then the irreducible Frobenius error due to limited perturbation dimensionality is:

$$\varepsilon_{\min} := \inf_{A\in\mathcal{M}} \|A - H\|_F^2.$$

Moreover, when the perturbation scale is $\|\delta_i\| = \ell \ll 1$, we have the lower bound:

$$\varepsilon_{\min} \ge c(n^2 - nd)\ell^2,$$

for some constant $c > 0$.

The subsequent irreducible-error lemma (Lemma 5.7) shows, however, that this invariance is ultimately limited by geometry when the sample size exceeds the input dimension.

6 Case Study on the 2-layer ReLU Network

This section instantiates the general results of Sections 4 and 5, as well as an analytical solution for $\nabla_x \Theta$ and $\mathcal{F}_{z|x}$, for the simplest neural net: a fully-connected network with a single hidden layer of *infinite* width and ReLU activations.

In this setting the infinite-width NTK corresponding to a 2-layer neural net with 1 infinitely-wide hidden layer and ReLU activation, analytically derived by (Arora et al. 2019a), as shown in Equation 14,

$$\Theta^{\infty}_{\text{2-ReLU}}(\mathbf{z}, \mathbf{x}) = \frac{\mathbf{z}^\top \mathbf{x}(\pi - \arccos(\mathbf{z}^\top \mathbf{x}))}{2\pi} \tag{14}$$

reveals elegant geometric properties of the kernel fibers in the input space formed by $\Theta^{\infty}_{\text{2-ReLU}}$ because it depends on $\mathbf{z}$ and $\mathbf{x}$ only through their inner product. We show in Lemma 6.1 that the kernel fibers of $\Theta^{\infty}_{\text{2-ReLU}}$ are level sets of $\mathbf{z}^\top \mathbf{x}$. Hence, membership in a fiber can be tested with a single dot product.

Lemma 6.1. *For an NTK Θ defined by a 2-Layer ReLU network, a vector $\mathbf{u}$ lies in the orthogonal subspace formed by $\nabla_{\mathbf{z}}\Theta(\mathbf{z}, \mathbf{x})$ if and only if $\mathbf{x}$ and $\mathbf{u}$ are orthogonal.*

Hence, for this architecture there is an *isomorphism* between (i) the subspace orthogonal to $\nabla_{\mathbf{z}}\Theta$ in parameter space and (ii) the subspace of input vectors orthogonal to $\mathbf{x}$. Moreover, $\nabla_{\mathbf{z}}\Theta$ itself points in the direction of z.

When studying curvature it is useful to examine the mixed derivative $\nabla_{\mathbf{z}}\nabla_{\mathbf{x}}\Theta(\mathbf{z}, \mathbf{x})$. Remarkably, Lemma 6.2 shows that this operator possesses only two distinct eigenvalues, with all but one being identical. This shows the curvature of this mixed hessian has remarkably low-rank with only 2 principal directions independent of the input space dimension.

Lemma 6.2. *For an NTK Θ defined by a 2-Layer ReLU network, $\nabla_{\mathbf{z}}\nabla_{\mathbf{x}}\Theta(z, x)$ will only have 2 distinct eigenvalues, with all but one being identical.*

Moreover, Lemma 6.3 reveals that the conditioning number – which in this case encapsulates the entire eigen spectrum – deteriorates as the angle between $\mathbf{z}$ and $\mathbf{x}$ shrinks:

Lemma 6.3. *The condition number of $\nabla_z \nabla_x \Theta$ becomes unstable the stronger the cosine similarity is between z and x in the feature space.*

Collectively, Lemmas 6.1–6.3 provide a more complete characterization of the geometric and spectral behavior of the NTK for the two-layer ReLU model.

7 Empirical Validation

In the following experiments, we empirically validate the claims made in Sections 4 and 5 using finite-width networks. While the present experiments are preliminary, they reveal measurable control over toy datasets and small scale benchmarks, and suggest a path towards scalable plug-and-play controls for neural net training. *Additional experimental results are provided in the Appendix.*

Gram Invariance

We first empirically validate how well our Gram preservation technique in Equation 13 approaches the lower bound described in Lemma 5.7. To do so, we consider a training set sampled from a Gaussian distribution and a 2-layer ReLU neural net, and construct its Gram matrix $\mathbf{H}$. Then we purposely perturb the original dataset with Gaussian noise with a covariance that is either (i) constructed by the local kernel fiber tangent space (Eq. 10), (ii) constructed to align with the NTK's gradient, or (iii) completely uniform $\mathcal{N}(0, I)$. Using the perturbed dataset, we construct a new Gram $\widehat{\mathbf{H}}$ and measure its difference from the original Gram matrix with the Frobenius norm: $||\widehat{\mathbf{H}} - \mathbf{H}||_F$.

The empirical results, as shown in Figure 3, show the mean and 95% C.I. of $||\widehat{\mathbf{H}} - \mathbf{H}||_F$ across datasets of varying volume, dimensionality, and magnitude of added noise from 100 different random trials. Our results shows 3 key findings: (a): noise aligned with the NTK's gradient diverges the Gram difference at a greater rate than a random Gaussian, (b): noise within the NTK's local tangent fiber yields a slower rate of Gram divergence than a random Gaussian, and (c): noise within the NTK's local tangent fiber lies close to the optimal lower bound. In short, these empirical findings (a) and (b) align precisely with our theoretical claims. Surprisingly, we see that the local tangent space, a first-order approximation of orthogonal noise, yields remarkably near-optimal results even with large magnitudes of noise.

Sculpting of Decision Boundaries

Next, we validate our key claim stemming from our analysis: perturbing training data from well-designed noise is sufficient alone for the sculpting of training trajectories of neural nets. To empirically demonstrate this claim, we empirically test our derived techniques to sharpen (Equation 11) and flatten (Equation 12) boundaries of Section 5. We sample training data from a sine curve, and purposefully remove training instances lying around $x = 0$. We then train a deep neural net under 3 different regimes: (i) standard gradient descent (SGD), (ii) SGD + perturbations to influence sharpening, and (iii) SGD + perturbations to influence flattening. If there were sufficiently sampled training instances, the ground-truth inference slope $||\nabla_{\mathbf{x}} f(x)||_2$ at $x = 0$ should be 1. Our results

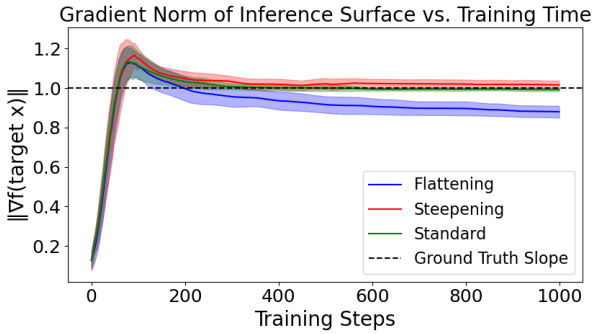
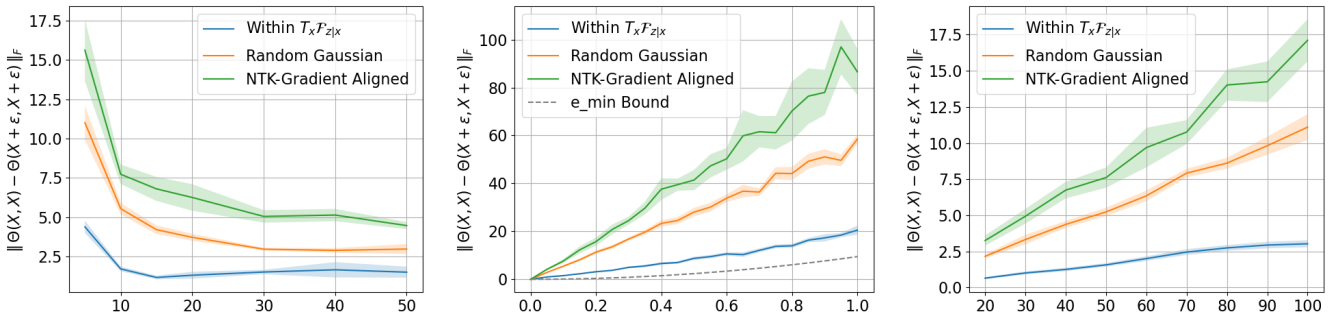


Figure 2: The difference between the inference slope at $x = 0$ when training a simple 2-layer neural net (5 trials) with (green) no augmentations, (blue) flattening augmentations, or (red) sharpening augmentations. This is evidence in support of our covariance construction can indeed affect learned inference surfaces with NTK-driven input perturbations.



(a) Input Dimension (b) Noise Magnitude (c) Training Set Volume

Figure 3: 2-Layer neural net with hidden width $w = 50$. The Frobenius norm of the difference between a training Gram matrix and a training Gram where training elements were perturbed with Gaussian noise with a covariance that is either (blue) constructed by the local kernel fiber tangent space (Eq. 10), (orange) uniform, or (green) constructed to align with the NTK's gradient. Across varying input dimensions, noise magnitudes, and training set volumes, we show evidence in support of our analysis that first-order approximations of noise that fall on the kernel fiber reduce the Frobenius norm. Additionally, noise that falls orthogonal to the kernel fiber exacerbates the Frobenius norm.

Augmentation Technique	$\nabla_{\mathbf{x}}\Theta$ **Alignment**	**Relative Level Set Deviation**
Rotation	$-0.062644_{\pm 0.023}$	$0.092406_{\pm 0.085}$
Hflip	$-0.080352_{\pm 0.020}$	$0.023625_{\pm 0.021}$
Vflip	$-0.092738_{\pm 0.021}$	$0.042702_{\pm 0.034}$
Cutout	$-0.008870_{\pm 0.016}$	$1.708538_{\pm 1.121}$
Mixup	$-0.104793_{\pm 0.031}$	$0.227315_{\pm 0.239}$
Uniform Gaussian	$-0.003340_{\pm 0.006}$	$0.211346_{\pm 0.086}$
$\boldsymbol{T_x\mathcal{F}_{z\|x}}$ (Eq. 10)	$-0.001090_{\pm 0.003}$	$0.022298_{\pm 0.010}$
$\boldsymbol{\nabla_{\mathbf{x}}\Theta}$ (Ground Truth)	$0.994879_{\pm 0.015}$	$0.189625_{\pm 0.149}$

Table 1: The mean cosine similarity ($\nabla_{\mathbf{x}}\Theta$ *Alignment*) and relative level set deviation (Equation 15) on instances in CIFAR-10 under different augmentation techniques against reference points z lying in the same class. Results computed with ResNet18. Existing augmentations like rotation, hflip, vflip yield low level set deviation, serving as poor techniques drivers for steerable training. On the otherhand, Cutout, Mixup, and uniform Gaussian yield higher level set deviation, yet on average do not align well with $\nabla_{\mathbf{x}}\Theta$, making them choatic drivers for steerable training.

in Figure 2 show that our baseline method (green line) learns the ground-truth slope even without any training data in that region. More interesting though, our proposed method to flatten (blue line) does indeed have a distinct trend in lowering the inference surface slope. For our method to sharpen (red line), its ability to sharpen the boundary is marginal, yet distributionally distinct from standard training. Thus, we show one can in fact steer the final inference surface by sculpting the training behavior with noise alone.

Measuring Steerability of Existing Augmentations

Lastly, we validate our theoretical claims on a sample of $1,000$ instances of CIFAR-10 (Krizhevsky, Hinton et al. 2009) with ResNet18 (He et al. 2016) against a suite of classical data augmentation techniques: random Gaussian noise, vertical/horizontal image flips, random image rotations, MixUp (Zhang et al. 2017), and CutOut (DeVries and Taylor 2017). In Table 1, for each instance x and its augmentation $\hat{x}$ according to one of the augmentation technique, we show the mean relative deviation of $\hat{x}$ from the x's level set:

$$\text{Relative Level Set Deviation} = \frac{\|\Theta(z,x) - \Theta(z,\hat{x})\|}{\Theta(z,x)} \tag{15}$$

as well as the cosine of its augmentation with the local NTK gradient $\nabla_{\mathbf{x}}\Theta$. We show that existing augmentation techniques, while pragmatically fruitful improving robustness, either (a) yield little difference in level set deviation, yet are not as orthogonal to $\nabla_{\mathbf{x}}\Theta$ as first order approximations of the Kernel Fiber $T_x\mathcal{F}_{z|x}$, or (b) yield great level set difference, but in a fashion that cannot be locally reasoned about. Further research in this direction to design novel augmentation strategies shows future promise in yielding precise control over training dynamics with NTK-driven data augmentations.

8 Conclusion

In this paper, we significantly expand the theoretical framework of Neural Tangent Kernel (NTK) analysis to address the complex dynamics of data augmentation that occur in modern neural network training. Building upon the classical infinite-width NTK paradigm, we introduced a stochastic differential equation (SDE) formulation that captures the impact of noise on the evolution of neural network functions. By formalizing the symmetries of the NTK as fiber bundles, we provided new insights into how these symmetries can shape training trajectories and improve convergence behavior. We further developed a detailed analysis of inference surface dynamics, demonstrating how data augmentations and noise can regularize the learning process and enhance generalization. We posit that further research extending this work shows promise in harnessing greater power in steering neural net training.

Acknowledgements

Results in this paper were obtained in part using a high-performance computing system acquired through NSF MRI grant DMS-1337943 to WPI. This research was supported by NSF under grant NRT-HDR-2021871.

References

Arora, S.; Du, S.; Hu, W.; Li, Z.; and Wang, R. 2019a. Fine-grained analysis of optimization and generalization for over-parameterized two-layer neural networks. In *International Conference on Machine Learning*, 322–332. PMLR.

Arora, S.; Du, S. S.; Hu, W.; Li, Z.; Salakhutdinov, R. R.; and Wang, R. 2019b. On exact computation with an infinitely wide neural net. *Advances in neural information processing systems*, 32.

Bordelon, B.; Canatar, A.; and Pehlevan, C. 2020. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, 1024–1034. PMLR.

Bronstein, M. M.; Bruna, J.; LeCun, Y.; Szlam, A.; and Vandergheynst, P. 2017. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4): 18–42.

DeVries, T.; and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Fort, S.; Hu, H.; and Lakshminarayanan, B. 2019. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*.

Gerken, J. E.; and Kessel, P. 2024. Emergent equivariance in deep ensembles. *arXiv preprint arXiv:2403.03103*.

Godfrey, C.; Brown, D.; Emerson, T.; and Kvinge, H. 2022. On the symmetries of deep learning models and their internal representations. *Advances in Neural Information Processing Systems*, 35: 11893–11905.

Halbouni, A.; Gunawan, T. S.; Habaebi, M. H.; Halbouni, M.; Kartiwi, M.; and Ahmad, R. 2022. Machine learning and deep learning approaches for cybersecurity: A review. *IEEE Access*, 10: 19572–19585.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hron, J.; Bahri, Y.; Sohl-Dickstein, J.; and Novak, R. 2020. Infinite attention: NNGP and NTK for deep attention networks. In *International Conference on Machine Learning*, 4376–4386. PMLR.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Jacot, A.; Gabriel, F.; and Hongler, C. 2018. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Li, Z.; Wang, R.; Yu, D.; Du, S. S.; Hu, W.; Salakhutdinov, R.; and Arora, S. 2019. Enhanced convolutional neural tangent kernels. *arXiv preprint arXiv:1911.00809*.

Martens, J. 2020. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146): 1–76.

Mei, S.; and Montanari, A. 2019. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*.

Miotto, R.; Wang, F.; Wang, S.; Jiang, X.; and Dudley, J. T. 2018. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6): 1236–1246.

Misof, P.; Kessel, P.; and Gerken, J. E. 2025. Equivariant Neural Tangent Kernels. In *Forty-second International Conference on Machine Learning*.

Novak, R.; Xiao, L.; Bahri, Y.; Lee, J.; Yang, G.; Hron, J.; Abolafia, D. A.; Pennington, J.; and Sohl-Dickstein, J. 2019. Neural tangent kernels for convolutional neural networks. *arXiv preprint arXiv:1902.04760*.

Perin, A.; and Deny, S. 2024. On the Ability of Deep Networks to Learn Symmetries from Data: A Neural Kernel Theory. *arXiv preprint arXiv:2412.11521*.

Refinetti, M.; Ingrosso, A.; and Goldt, S. 2023. Neural networks trained with sgd learn distributions of increasing complexity. In *International Conference on Machine Learning*, 28843–28863. PMLR.

Shan, H.; and Bordelon, B. 2021. A theory of neural tangent kernel alignment and its influence on training. *arXiv preprint arXiv:2105.14301*.

Tsilivis, N.; and Kempe, J. 2022. What can the neural tangent kernel tell us about adversarial robustness? *Advances in Neural Information Processing Systems*, 35: 18116–18130.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yang, G. 2020. Tensor programs II: Neural tangent kernel for any architecture. *arXiv preprint arXiv:2006.14548*.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.