A Special Case of Nonlinear Extrapolation Under ReLU and the Neural Tangent Kernel

Abiel J. Kim

November 25, 2025

Abstract

It has been demonstrated both theoretically and empirically that the ReLU MLP tends to extrapolate linearly for an out-of-distribution evaluation point. The machine learning literature provides ample analysis with respect to the mechanisms to which linearity is induced. However, the analysis of extrapolation at the origin under the NTK regime remains a more unexplored special case. In particular, the infinite-dimensional feature map induced by the neural tangent kernel is not translationally invariant. This means that the study of an out-of-distribution evaluation point very far from the origin is not equivalent to the evaluation of a point very near the origin. And since the feature map is rotation invariant, these two special cases may represent the most canonically extreme bounds of ReLU NTK extrapolation. Ultimately, it is this loose recognition of the two special cases of extrapolation that motivate the discovery of quadratic extrapolation for an evaluation close to the origin.

1 Introduction

The work of Xu et al. [11] proves that an over-parameterized ReLU-activated multilayer perceptron (MLP) will extrapolate linearly when evaluated along any direction very distant from the origin. They formally prove extrapolative linearity by analysis of the learned regressor's functional form in the *neural tangent kernel* (NTK) *reproducing kernel hilbert space* (RKHS) [3]. And, since the infinite dimensional feature map induced by the neural tangent kernel is rotation invariant, the analysis covers the generalizable case of an evaluation point very distant from the origin. However, it is not difficult to recognize that the same feature map is not translation invariant. It is by a geometric reasoning that the origin of the RKHS must be a distinct special case whose analysis departs from Theorem 1 of Xu et al. [11]. That is, in the limit of a large relative distance between the training point set and the evaluation point, one observes that there must be two special locations of the evaluation point with respect to the NTK induced feature map: A location casted along a singular feature direction, and a location which intersects all feature directions.

It is this recognition of the distinguishable cases that motivates the extrapolative analysis at the origin location. The non translation invariance of the feature map implies that the extrapolative analysis at the origin and far from origin are not equivalent problems. It can be reasoned that they are two canonical cases of a more complete analysis of extrapolation. However, inducing extrapolation at the origin must be done carefully to ensure that the evaluation data is pushed out of the support of the training distribution space. This is achieved by this paper's definition of a labeled training set, which is formally presented in the problem setup of section 2. The desired effect of said definition is to induce a problem setup where all members of the training set are sent infinitely far away from the origin whilst fixing the evaluation data at the origin. Under this variant setting, we state Theorem 1, which discovers that an over-parameterized neural network extrapolates quadratically when evaluated near the origin. This finding contrasts, but does not conflict with, Xu et al. [11], which contrastingly concerns itself an evaluation point far from the origin.

The paper is organized as follows. The proof of Theorem 1 is presented in §A.4 and will depend on the results of Lemmas 1 and 2, which are proven with continuity in §A.2 and §A.3 respectively. Our problem setup induces a special case of the NTK gram matrix which must be studied in §A.1 to set the stage for the remainder of the mathematics.

2 Preliminaries

Background on NTK: Suppose that a neural network performs nonlinear regression $f(\boldsymbol{\theta}, \boldsymbol{x}) : \mathcal{X} \to \mathbb{R}$ where $\boldsymbol{\theta}$ is a vectorization of the network parameters and $\boldsymbol{x} \in \mathcal{X}$. Let there be n training points which form a labeled set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. If we train the network on the labeled set to minimize the squared loss function $\frac{1}{2}\sum_{i=1}^n \left(f(\boldsymbol{\theta}, \boldsymbol{x}_i) - y_i\right)^2$ via gradient descent, then we can derive a kernel method from the network by first considering an affine approximation of the network output in parameter space. If we denote the time-dependent parameter vector induced by gradient descent as $\boldsymbol{\theta}^{(t)}$ for some iteration t, then we define the feature map $\phi(\boldsymbol{x})$ as the gradient of the network output with respect to $\boldsymbol{\theta}$ evaluated at $\boldsymbol{\theta}^{(0)}$ denoted as $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(0)}, \boldsymbol{x})$. The corresponding kernel, called the *neural tangent kernel* (NTK), is then an affine model that is linear in the network parameters. Under particular constraints such as the infinite width and infinitesimal learning rate, the NTK becomes an expectation:

$$NTK(\boldsymbol{x}_i, \boldsymbol{x}_j) = \mathbb{E}_{\boldsymbol{\theta}\sim\mathcal{N}} \left\langle \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(0)}, \boldsymbol{x}_i), \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(0)}, \boldsymbol{x}_j) \right\rangle,$$

where the expectation emerges by the law of large numbers induced by the network's infinite width. Interestingly, the affine approximation is correct under the NTK constraints in parameter space, and is closely tied to the network's notion of *lazy training*. Ultimately, since training is linear in the often high-dimensional, possibly infinite, feature space, the neural network behaves as an affine kernel regression. We take all such pairwise NTK evaluations from the labeled training set to produce the positive semi-definite NTK gram matrix denoted as NTK_{train}.

Background on Neural Network Extrapolation: Xu et al. [11] builds on the established results of the NTK equivalence between neural network training and kernel regression to more precisely analyze extrapolation. However, using the NTK directly requires analysis of the point-wise form as a kernel regression fit over the labeled training set. It can be more advantageous to work in the NTK induced feature space instead to derive a functional representation of the learned network, which may be more analytically manageable. This is precisely the route they take and formalize this equivalence between point-wise NTK regression and the learned function in the NTK induced feature space in their Lemma 2:

$$\begin{aligned} f_{NTK}(\boldsymbol{x}) &= \phi(\boldsymbol{x})^\top \boldsymbol{\beta}_{\text{NTK}} \\ \text{where} \quad \boldsymbol{\beta}_{\text{NTK}} &= \min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_2 \\ &\text{s.t.} \quad \phi(\boldsymbol{x})^\top \boldsymbol{\beta} = y_i \quad \text{for } i = 1, \ldots, n, \end{aligned}$$

where $f_{NTK}(\boldsymbol{x}) = \phi(\boldsymbol{x})^\top \boldsymbol{\beta}_{\text{NTK}}$ is the min-norm functional form equivalent to NTK kernel regression fitted over the training data for any $\boldsymbol{x} \in \mathcal{X}$. Further, they derive the precise closed-form of the NTK induced feature map for a ReLU two-layer MLP in their Lemma 3:

$$\phi(\boldsymbol{x}) = c'\Big(\boldsymbol{x} \cdot \mathbb{I}\Big(\boldsymbol{w}^{(k)\top} \boldsymbol{x} \geq 0\Big),\ \boldsymbol{w}^{(k)\top} \boldsymbol{x} \cdot \mathbb{I}\Big(\boldsymbol{w}^{(k)\top} \boldsymbol{x} \geq 0\Big),\ \ldots\Big),$$

where c' is a constant, $\mathbb{I}$ is the Heaviside indicator function, and $\boldsymbol{w}^{(k)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ for $k \to \infty$. By analyzing the functional representation in the NTK RKHS, they discovered that for a labeled training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and evaluation point $\boldsymbol{x}_0 = t\boldsymbol{v}$ for any direction $\boldsymbol{v} \in \mathbb{R}^d$, the network converges to a linear function.

Problem Setup: Our problem setup inherits from both Jacot, Gabriel, and Hongler [3] and Xu et al. [11] primarily through the notation of the latter for compatibility. Let $\mathcal{X}$ be a d-dimensional Euclidean input space and φ be a set of n training inputs such that $\varphi = \{\boldsymbol{x}_i\}_{i=1}^n$ with $\boldsymbol{x}_i \in \mathcal{X}$ for $i \in [n]$. If we translate φ by the vector $-t\boldsymbol{v}_\varphi$ for any direction $\boldsymbol{v}_\varphi$, then we will have formed a new set $\varphi^\infty = \{\boldsymbol{x}_i - t\boldsymbol{v}_\varphi : \boldsymbol{x}_i \in \varphi\}$ where a member is denoted $\boldsymbol{x}_i^\infty = \boldsymbol{x}_i - t\boldsymbol{v}_\varphi$ for any $\boldsymbol{x}_i^\infty \in \varphi^\infty$. The labeled training set can then be constructed as $\{(\boldsymbol{x}_i^\infty, y_i^\infty)\}_{i=1}^n$ where $y_i^\infty = g(\boldsymbol{x}_i^\infty)$ for target function $g : \mathcal{X} \to \mathbb{R}$. We train a single-output two-layer ReLU MLP $f_{NTK} : \mathcal{X} \to \mathbb{R}$ in the NTK regime using gradient descent to minimize the squared loss function over the labeled training set. We reintroduce the hat notation which denotes a data vector augmented with bias term: $\hat{\boldsymbol{x}} = [\boldsymbol{x}|1]$. We also introduce the check notation which denotes the explicit exclusion of the bias weight with respect to the k-th hidden neuron, $\check{\boldsymbol{w}}^{(k)} \in \mathbb{R}^d$.

Clarifying the Training Data: Please agree that the definition of the labeled training set, which is constructed from φ^∞, facilitates an analysis of extrapolation at the origin location. If extrapolation can be configured by defining the labeled training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and the evaluation point $\boldsymbol{x}_0 = t\boldsymbol{v}$ for any direction $\boldsymbol{v}$, then it is not difficult to attain the "inverted" setup of the new training set $\{(\boldsymbol{x}_i - t\boldsymbol{v},\ g(\boldsymbol{x}_i - t\boldsymbol{v}))\}_{i=1}^n$ and evaluation point $\boldsymbol{x}_0 = \boldsymbol{0}$. Critically, the coordinate shift preserves the sufficient condition that induces extrapolation as $t \to \infty$.

Clarifying the Notation: This paper refers to the set φ as a (point) realization, which is a convention related to point processes and stochastic geometry [1]. Since the NTK regime deals with finite datasets, it may be useful to explicitly describe or analyze φ as a point process in a later related work. The set φ may be ascribed an underlying mathematical data generator for the purposes of a neural scaling law analysis, for instance.

3 Theoretical Contributions

Remark 1. *If all n inputs of the training set φ^∞ are located infinitely far from the origin along the same direction, then the asymptotic pseudo-inverse of the NTK gram matrix is a difference between the identity and all-ones matrix: $\frac{1}{\delta}\boldsymbol{I} - \frac{t^2\kappa}{\delta(n\kappa t^2+\delta)}\boldsymbol{J}$, where $\delta \to 0$, $t \to \infty$, and κ is a constant.*

Special Case of the NTK Gram: We discover this closed-form in §A.1 by first recognizing that under the definition of φ^∞, the indicators for any training input become *input agnostic* insofar that the indicating logic strictly depends on a feature direction $\boldsymbol{w}$ and training direction $\boldsymbol{v}_\varphi$. The definition of φ^∞ induces the otherwise singular asymptotic NTK gram matrix $\kappa t^2\boldsymbol{J}$. We then use Tikhonov regularization to pseudo-invert this asymptotic NTK gram expressed as $\left(\kappa t^2\boldsymbol{J} + \boldsymbol{\Gamma}\right)^{-1}$. We leverage this special case of the asymptotic NTK gram matrix induced by φ^∞ and its pseudo-inverse to express the components of $\boldsymbol{\beta}_{NTK}$ induced by training inputs φ^∞ that are distant from the origin. These results are then used to prove Lemma 2 and ultimately Theorem 1.

Theorem 1. *An over-parameterized two-layer ReLU MLP $f_{NTK} : \mathbb{R}^d \to \mathbb{R}$ that is trained on a labeled set $\{(\boldsymbol{x}_i^\infty, y_i^\infty)\}_{i=1}^n$ with $\boldsymbol{x}_i^\infty = \boldsymbol{x}_i - t\boldsymbol{v}_\varphi$ for $\boldsymbol{x}_i \in \mathcal{X}$ and any direction $\boldsymbol{v}_\varphi$ in the NTK regime minimizing squared loss will converge to a quadratic extrapolator when evaluated at a point near the origin $\mathbf{0}$ as $t \to \infty$.*

Theorem 1 Proof Sketch: Theorem 1 is the main contribution of this paper and states that an extremely wide NTK predictor with ReLU activations that is trained on a dataset which is extremely distant from the origin will converge to a quadratic extrapolator when evaluated near the origin. That is, the Theorem 1 states that the predictor's first and second directional derivatives exist and all higher-order derivatives are 0. And the proof of Theorem 1, which is presented in §A.4, depends on the results of Lemmas 1 and 2. Lemma 1 is a generalized algebraic manipulation and states that the directional derivative of the NTK predictor can be expressed in terms of the derivatives of the indicator. The significance of Lemma 1 is most clear when we leverage the Dirac-delta's so called *sifting property*, also known as the *sampling property*. We note that the derivative of the Heaviside indicator is the Dirac-delta, and applies itself nicely when viewing the predictor's derivative as an integral. Lemma 2 completes the second half of the Theorem 1 proof by stating that the partial derivatives of the beta components with respect to the bias component of a feature direction $\boldsymbol{w}_{d+1}$ are always 0 for any order derivative past the second. The significance of Lemma 2 is clear when we see in §A.2 that the z-th derivative of the predictor depends on the $(z-1)$-th and $(z-2)$-th partial derivatives of the beta components. It is not difficult to see that the quadratic-order persists when taking the z-th derivative of f_{NTK}.

Lemma 1. *The feature map of the z-th directional derivative of f_{NTK} for any direction $\boldsymbol{v}_0$ can be expressed in terms of the z-th and $(z-1)$-th directional derivatives of the indicator for $\boldsymbol{v}_0$ such that:*

$$D_{\boldsymbol{v}}^z f_{NTK}(\boldsymbol{x}_0) = \boldsymbol{\beta}_{NTK}^\top \left(\hat{\boldsymbol{x}}_0 \cdot D_{\boldsymbol{v}}^z \mathbb{I}^{(k)} - z\hat{\boldsymbol{v}} \cdot D_{\boldsymbol{v}}^{z-1} \mathbb{I}^{(k)},\ \boldsymbol{w}^{(k)\top}\hat{\boldsymbol{x}}_0 \cdot D_{\boldsymbol{v}}^z \mathbb{I}^{(k)} - z\boldsymbol{w}^{(k)\top}\hat{\boldsymbol{v}} \cdot D_{\boldsymbol{v}}^{z-1} \mathbb{I}^{(k)},\ ...\right)$$

Lemma 2. *The components of the NTK representation coefficient $\boldsymbol{\beta}_{NTK}$ induced by a training input set $\varphi^\infty = \{\boldsymbol{x}_i^\infty\}_{i=1}^n$ where $\boldsymbol{x}_i^\infty = \boldsymbol{x}_i - t\boldsymbol{v}_\varphi$ for some $\boldsymbol{x}_i \in \mathcal{X}$ and any direction $\boldsymbol{v}_\varphi$ are constant with respect to the bias component of any given feature direction $\boldsymbol{w}_{d+1}$ such that:*

$$\frac{\partial^z \boldsymbol{\beta}_{\boldsymbol{w}}^1}{\partial \boldsymbol{w}_{d+1}^z}, \frac{\partial^z \boldsymbol{\beta}_{\boldsymbol{w}}^2}{\partial \boldsymbol{w}_{d+1}^z} = 0 \text{ for all } z \geq 1.$$

4 Conclusion

This paper proves that the over-parameterized ReLU MLP extrapolates non-linearly for the special case at origin in the RKHS. More specifically, this paper finds that at the origin, neural network extrapolation behaves like a quadratic function. However, the quadratic function is highly dependent on the degree of similarity between vector orientations $\boldsymbol{w}$ - which represents the direction of a feature - and $\boldsymbol{v}$ - the vector which defines the evaluation point. If, for instance, the two orientations are orthogonal, then the second derivative is unconditionally zero. The second derivative may also be zero dependent on the beta components, i.e., if the beta 1 component is orthogonal to $\boldsymbol{v}$ and the beta 2 component is zero; However, this condition is less strict. The results are distinct from but complementary to the existing ML literature which primarily concern the linearity of neural network extrapolation. That is, since the feature map induced by the neural tangent kernel is not translation invariant, extrapolation at a point far from the origin is not equivalent to extrapolation a point close to the origin. We prove our results by determining a closed-form of the asymptotic pseudo-inverse NTK gram matrix to determine the components of $\boldsymbol{\beta}_{NTK}$ induced by the definition of φ^∞. Then we discover a neat algebraic trick in Lemma 1 to rewrite the directional derivative of the predictor as partial derivatives of the beta components using the distributional derivative equivalence to the directional derivative of the indicator.

References

[1] Sung Nok Chiu et al. *Stochastic geometry and its applications*. John Wiley & Sons, 2013.

[2] Martin Haenggi. *Stochastic geometry for wireless networks*. Cambridge University Press, 2013.

[3] Arthur Jacot, Franck Gabriel, and Clément Hongler. "Neural Tangent Kernel: Convergence and Generalization in Neural Networks". In: *CoRR* abs/1806.07572 (2018). arXiv: 1806.07572. URL: http://arxiv.org/abs/1806.07572.

[4] Sun Yuan Kung. *Kernel methods and machine learning*. Cambridge University Press, 2014.

[5] Pierluigi Maponi. "The solution of linear systems by using the Sherman–Morrison formula". In: *Linear algebra and its applications* 420.2-3 (2007), pp. 276–294.

[6] AR Murugan, CG Moorthy, and CT Ramasamy. "A definition of dirac delta functions". In: *Adv. Math. Sci. J* 9 (2020), pp. 1213–1220.

[7] ABO OM. "Ridge regression and inverse problems". In: *Stockholm University, Department of Mathematics* (2001).

[8] Akshay Rangamani, Lorenzo Rosasco, and Tomaso Poggio. *For interpolating kernel machines, minimizing the norm of the ERM solution minimizes stability*. 2020. arXiv: 2006.15522 [stat.ML]. URL: https://arxiv.org/abs/2006.15522.

[9] A Salah. "delta function and Heaviside function". In: *Indian Institute of space science and technology* (2015).

[10] Eric W Weisstein. "Heaviside step function". In: *https://mathworld. wolfram. com/* (2002).

[11] Keyulu Xu et al. "How Neural Networks Extrapolate: From Feedforward to Graph Neural Networks". In: *CoRR* abs/2009.11848 (2020). arXiv: 2009.11848. URL: https://arxiv.org/abs/2009.11848.

A Proofs

A.1 Special Case of the NTK Gram Matrix

We begin our analysis by making clear the form of $\boldsymbol{\beta}$, which is the coefficient vector in the NTK RKHS that is fit over the labeled training data. We begin with the point-wise form of NTK regression to write $\boldsymbol{\beta}$ in terms of the NTK gram:

$$f_{NTK}(\hat{\boldsymbol{x}}) = \left(\langle\phi(\hat{\boldsymbol{x}}), \phi(\hat{\boldsymbol{x}}_1^\infty)\rangle, \ldots, \langle\phi(\hat{\boldsymbol{x}}), \phi(\hat{\boldsymbol{x}}_n^\infty)\rangle\right)^\top \cdot \boldsymbol{NTK}_{train}^{-1}\boldsymbol{Y} \tag{1}$$

$$= \phi(\hat{\boldsymbol{x}})^\top \boldsymbol{\Phi}_{train}^\top \boldsymbol{NTK}_{train}^{-1}\boldsymbol{Y} \tag{2}$$

$$= \phi(\hat{\boldsymbol{x}})^\top \boldsymbol{\beta}. \tag{3}$$

Attaining a closed form expression of $\boldsymbol{NTK}_{train}^{-1}$ is a desirable but non-trivial analysis. Fortunately, later in this section, we will see how the definition of φ^∞ induces the NTK gram to a closed-form asymptotic pseudo-inverse. But first, we recognize the application of Tikhonov regularization, which ensures the invertibility of the NTK gram matrix and induces a choice of $\boldsymbol{\beta}$ equivalent to the min-norm definition of the unique $\boldsymbol{\beta}_{NTK}$. Tikhonov regularization was chosen for its simple usage but is also an approach supported by Rangamani, Rosasco, and Poggio [8]. We express $\boldsymbol{\beta}_{NTK}$ in terms of the Tikhonov regularized NTK gram matrix:

$$\boldsymbol{\beta}_{NTK} = \boldsymbol{\Phi}_{train}^\top \left(\boldsymbol{NTK}_{train} + \boldsymbol{\Gamma}\right)^{-1}\boldsymbol{Y} \tag{4}$$

for Tikhonov matrix $\boldsymbol{\Gamma} = \delta\boldsymbol{I}, \delta \to 0^+$. Before we solve for the pseudo-inverse of $\boldsymbol{NTK}_{train}$, we take note of the induced behavior of the indication function for a training data point under the definition of φ^∞, where we find that the indication depends solely on the dot product between any particular feature direction $\boldsymbol{w}$ and the special $\boldsymbol{v}_\varphi$ that translates φ. In other words, under the definition of φ^∞, ReLU indicators for any training data point $\boldsymbol{x}_i^\infty \in \varphi^\infty$ become input agnostic insofar that they become independent from $\boldsymbol{x}_i \in \varphi$:

$$\mathbb{I}\left(\boldsymbol{w}^\top \hat{\boldsymbol{x}}_i^\infty \geq 0\right) \tag{5}$$

$$= \mathbb{I}\left(\boldsymbol{w}^\top (\hat{\boldsymbol{x}}_i - t\hat{\boldsymbol{v}}) \geq 0\right) \tag{6}$$

$$= \mathbb{I}\left(\boldsymbol{w}^\top (-\hat{\boldsymbol{v}}) \geq 0\right). \tag{7}$$

The independence that arises between the indicators and the training inputs is a crucial insight and will be a recurring assistance that enables the pseudo-inversion of the NTK gram. Speaking of which, by definition of the neural tangent

kernel, the (i, j)-th entry of $\boldsymbol{NTK}_{train}$ can be expressed as:

$$\boldsymbol{NTK}_{train}[i,j] = \left\langle \phi(\hat{\boldsymbol{x}}_i^{\infty}), \phi(\hat{\boldsymbol{x}}_j^{\infty}) \right\rangle \tag{8}$$

$$= \int \hat{\boldsymbol{x}}_i^{\infty} \cdot \hat{\boldsymbol{x}}_j^{\infty} \cdot \mathbb{I}\left(\boldsymbol{w}^{\top} \hat{\boldsymbol{x}}_i^{\infty} \geq 0\right) \cdot \mathbb{I}\left(\boldsymbol{w}^{\top} \hat{\boldsymbol{x}}_j^{\infty} \geq 0\right) \tag{9}$$

$$+ (\boldsymbol{w}^{(k)^{\top}} \hat{\boldsymbol{x}}_i^{\infty}) \cdot (\boldsymbol{w}^{(k)^{\top}} \hat{\boldsymbol{x}}_j^{\infty}) \cdot \mathbb{I}\left(\boldsymbol{w}^{\top} \hat{\boldsymbol{x}}_i^{\infty} \geq 0\right) \cdot \mathbb{I}\left(\boldsymbol{w}^{\top} \hat{\boldsymbol{x}}_j^{\infty} \geq 0\right) d\mathbb{P}(\boldsymbol{w}) \tag{10}$$

for any pair $(\boldsymbol{x_i^{\infty}}, \boldsymbol{x_j^{\infty}})$ taken from the labeled training set. We observe the emergence of an indication pair in lines (9)-(10). But since indicators become input agnostic, we greatly simplify their indicating logic using equation (7):

$$\mathbb{I}\left(\boldsymbol{w}^{\top} \hat{\boldsymbol{x}}_i^{\infty} \geq 0\right) \cdot \mathbb{I}\left(\boldsymbol{w}^{\top} \hat{\boldsymbol{x}}_j^{\infty} \geq 0\right) \tag{11}$$

$$= \mathbb{I}\left(\boldsymbol{w}^{\top} \hat{\boldsymbol{x}}_i - t(\boldsymbol{w}^{\top} \hat{\boldsymbol{v}}) \geq 0\right) \cdot \mathbb{I}\left(\boldsymbol{w}^{\top} \hat{\boldsymbol{x}}_j - t(\boldsymbol{w}^{\top} \hat{\boldsymbol{v}}) \geq 0\right) \tag{12}$$

$$= \mathbb{I}\left(\boldsymbol{w}^{\top}(-\hat{\boldsymbol{v}}) \geq 0\right)^2 \tag{13}$$

$$= \mathbb{I}\left(\boldsymbol{w}^{\top}(-\hat{\boldsymbol{v}}) \geq 0\right). \tag{14}$$

Then we apply the definition of φ^{∞} to expand the dot product in equation (9):

$$\hat{\boldsymbol{x}}_i^{\infty} \cdot \hat{\boldsymbol{x}}_j^{\infty} \tag{15}$$

$$= (\hat{\boldsymbol{x}}_i - t\hat{\boldsymbol{v}}) \cdot (\hat{\boldsymbol{x}}_j - t\hat{\boldsymbol{v}}) \tag{16}$$

$$= \hat{\boldsymbol{v}}^2 t^2 - (\hat{\boldsymbol{x}}_i + \hat{\boldsymbol{x}}_j) \cdot \hat{\boldsymbol{v}} t + \hat{\boldsymbol{x}}_i \cdot \hat{\boldsymbol{x}}_j, \tag{17}$$

as well as the dot product in equation (10):

$$(\boldsymbol{w}^{\top} \hat{\boldsymbol{x}}_i^{\infty}) \cdot (\boldsymbol{w}^{\top} \hat{\boldsymbol{x}}_j^{\infty}) \tag{18}$$

$$= (\boldsymbol{w}^{\top}(\hat{\boldsymbol{x}}_i - t\hat{\boldsymbol{v}})) \cdot (\boldsymbol{w}^{\top}(\hat{\boldsymbol{x}}_j - t\hat{\boldsymbol{v}}) \tag{19}$$

$$= (\boldsymbol{w}^{\top} \hat{\boldsymbol{v}})^2 t^2 - \boldsymbol{w}^{\top} \hat{\boldsymbol{v}} (\boldsymbol{w}^{\top} \hat{\boldsymbol{x}}_i + \boldsymbol{w}^{\top} \hat{\boldsymbol{x}}_j) t + (\boldsymbol{w}^{\top} \hat{\boldsymbol{x}}_i) \cdot (\boldsymbol{w}^{\top} \hat{\boldsymbol{x}}_j), \tag{20}$$

to rewrite the (i, j)-th entry of the NTK gram matrix using lines (14), (17), and (20) as:

$$\boldsymbol{NTK}_{train}[i,j] \tag{21}$$

$$= t^2 \int (\hat{\boldsymbol{v}}^2 + (\boldsymbol{w}^{\top} \hat{\boldsymbol{v}})^2) \cdot \mathbb{I}\left(\boldsymbol{w}^{\top}(-\hat{\boldsymbol{v}}) \geq 0\right) d\mathbb{P}(\boldsymbol{w}) \tag{22}$$

$$- t \int ((\hat{\boldsymbol{x}}_i + \hat{\boldsymbol{x}}_j) \cdot \hat{\boldsymbol{v}} + \boldsymbol{w}^{\top} \hat{\boldsymbol{v}} (\boldsymbol{w}^{\top} \hat{\boldsymbol{x}}_i + \boldsymbol{w}^{\top} \hat{\boldsymbol{x}}_j)) \cdot \mathbb{I}\left(\boldsymbol{w}^{\top}(-\hat{\boldsymbol{v}}) \geq 0\right) d\mathbb{P}(\boldsymbol{w}) \tag{23}$$

$$+ \int (\hat{\boldsymbol{x}}_i \cdot \hat{\boldsymbol{x}}_j + (\boldsymbol{w}^{\top} \hat{\boldsymbol{x}}_i) \cdot (\boldsymbol{w}^{\top} \hat{\boldsymbol{x}}_j)) \cdot \mathbb{I}\left(\boldsymbol{w}^{\top}(-\hat{\boldsymbol{v}}) \geq 0\right) d\mathbb{P}(\boldsymbol{w}). \tag{24}$$

The quadratic form that emerges in lines (22)-(24) is a direct consequence of applying the definition of φ^{∞}; It defines the NTK gram matrix induced by the limiting training set. It is a beautiful structure because the leading-order term is the only term in the quadratic form that does not depend on indices i and j. Without this particular structure, pseudo-inverting the matrix $(\boldsymbol{NTK}_{train} + \boldsymbol{\Gamma})$ for a closed-form would be a more difficult analysis. Since line (22) is the leading order term, the resulting form intuitively suggests that as φ is shifted further from the origin along some direction $\boldsymbol{v}_{\varphi}$, the kernel regression solution depends less on the inputs of φ and more on the direction $\boldsymbol{v}_{\varphi}$:

$$\boldsymbol{NTK}_{train}[i,j] \asymp t^2 \kappa, \tag{25}$$

where κ is a constant equal to the integral of line (22). Therefore, in the limit as $t \to \infty$, we find that the (i, j)-th entry of the NTK gram does not depend on φ. The asymptotic form is then a constant matrix, meaning that $\boldsymbol{NTK}_{train}[i,j]$ is constant for any pair (i, j). We can finally invert the regularized NTK gram from line (4) as:

$$(\boldsymbol{NTK}_{train} + \boldsymbol{\Gamma})^{-1} \tag{26}$$

$$\asymp \left(t^2 \kappa \boldsymbol{J} + \boldsymbol{\Gamma}\right)^{-1} \tag{27}$$

$$= \left(\delta \boldsymbol{I} + (t^2 \mathbf{1})(\kappa \mathbf{1})^{\top}\right)^{-1} \tag{28}$$

$$= \frac{1}{\delta} \boldsymbol{I} - \frac{t^2 \kappa}{\delta(n\kappa t^2 + \delta)} \boldsymbol{J}, \tag{29}$$

where $\boldsymbol{J}[i,j] = 1$ for any pair of indices (i,j). The penultimate equality has $\delta\boldsymbol{I}$ from our definition of $\boldsymbol{\Gamma}$ with the outer product between $t^2\mathbf{1}$ and $\kappa\mathbf{1}$. In the final equality, one inverts the matrix using the Sherman-Morrison formula [5]. It follows from line (29) that the (i,j)-th entry of the NTK gram asymptotic pseudo-inverse is:

$$\left(\boldsymbol{NTK}_{train} + \boldsymbol{\Gamma}\right)^{-1}[i,j] \asymp \begin{cases} -\frac{\kappa t^2}{\delta(n\kappa t^2+\delta)}, & \text{if } i \neq j \\ \frac{1}{\delta} - \frac{\kappa t^2}{\delta(n\kappa t^2+\delta)}, & \text{if } i = j \end{cases} \tag{30}$$

Using the piecewise definition of equation (30), let $\boldsymbol{\alpha}_{NTK} \asymp \left(\frac{1}{\delta}\boldsymbol{I} - \frac{t^2\kappa}{\delta(n\kappa t^2+\delta)}\boldsymbol{J}\right)\boldsymbol{Y}$ denote the matrix-vector product between the label vector $\boldsymbol{Y}$ and the asymptotic pseudo-inverse. Note that $\boldsymbol{\alpha}_{NTK}$ is sub-scripted as such so that the applied regularization is explicit. It is not difficult to calculate the closed-form of the i-th entry of $\boldsymbol{\alpha}_{NTK}$:

$$\boldsymbol{\alpha}_{NTK}[i] = \left(\frac{1}{\delta}\boldsymbol{I} - \frac{t^2\kappa}{\delta(n\kappa t^2+\delta)}\boldsymbol{J}\right)[i] \cdot \boldsymbol{Y} \tag{31}$$

$$= \sum_{j=1}^{n}\left(\frac{1}{\delta}\boldsymbol{I} - \frac{t^2\kappa}{\delta(n\kappa t^2+\delta)}\boldsymbol{J}\right)[i,j] \cdot \boldsymbol{Y}[j] \tag{32}$$

$$= \sum_{j=1}^{n}\left(-\frac{t^2\kappa}{\delta(n\kappa t^2+\delta)}\right) g(\hat{\boldsymbol{x}}_j^\infty) + \frac{1}{\delta}g(\hat{\boldsymbol{x}}_i^\infty) \tag{33}$$

$$= -\frac{t^2\kappa}{\delta(n\kappa t^2+\delta)}\sum_{j=1}^{n} g(\hat{\boldsymbol{x}}_j^\infty) + \frac{1}{\delta}g(\hat{\boldsymbol{x}}_i^\infty). \tag{34}$$

And, it should be made clear the values of the $\boldsymbol{\beta}$ components. There are two components associated with a feature direction $\boldsymbol{w}^{(k)}$ for any k. We follow the notation of Xu et al. [11] and denote the first (vector) beta component as $\boldsymbol{\beta}_{\boldsymbol{w}}^1$ and the second (scalar) beta component as $\boldsymbol{\beta}_{\boldsymbol{w}}^2$ denoting $\boldsymbol{w}$ as a shorthand for any particular $\boldsymbol{w}^{(k)}$. See line (38) below:

$$= \boldsymbol{\Phi}_{train}^\top \boldsymbol{\alpha} \tag{35}$$

$$= \boldsymbol{\alpha}_1\phi(\hat{\boldsymbol{x}}_1^\infty) + \boldsymbol{\alpha}_2\phi(\hat{\boldsymbol{x}}_2^\infty) + ... + \boldsymbol{\alpha}_n\phi(\hat{\boldsymbol{x}}_n^\infty) \tag{36}$$

$$= \boldsymbol{\alpha}_1\begin{bmatrix} \hat{\boldsymbol{x}}_1^\infty \cdot \mathbb{I}\left(\boldsymbol{w}^{(k)^\top}\hat{\boldsymbol{x}}_1^\infty \geq 0\right) \\ \boldsymbol{w}^{(k)^\top}\hat{\boldsymbol{x}}_1^\infty \cdot \mathbb{I}\left(\boldsymbol{w}^{(k)^\top}\hat{\boldsymbol{x}}_1^\infty \geq 0\right) \\ \vdots \end{bmatrix} + ... + \boldsymbol{\alpha}_n\begin{bmatrix} \hat{\boldsymbol{x}}_n^\infty \cdot \mathbb{I}\left(\boldsymbol{w}^{(k)^\top}\hat{\boldsymbol{x}}_n^\infty \geq 0\right) \\ \boldsymbol{w}^{(k)^\top}\hat{\boldsymbol{x}}_n^\infty \cdot \mathbb{I}\left(\boldsymbol{w}^{(k)^\top}\hat{\boldsymbol{x}}_n^\infty \geq 0\right) \\ \vdots \end{bmatrix} \tag{37}$$

$$= \begin{bmatrix} \boldsymbol{\alpha}_1\hat{\boldsymbol{x}}_1^\infty\mathbb{I}\left(\boldsymbol{w}^{(k)^\top}\hat{\boldsymbol{x}}_1^\infty \geq 0\right) + ... + \boldsymbol{\alpha}_n\hat{\boldsymbol{x}}_n^\infty\mathbb{I}\left(\boldsymbol{w}^{(k)^\top}\hat{\boldsymbol{x}}_n^\infty \geq 0\right) \\ \boldsymbol{\alpha}_1\boldsymbol{w}^{(k)^\top}\hat{\boldsymbol{x}}_1^\infty\mathbb{I}\left(\boldsymbol{w}^{(k)^\top}\hat{\boldsymbol{x}}_1^\infty \geq 0\right) + ... + \boldsymbol{\alpha}_n\boldsymbol{w}^{(k)^\top}\hat{\boldsymbol{x}}_n^\infty\mathbb{I}\left(\boldsymbol{w}^{(k)^\top}\hat{\boldsymbol{x}}_n^\infty \geq 0\right) \\ \boldsymbol{\alpha}_1\hat{\boldsymbol{x}}_1^\infty\mathbb{I}\left(\boldsymbol{w}^{(k+1)^\top}\hat{\boldsymbol{x}}_1^\infty \geq 0\right) + ... + \boldsymbol{\alpha}_n\hat{\boldsymbol{x}}_n^\infty\mathbb{I}\left(\boldsymbol{w}^{(k+1)^\top}\hat{\boldsymbol{x}}_n^\infty \geq 0\right) \\ \boldsymbol{\alpha}_1\boldsymbol{w}^{(k+1)^\top}\hat{\boldsymbol{x}}_1^\infty\mathbb{I}\left(\boldsymbol{w}^{(k+1)^\top}\hat{\boldsymbol{x}}_1^\infty \geq 0\right) + ... + \boldsymbol{\alpha}_n\boldsymbol{w}^{(k+1)^\top}\hat{\boldsymbol{x}}_n^\infty\mathbb{I}\left(\boldsymbol{w}^{(k+1)^\top}\hat{\boldsymbol{x}}_n^\infty \geq 0\right) \\ \vdots \end{bmatrix}. \tag{38}$$

It follows from line (38) that the components of $\boldsymbol{\beta}_{NTK}$ can be written as:

$$\boldsymbol{\beta}_{\boldsymbol{w}}^1 = \sum_{i=1}^{n}\boldsymbol{\alpha}_{NTK}[i]\,\hat{\boldsymbol{x}}_i^\infty\,\mathbb{I}\left(\boldsymbol{w}^\top\hat{\boldsymbol{x}}_i^\infty \geq 0\right) \tag{39}$$

$$\boldsymbol{\beta}_{\boldsymbol{w}}^2 = \sum_{i=1}^{n}\boldsymbol{\alpha}_{NTK}[i]\,\boldsymbol{w}^\top\hat{\boldsymbol{x}}_i^\infty\,\mathbb{I}\left(\boldsymbol{w}^\top\hat{\boldsymbol{x}}_i^\infty \geq 0\right). \tag{40}$$

Lastly, we use lines (7), (34), and the definition of φ^∞, to finally rewrite equations (39)-(40) for a closed-form of the first and second beta components that are induced by the definition of φ^∞:

$$\boldsymbol{\beta}_{\boldsymbol{w}}^1 = \mathbb{I}\left(\boldsymbol{w}^\top(-\hat{\boldsymbol{v}}) \geq 0\right) \cdot \left(C(t,\delta,\kappa)\sum_{j=1}^{n} g(\hat{\boldsymbol{x}}_i^\infty)\sum_{i=1}^{n}\hat{\boldsymbol{x}}_i^\infty + \frac{1}{\delta}\sum_{i=1}^{n}\hat{\boldsymbol{x}}_i^\infty g(\hat{\boldsymbol{x}}_i^\infty)\right) \tag{41}$$

$$\boldsymbol{\beta}_{\boldsymbol{w}}^2 = \mathbb{I}\left(\boldsymbol{w}^\top(-\hat{\boldsymbol{v}}) \geq 0\right) \cdot \left(C(t,\delta,\kappa)\sum_{j=1}^{n} g(\hat{\boldsymbol{x}}_i^\infty)\sum_{i=1}^{n}\boldsymbol{w}^\top\hat{\boldsymbol{x}}_i^\infty + \frac{1}{\delta}\sum_{i=1}^{n}\boldsymbol{w}^\top\hat{\boldsymbol{x}}_i^\infty g(\hat{\boldsymbol{x}}_i^\infty)\right) \tag{42}$$

where $C(t, \delta, \kappa) = -\frac{t^2\kappa}{\delta(n\kappa t^2+\delta)}$ with $t \to \infty$, $\delta \to 0^+$, and κ depending on $\boldsymbol{w}$. One final note as an aside is that we can digress into a separate but related analysis if we take the target g to be linear, i.e., we can apply the equivalence $g(\hat{\boldsymbol{x}}_i^{\infty}) = g(\hat{\boldsymbol{x}}_i) - tg(\hat{\boldsymbol{v}}_{\varphi})$ to lines (41)-(42) and analyze unrelated forms. Although this is irrelevant to the paper, it may lead to an alternate proof of somewhat interesting findings.

A.2 Proof of Lemma 1

If $\boldsymbol{x}_0 \in \mathcal{X}$ is an evaluation point then let $\boldsymbol{x}_1 = \boldsymbol{x}_0 + h\boldsymbol{v}$ for some direction $\boldsymbol{v}$. We can compute the z-th directional derivative of f_{NTK} recursively using the standard limit definition:

$$D_{\boldsymbol{v}}^z f_{NTK}(\boldsymbol{x}_0) = \lim_{h\to 0} \frac{D_{\boldsymbol{v}}^{z-1} f_{NTK}(\boldsymbol{x}_1) - D_{\boldsymbol{v}}^{z-1} f_{NTK}(\boldsymbol{x}_0)}{h}. \tag{43}$$

Using the definition of f_{NTK} we can expand the numerator of equation (43) for the first directional derivative:

$$f_{NTK}(\hat{\boldsymbol{x}}_1) - f_{NTK}(\hat{\boldsymbol{x}}_0) \tag{44}$$

$$= \boldsymbol{\beta}_{NTK}^{\top} \left(\hat{\boldsymbol{x}}_1 \cdot \mathbb{I}_1^{(k)} - \hat{\boldsymbol{x}}_0 \cdot \mathbb{I}_0^{(k)}, {\boldsymbol{w}^{(k)}}^{\top} \hat{\boldsymbol{x}}_1 \cdot \mathbb{I}_1^{(k)} - {\boldsymbol{w}^{(k)}}^{\top} \hat{\boldsymbol{x}}_0 \cdot \mathbb{I}_0^{(k)}, \dots\right), \tag{45}$$

where $\mathbb{I}_0^{(k)} = \mathbb{I}\left({\boldsymbol{w}^{(k)}}^{\top} \hat{\boldsymbol{x}}_0 \geq 0\right)$, $\mathbb{I}_1^{(k)} = \mathbb{I}\left({\boldsymbol{w}^{(k)}}^{\top} \hat{\boldsymbol{x}}_1 \geq 0\right)$, ... are defined for notational brevity. The numerator for the second directional derivative is similarly expanded, omitting an h which has been factored out:

$$(f_{NTK}(\hat{\boldsymbol{x}}_2) - f_{NTK}(\hat{\boldsymbol{x}}_1)) - (f_{NTK}(\hat{\boldsymbol{x}}_1) - f_{NTK}(\hat{\boldsymbol{x}}_0)) \tag{46}$$

$$= \boldsymbol{\beta}_{NTK}^{\top} \left((\hat{\boldsymbol{x}}_2 \cdot \mathbb{I}_2^{(k)} - \hat{\boldsymbol{x}}_1 \cdot \mathbb{I}_1^{(k)}) - (\hat{\boldsymbol{x}}_1 \cdot \mathbb{I}_1^{(k)} - \hat{\boldsymbol{x}}_0 \cdot \mathbb{I}_0^{(k)}), \right. \tag{47}$$

$$\left.({\boldsymbol{w}^{(k)}}^{\top} \hat{\boldsymbol{x}}_2 \cdot \mathbb{I}_2^{(k)} - {\boldsymbol{w}^{(k)}}^{\top} \hat{\boldsymbol{x}}_1 \cdot \mathbb{I}_1^{(k)}) - {\boldsymbol{w}^{(k)}}^{\top} \hat{\boldsymbol{x}}_1 \cdot \mathbb{I}_1^{(k)} - {\boldsymbol{w}^{(k)}}^{\top} \hat{\boldsymbol{x}}_0 \cdot \mathbb{I}_0^{(k)}, \dots\right) \tag{48}$$

$$= \boldsymbol{\beta}_{NTK}^{\top} \left(\hat{\boldsymbol{x}}_2 \cdot \mathbb{I}_2^{(k)} - 2\hat{\boldsymbol{x}}_1 \cdot \mathbb{I}_1^{(k)} + \hat{\boldsymbol{x}}_0 \cdot \mathbb{I}_0^{(k)}, \right. \tag{49}$$

$$\left.{\boldsymbol{w}^{(k)}}^{\top} \hat{\boldsymbol{x}}_2 \cdot \mathbb{I}_2^{(k)} - 2{\boldsymbol{w}^{(k)}}^{\top} \hat{\boldsymbol{x}}_1 \cdot \mathbb{I}_1^{(k)} + {\boldsymbol{w}^{(k)}}^{\top} \hat{\boldsymbol{x}}_0 \cdot \mathbb{I}_0^{(k)}, \dots\right), \tag{50}$$

and so forth. The point is that the z-th directional derivative of f_{NTK} will contain the terms $\boldsymbol{x}_0, \boldsymbol{x}_1, ..., \boldsymbol{x}_z$ where $\boldsymbol{x}_z = \boldsymbol{x}_0 + zh\boldsymbol{v}$ where we repeatedly differentiate along the same direction $\boldsymbol{v}$.

Next, let $\Sigma^{(z)}_{\hat{\boldsymbol{x}},\mathbb{I}^{(k)}}$ be defined as:

$$\Sigma^{(z)}_{\hat{\boldsymbol{x}},\mathbb{I}^{(k)}} = P_z^{(z)} \hat{\boldsymbol{x}}_z \mathbb{I}_z^{(k)} + P_{z-1}^{(z)} \hat{\boldsymbol{x}}_{z-1} \mathbb{I}_{z-1}^{(k)} + ... + P_0^{(z)} \hat{\boldsymbol{x}}_0 \mathbb{I}_0^{(k)}, \tag{51}$$

where the coefficients $P_z^{(z)}, P_{z-1}^{(z)}, ..., P_0^{(z)}$ represent the sign-alternating Pascal coefficients of the z-th line in a 0-indexed Pascal triangle, e.g., $P_1^{(1)} = 1$ and $P_0^{(1)} = -1$. We can now generally rewrite the z-th directional derivative of f_{NTK} using equation (51) as:

$$D_{\boldsymbol{v}}^z f(\hat{\boldsymbol{x}}_0) = \lim_{h\to 0} \frac{\boldsymbol{\beta}_{NTK}^{\top} \left(\Sigma^{(z)}_{\hat{\boldsymbol{x}},\mathbb{I}^{(k)}}, {\boldsymbol{w}^{(k)}}^{\top} \Sigma^{(z)}_{\hat{\boldsymbol{x}},\mathbb{I}^{(k)}}, ...\right)}{h^z}. \tag{52}$$

The last definition in preparation for the proof of Lemma 1 will be the sum $\Sigma^{(z)}_{\mathbb{I}^{(k)}}$, defined in terms of the indicators:

$$\Sigma^{(z)}_{\mathbb{I}^{(k)}} = P_z^{(z)} \mathbb{I}_z^{(k)} + P_{z-1}^{(z)} \mathbb{I}_{z-1}^{(k)} + ... + P_0^{(z)} \mathbb{I}_0^{(k)}. \tag{53}$$

Lemma 1. *The feature map of the z-th directional derivative of f_{NTK} for any direction $\boldsymbol{v}_0$ can be expressed in terms of the z-th and $(z-1)$-th directional derivatives of the indicator for $\boldsymbol{v}_0$ such that:*

$$D_{\boldsymbol{v}}^z f_{NTK}(\boldsymbol{x}_0) = \boldsymbol{\beta}_{NTK}^{\top} \left(\hat{\boldsymbol{x}}_0 \cdot D_{\boldsymbol{v}}^z \mathbb{I}^{(k)} - z\hat{\boldsymbol{v}} \cdot D_{\boldsymbol{v}}^{z-1} \mathbb{I}^{(k)}, {\boldsymbol{w}^{(k)}}^{\top} \hat{\boldsymbol{x}}_0 \cdot D_{\boldsymbol{v}}^z \mathbb{I}^{(k)} - z{\boldsymbol{w}^{(k)}}^{\top} \hat{\boldsymbol{v}} \cdot D_{\boldsymbol{v}}^{z-1} \mathbb{I}^{(k)}, ...\right)$$

Proof. The first term of $\Sigma^{(z)}_{\hat{\boldsymbol{x}},\mathbb{I}^{(k)}}$ is $P_z^{(z)} \hat{\boldsymbol{x}}_z \mathbb{I}_z^{(k)}$. Since $P_z^{(z)}$ always lies on the left edge of the Pascal triangle, we always have $P_z^{(z)} \hat{\boldsymbol{x}}_z \mathbb{I}_z^{(k)} = \hat{\boldsymbol{x}}_z \mathbb{I}_z^{(k)}$. We use the trick

$$\hat{\boldsymbol{x}}_z \mathbb{I}_z^{(k)} \tag{54}$$

$$= \hat{\boldsymbol{x}}_z (\mathbb{I}_z^{(k)} + (\Sigma^{(z)}_{\mathbb{I}^{(k)}} - \mathbb{I}_z^{(k)}) - (\Sigma^{(z)}_{\mathbb{I}^{(k)}} - \mathbb{I}_z^{(k)})) \tag{55}$$

$$= \hat{\boldsymbol{x}}_z \cdot \Sigma^{(z)}_{\mathbb{I}^{(k)}} - \hat{\boldsymbol{x}}_z \cdot (\Sigma^{(z)}_{\mathbb{I}^{(k)}} - \mathbb{I}_z^{(k)}) \tag{56}$$

so that

$$\Sigma^{(z)}_{\hat{\boldsymbol{x}},\mathbb{I}^{(k)}} \tag{57}$$

$$= \hat{\boldsymbol{x}}_z \cdot \Sigma^{(z)}_{\mathbb{I}^{(k)}} - \hat{\boldsymbol{x}}_z \cdot (\Sigma^{(z)}_{\mathbb{I}^{(k)}} - \mathbb{I}^{(k)}_z) + (\Sigma^{(z)}_{\hat{\boldsymbol{x}},\mathbb{I}^{(k)}} - \hat{\boldsymbol{x}}_z \mathbb{I}^{(k)}_z) \tag{58}$$

$$= \hat{\boldsymbol{x}}_z \cdot \Sigma^{(z)}_{\mathbb{I}^{(k)}} + \Sigma^{(z)}_{\hat{\boldsymbol{x}},\mathbb{I}^{(k)}} - \hat{\boldsymbol{x}}_z \cdot \Sigma^{(z)}_{\mathbb{I}^{(k)}} \tag{59}$$

$$= \hat{\boldsymbol{x}}_z \cdot \Sigma^{(z)}_{\mathbb{I}^{(k)}} + (P^{(z)}_{z-1} \hat{\boldsymbol{x}}_{z-1} \mathbb{I}^{(k)}_{z-1} - P^{(z)}_{z-1} \hat{\boldsymbol{x}}_z \mathbb{I}^{(k)}_{z-1}) + ... + (P^{(z)}_0 \hat{\boldsymbol{x}}_0 \mathbb{I}^{(k)}_0 - P^{(z)}_0 \hat{\boldsymbol{x}}_z \mathbb{I}^{(k)}_0) \tag{60}$$

$$= \hat{\boldsymbol{x}}_z \cdot \Sigma^{(z)}_{\mathbb{I}^{(k)}} + P^{(z)}_{z-1} \mathbb{I}^{(k)}_{z-1} (\hat{\boldsymbol{x}}_{z-1} - \hat{\boldsymbol{x}}_z) + ... + P^{(z)}_0 \mathbb{I}^{(k)}_0 (\hat{\boldsymbol{x}}_0 - \hat{\boldsymbol{x}}_z) \tag{61}$$

$$= \hat{\boldsymbol{x}}_z \cdot \Sigma^{(z)}_{\mathbb{I}^{(k)}} + P^{(z)}_{z-1} \mathbb{I}^{(k)}_{z-1} ([-h\boldsymbol{v} \,|\, 0]) + ... + P^{(z)}_0 \mathbb{I}^{(k)}_0 ([-zh\boldsymbol{v} \,|\, 0]) \tag{62}$$

$$= \hat{\boldsymbol{x}}_z \cdot \Sigma^{(z)}_{\mathbb{I}^{(k)}} - \left(\sum_{i=0}^{z-1} P^{(z)}_i \mathbb{I}^{(k)}_i (z-i) h[\boldsymbol{v} \,|\, 0] \right) \tag{63}$$

$$= \hat{\boldsymbol{x}}_z \cdot \Sigma^{(z)}_{\mathbb{I}^{(k)}} - \left(h[\boldsymbol{v} \,|\, 0] \sum_{i=0}^{z-1} P^{(z)}_i \mathbb{I}^{(k)}_i (z-i) \right) \tag{64}$$

$$= \hat{\boldsymbol{x}}_z \cdot \Sigma^{(z)}_{\mathbb{I}^{(k)}} - \left(h[\boldsymbol{v} \,|\, 0] \sum_{i=0}^{z-1} z P^{(z-1)}_i \mathbb{I}^{(k)}_i \right) \tag{65}$$

$$= \hat{\boldsymbol{x}}_z \cdot \Sigma^{(z)}_{\mathbb{I}^{(k)}} - zh[\boldsymbol{v} \,|\, 0] \cdot \Sigma^{(z-1)}_{\mathbb{I}^{(k)}} \tag{66}$$

where we use the algebraic trick of lines (54)-(56) to get a secondary trick on line (59), which would be otherwise more difficult to see. Then, lines (60)-(62) follow from the definitions of lines (51) and (53). Another critical step in the proof sequence is the penultimate equality which realizes the equivalence between $P^{(z)}_i(z-i)$ and $zP^{(z-1)}_i$ for $i = 0, ..., z-1$. This equivalence finds a correspondence from a coefficient on the z-th line of the Pascal triangle to the number on the left in the previous $(z-1)$-th line. And since the equivalence is defined for $i = 0, ..., z-1$, it is well-defined because the correspondence from a coefficient $P^{(z)}_i$ to the preceding coefficient $P^{(z-1)}_i$ on the left of the triangle is only undefined when $i = z$ which would be out of bounds with respect to an indexing error on the $(z-1)$-th line. Also note that this equivalence implicitly requires $z \geq 1$ since we are computing derivatives.

The significance of the result on line (66) is that we can reformulate equation (52) to be expressed in terms of the definition (53), effectively re-expressing the binomial expansion coefficients, which comes from the limit definition of the directional derivative of f_{NTK}, in terms of the indicators. And since we only manipulated equation (51), the limit is now taken with respect to the indicators. The point is that we can now write the z-th derivative of f_{NTK} in terms of the z-th and $(z-1)$-th directional derivatives of $\mathbb{I}$:

$$D^z_{\boldsymbol{v}} f(\hat{\boldsymbol{x}}_0) \tag{67}$$

$$= \lim_{h \to 0} \frac{\boldsymbol{\beta}^\top_{\text{NTK}}}{h^z} \big(\hat{\boldsymbol{x}}_z \cdot \Sigma^{(z)}_{\mathbb{I}^{(k)}} - zh\hat{\boldsymbol{v}} \cdot \Sigma^{(z-1)}_{\mathbb{I}^{(k)}},\, \boldsymbol{w}^{(k)\top} (\hat{\boldsymbol{x}}_z \cdot \Sigma^{(z)}_{\mathbb{I}^{(k)}} - zh\hat{\boldsymbol{v}} \cdot \Sigma^{(z-1)}_{\mathbb{I}^{(k)}}),\, ... \big) \tag{68}$$

$$= \lim_{h \to 0} \boldsymbol{\beta}^\top_{\text{NTK}} \big(\hat{\boldsymbol{x}}_z \cdot (\Sigma^{(z)}_{\mathbb{I}^{(k)}}/h^z) - z\hat{\boldsymbol{v}} \cdot (\Sigma^{(z-1)}_{\mathbb{I}^{(k)}}/h^{z-1}),\, \boldsymbol{w}^{(k)\top} (\hat{\boldsymbol{x}}_z \cdot (\Sigma^{(z)}_{\mathbb{I}^{(k)}}/h^z) - z\hat{\boldsymbol{v}} \cdot (\Sigma^{(z-1)}_{\mathbb{I}^{(k)}}/h^{z-1})),\, ... \big) \tag{69}$$

$$= \boldsymbol{\beta}^\top_{\text{NTK}} \left(\hat{\boldsymbol{x}}_0 \cdot D^z_{\boldsymbol{v}}\, \mathbb{I}^{(k)} - z\hat{\boldsymbol{v}} \cdot D^{z-1}_{\boldsymbol{v}}\, \mathbb{I}^{(k)},\, \boldsymbol{w}^{(k)\top} \hat{\boldsymbol{x}}_0 \cdot D^z_{\boldsymbol{v}}\, \mathbb{I}^{(k)} - z\boldsymbol{w}^{(k)\top} \hat{\boldsymbol{v}} \cdot D^{z-1}_{\boldsymbol{v}}\, \mathbb{I}^{(k)},\, ... \right) \tag{70}$$

where line (70) follows from the limit definition of the directional derivative of the indicator evaluated at $\hat{\boldsymbol{x}}_0$ which completes our proof of Lemma 1. □

Let us look closer at line (70). It is well known that the derivative of the indicator (Heaviside function) $\mathbb{I}$ - or any such step function for this matter - does not classically have a well-defined derivative. This fact makes the analysis beyond equation (70) difficult because we are interested in the derivative of the indicator evaluated at $\boldsymbol{x}_0 = \mathbf{0}$, which is precisely where the discontinuity exists.

Fortunately, we have a workaround. By generalizing the notion of the indicator's derivative, we can consider the distributional derivative of the indicator, which is the Dirac-delta function (impulse spike located at $\boldsymbol{x}_0 = \mathbf{0}$). This is a similar workaround to how we pseudo-inverted the otherwise singular constant matrix $\boldsymbol{J}$ from equation (27) by generalizing the notion of the matrix inverse. Using chain rule, the directional derivative of $\mathbb{I}$ evaluted at $\boldsymbol{x}_0 = \mathbf{0}$ is:

$$D^z_{\boldsymbol{v}}\, \mathbb{I} \left(\boldsymbol{w}^\top \hat{\boldsymbol{x}}_0 \geq 0 \right) = \langle \tilde{\boldsymbol{w}}, \boldsymbol{v} \rangle^z \cdot \delta^{(z-1)}(\boldsymbol{w}_{d+1}). \tag{71}$$

Equation (71) gives us a cleaner expression for the z-th derivative of f_{NTK} by the sifting property of the Dirac-delta. Continuing from equation (70), we use line (71) to get:

$$= \int (\boldsymbol{\beta}^1_{\boldsymbol{w}})_{d+1} \langle \check{\boldsymbol{w}}, \boldsymbol{v} \rangle^z \cdot \delta^{(z-1)}(\boldsymbol{w}_{d+1}) d\mathbb{P}(\boldsymbol{w}) - \int z \langle \boldsymbol{\beta}^1_{\boldsymbol{w}} \hat{\boldsymbol{v}} \rangle \langle \check{\boldsymbol{w}}, \boldsymbol{v} \rangle^{z-1} \cdot \delta^{(z-2)}(\boldsymbol{w}_{d+1}) d\mathbb{P}(\boldsymbol{w}) \tag{72}$$

$$+ \int \boldsymbol{\beta}^2_{\boldsymbol{w}} \boldsymbol{w}_{d+1} \langle \check{\boldsymbol{w}}, \boldsymbol{v} \rangle^z \cdot \delta^{(z-1)}(\boldsymbol{w}_{d+1}) d\mathbb{P}(\boldsymbol{w}) - \int z \boldsymbol{\beta}^2_{\boldsymbol{w}} \langle \check{\boldsymbol{w}}, \boldsymbol{v} \rangle^z \cdot \delta^{(z-2)}(\boldsymbol{w}_{d+1}) d\mathbb{P}(\boldsymbol{w}) \tag{73}$$

$$= (-1)^{z-1} \langle \check{\boldsymbol{w}}, \boldsymbol{v} \rangle^z \left[\frac{\partial^{z-1}}{\partial \boldsymbol{w}^{z-1}_{d+1}} (\boldsymbol{\beta}^1_{\boldsymbol{w}})_{d+1} \right]_{\boldsymbol{w}_{d+1}=0} + (-1)^{z-1} z \langle \check{\boldsymbol{w}}, \boldsymbol{v} \rangle^{z-1} \left[\frac{\partial^{z-2}}{\partial \boldsymbol{w}^{z-2}_{d+1}} \langle \boldsymbol{\beta}^1_{\boldsymbol{w}} \hat{\boldsymbol{v}} \rangle \right]_{\boldsymbol{w}_{d+1}=0} \tag{74}$$

$$+ (-1)^{z-1} \langle \check{\boldsymbol{w}}, \boldsymbol{v} \rangle^z \left[\frac{\partial^{z-1}}{\partial \boldsymbol{w}^{z-1}_{d+1}} \boldsymbol{\beta}^2_{\boldsymbol{w}} \boldsymbol{w}_{d+1} \right]_{\boldsymbol{w}_{d+1}=0} + (-1)^{z-1} z \langle \check{\boldsymbol{w}}, \boldsymbol{v} \rangle^z \left[\frac{\partial^{z-2}}{\partial \boldsymbol{w}^{z-2}_{d+1}} \boldsymbol{\beta}^2_{\boldsymbol{w}} \right]_{\boldsymbol{w}_{d+1}=0} \tag{75}$$

which rewrites the derivative of f_{NTK} in terms of the high derivatives of the beta components. The check notation $\check{\boldsymbol{w}}$ emphasizes that the direction vector no longer depends on the bias component, $\boldsymbol{w}_{d+1}$. Actually, the equivalence between $\check{\boldsymbol{w}}^\top \boldsymbol{v}$ and $\boldsymbol{w}^\top \hat{\boldsymbol{v}}$ is a useful consideration.

At this point, the emergence of the Dirac impulse suggests that nonlinearity may be preserved in high orders. Nonlinearity, if preserved, depends closely on the relationship between directions $\boldsymbol{w}$ and $\hat{\boldsymbol{v}}$ as well as the the derivative of the $\boldsymbol{\beta}_{NTK}$ components with respect to $\boldsymbol{w}_{d+1}$. However, the precise degree of nonlinearity remains unknown because we have not yet accounted for the influence of φ^∞. The upcoming section demonstrates that solving for the derivative of the $\boldsymbol{\beta}_{NTK}$ components simultaneously accounts for the positions of φ^∞.

A.3 Proof of Lemma 2

Following lines (74)-(75), we must solve for the partial derivatives of the beta components. We will first consider the terms with a dependence on $\boldsymbol{w}$; Recall the forms of the beta components from section 3 equations (41)-(42):

$$\boldsymbol{\beta}^1_{\boldsymbol{w}} = \mathbb{I}\left(\boldsymbol{w}^\top(-\hat{\boldsymbol{v}}) \geq 0\right) \cdot \left(C(t,\delta,\kappa) \sum_{j=1}^n g(\hat{\boldsymbol{x}}^\infty_i) \sum_{i=1}^n \hat{\boldsymbol{x}}^\infty_i + \frac{1}{\delta} \sum_{i=1}^n \hat{\boldsymbol{x}}^\infty_i g(\hat{\boldsymbol{x}}^\infty_i) \right)$$

$$\boldsymbol{\beta}^2_{\boldsymbol{w}} = \mathbb{I}\left(\boldsymbol{w}^\top(-\hat{\boldsymbol{v}}) \geq 0\right) \cdot \left(C(t,\delta,\kappa) \sum_{j=1}^n g(\hat{\boldsymbol{x}}^\infty_i) \sum_{i=1}^n \boldsymbol{w}^\top \hat{\boldsymbol{x}}^\infty_i + \frac{1}{\delta} \sum_{i=1}^n \boldsymbol{w}^\top \hat{\boldsymbol{x}}^\infty_i g(\hat{\boldsymbol{x}}^\infty_i) \right).$$

Firstly, the partial derivative of the indicator is a trivial analysis. We recall that the distributional derivative of the indicator is the Dirac-delta function. And, since the bias component of a direction vector $\hat{\boldsymbol{v}}_{d+1}$ is 0, it is not difficult to see that the z-th partial derivative of the indicator is 0 with respect to $\boldsymbol{w}_{d+1}$ for all $z \geq 1$.

$$\frac{\partial}{\partial \boldsymbol{w}_{d+1}} \mathbb{I}\left(\boldsymbol{w}^\top(-\hat{\boldsymbol{v}}) \geq 0\right) = \delta(\boldsymbol{w}^\top(-\hat{\boldsymbol{v}})) \cdot (-\hat{\boldsymbol{v}}_{d+1}) \tag{76}$$

Secondly, the partial derivative of the dot product $\boldsymbol{w}^\top \hat{\boldsymbol{x}}^\infty_i$ is also trivial to solve. For clarity, we apply the definition of φ^∞ before computing the partial derivative. Since the bias components of a data point and direction vector are 1 and 0 respectively, it is clear to see that the partial derivative equals 1.

$$\frac{\partial}{\partial \boldsymbol{w}_{d+1}} \boldsymbol{w}^\top \hat{\boldsymbol{x}}^\infty_i \tag{77}$$

$$= \frac{\partial}{\partial \boldsymbol{w}_{d+1}} (\boldsymbol{w}^\top \hat{\boldsymbol{x}}_i - t(\boldsymbol{w}^\top \hat{\boldsymbol{v}})) \tag{78}$$

$$= (\hat{\boldsymbol{x}}_i)_{d+1} - t\hat{\boldsymbol{v}}_{d+1} \tag{79}$$

Lastly, we want to discover the partial derivative of the constant $C(t, \delta, \kappa)$. Recalling its definition, κ is the only term in $C(t, \delta, \kappa)$ that depends on $\boldsymbol{w}$. We find that the z-th derivative of κ is 0 for any $z \geq 1$:

$$\frac{\partial^z \kappa}{\partial \boldsymbol{w}_{d+1}^z} \tag{80}$$

$$= \frac{\partial^z}{\partial \boldsymbol{w}_{d+1}^z} \int (\hat{\boldsymbol{v}}^2 + (\boldsymbol{w}^\top \hat{\boldsymbol{v}})^2) \cdot \mathbb{I}\left(\boldsymbol{w}^\top(-\hat{\boldsymbol{v}}) \geq 0\right) d\mathbb{P}(\boldsymbol{w}) \tag{81}$$

$$= \int \frac{\partial^z}{\partial \boldsymbol{w}_{d+1}^z} (\hat{\boldsymbol{v}}^2 + (\boldsymbol{w}^\top \hat{\boldsymbol{v}})^2) \cdot \mathbb{I}\left(\boldsymbol{w}^\top(-\hat{\boldsymbol{v}}) \geq 0\right) d\mathbb{P}(\boldsymbol{w}) \tag{82}$$

$$= \int \frac{\partial^{z-1}}{\partial \boldsymbol{w}_{d+1}^{z-1}} \left(2\left(\boldsymbol{w}^\top \hat{\boldsymbol{v}}\right) \mathbb{I}\left(\boldsymbol{w}^\top(-\hat{\boldsymbol{v}}) \geq 0\right) \hat{\boldsymbol{v}}_{d+1} - \left(\hat{\boldsymbol{v}}^2 + (\boldsymbol{w}^\top \hat{\boldsymbol{v}})^2\right) \delta\left(\boldsymbol{w}^\top(-\hat{\boldsymbol{v}})\right) \hat{\boldsymbol{v}}_{d+1}\right) d\mathbb{P}(\boldsymbol{w}), \tag{83}$$

where the third equality is 0 from the fact that $\hat{\boldsymbol{v}}_{d+1}$ is 0. We are now prepared to differentiate the beta components.

Lemma 2. *The components of the NTK representation coefficient $\boldsymbol{\beta}_{NTK}$ induced by a training input set $\varphi^\infty = \{\boldsymbol{x}_i^\infty\}_{i=1}^n$ where $\boldsymbol{x}_i^\infty = \boldsymbol{x}_i - t\boldsymbol{v}_\varphi$ for some $\boldsymbol{x}_i \in \mathcal{X}$ and any direction $\boldsymbol{v}_\varphi$ are constant with respect to the bias component of any given feature direction $\boldsymbol{w}_{d+1}$ such that:*

$$\frac{\partial^z \boldsymbol{\beta}_{\boldsymbol{w}}^1}{\partial \boldsymbol{w}_{d+1}^z}, \frac{\partial^z \boldsymbol{\beta}_{\boldsymbol{w}}^2}{\partial \boldsymbol{w}_{d+1}^z} = 0 \text{ for all } z \geq 1.$$

Proof. Differentiating the first beta component is relatively straightforward. By product rule, we analyze the derivative of the indication on the LHS and the sum on the RHS. We already know that the derivative of the indicator for a training point induced by φ^∞ is 0. We also know that the derivative of kappa is 0. And, since no other terms depend on $\boldsymbol{w}$ the z-th derivative of the first beta component with respect to $\boldsymbol{w}_{d+1}$ is simply 0 for all $z \geq 1$:

$$\frac{\partial \boldsymbol{\beta}_{\boldsymbol{w}}^1}{\partial \boldsymbol{w}_{d+1}} = \left(\sum_{j=1}^n g(\hat{\boldsymbol{x}}_j^\infty) \sum_{i=1}^n \hat{\boldsymbol{x}}_i^\infty \frac{\partial C(t, \delta, \kappa)}{\partial \boldsymbol{w}_{d+1}}\right) \cdot \mathbb{I}\big(\boldsymbol{w}^\top(-\hat{\boldsymbol{v}}) \geq 0\big). \tag{84}$$

where

$$\frac{\partial C(t, \delta, \kappa)}{\partial \boldsymbol{w}_{d+1}} = (t^2\kappa)(n\delta t^2 \frac{\partial \kappa}{\partial \boldsymbol{w}_{d+1}}) + (t^2 \frac{\partial \kappa}{\partial \boldsymbol{w}_{d+1}})(\delta(n\kappa t^2 + \delta)). \tag{85}$$

Similarly, we differentiate the second beta component by product rule. We observe the summation on the RHS where the dependence on $\boldsymbol{w}$ is more elaborate. Using equation (79) we can see that the derivative of the dot product in the second term of the summation reduces to 1. Then, for the first term, we once again leverage equation (79) and the fact that the derivative of kappa is 0 to discover by a straightforward algebraic manipulation that the z-th derivative of the second beta component approaches 0 for all $z \geq 1$:

$$\frac{\partial \boldsymbol{\beta}_{\boldsymbol{w}}^2}{\partial \boldsymbol{w}_{d+1}} \tag{86}$$

$$= \left(C(t, \delta, \kappa) \sum_{j=1}^n g(\hat{\boldsymbol{x}}_i^\infty)\, n + \frac{1}{\delta} \sum_{i=1}^n g(\hat{\boldsymbol{x}}_i^\infty)\right) \cdot \mathbb{I}\left(\boldsymbol{w}^\top(-\hat{\boldsymbol{v}}) \geq 0\right) \tag{87}$$

$$= \left(C(t, \delta, \kappa) \sum_{j=1}^n g(\hat{\boldsymbol{x}}_i^\infty)\, n + \frac{1}{\delta} \sum_{i=1}^n g(\hat{\boldsymbol{x}}_i^\infty)\right) \mathbb{I}\left(\boldsymbol{w}^\top(-\hat{\boldsymbol{v}}) \geq 0\right) \tag{88}$$

$$= \left(-\frac{n\kappa g_{sum} t^2}{\delta(n\kappa t^2 + \delta)} + \frac{1}{\delta} \sum_{i=1}^n g(\hat{\boldsymbol{x}}_i^\infty)\right) \mathbb{I}\left(\boldsymbol{w}^\top(-\hat{\boldsymbol{v}}) \geq 0\right) \tag{89}$$

$$= \left(\frac{g_{sum}}{\delta} - \frac{n\kappa g_{sum} t^2}{\delta(n\kappa t^2 + \delta)}\right) \mathbb{I}\left(\boldsymbol{w}^\top(-\hat{\boldsymbol{v}}) \geq 0\right) \tag{90}$$

$$= \frac{g_{sum}\delta}{\delta(n\kappa t^2 + \delta)} \mathbb{I}\left(\boldsymbol{w}^\top(-\hat{\boldsymbol{v}}) \geq 0\right) \tag{91}$$

$$= \frac{g_{sum}}{n\kappa t^2 + \delta} \mathbb{I}\left(\boldsymbol{w}^\top(-\hat{\boldsymbol{v}}) \geq 0\right). \tag{92}$$

Inspecting the final equality, it is not difficult to see that the first derivative of the second beta component approaches 0 as $\delta \to 0^+$ and $t \to \infty$. Furthermore, since the derivatives of the indicator and kappa are both 0, it is clear to see that by chain rule, the second derivative of the second beta component is also 0. Therefore, the z-th derivative of the second beta component is 0 for all $z \geq 1$. This completes our proof of Lemma 2. □

A.4 Proof of Theorem 1

Theorem 1. *An over-parameterized two-layer ReLU MLP $f_{NTK} : \mathbb{R}^d \to \mathbb{R}$ that is trained on a labeled set $\{(\boldsymbol{x}_i^\infty, y_i^\infty)\}_{i=1}^n$ with $\boldsymbol{x}_i^\infty = \boldsymbol{x}_i - t\boldsymbol{v}_\varphi$ for $\boldsymbol{x}_i \in \mathcal{X}$ and any direction $\boldsymbol{v}_\varphi$ in the NTK regime minimizing squared loss will converge to a quadratic extrapolator when evaluated at a point near the origin $\mathbf{0}$ as $t \to \infty$.*

Proof. Under the definition of φ^∞, Lemma 2 states that $\frac{\partial^z \boldsymbol{\beta}_{\boldsymbol{w}}^1}{\partial \boldsymbol{w}_{d+1}^z}$ and $\frac{\partial^z \boldsymbol{\beta}_{\boldsymbol{w}}^2}{\partial \boldsymbol{w}_{d+1}^z}$ are 0 for orders $z \geq 1$. But since Lemma 1 shows that $D_{\boldsymbol{v_0}}^z f_{NTK}$ for any direction $\boldsymbol{v_0}$ actually depends on the lower ordered $(z-1)$-th and $(z-2)$-th derivatives $\frac{\partial^{z-1} \boldsymbol{\beta}_{\boldsymbol{w}}^1}{\partial \boldsymbol{w}_{d+1}^{z-1}}$, $\frac{\partial^{z-2} \boldsymbol{\beta}_{\boldsymbol{w}}^1}{\partial \boldsymbol{w}_{d+1}^{z-2}}$, $\frac{\partial^{z-1} \boldsymbol{\beta}_{\boldsymbol{w}}^2}{\partial \boldsymbol{w}_{d+1}^{z-1}}$, and $\frac{\partial^{z-2} \boldsymbol{\beta}_{\boldsymbol{w}}^2}{\partial \boldsymbol{w}_{d+1}^{z-2}}$, it is not difficult to see that the third and all higher order derivatives are automatically 0. Then, taking $z = 1$ we simplify equation (74) to get an examinable form of the first derivative:

$$
\begin{aligned}
&D_{\boldsymbol{v_0}} f_{NTK}(\hat{\mathbf{0}}) \\
&= \langle \check{\boldsymbol{w}}, \boldsymbol{v} \rangle \left[(\boldsymbol{\beta}_{\boldsymbol{w}}^1)_{d+1} \right]_{\boldsymbol{w}_{d+1}=0} - \int \boldsymbol{\beta}_{\boldsymbol{w}}^{1\top} \hat{\boldsymbol{v}} \cdot \mathbb{I}\left(\boldsymbol{w}_{d+1} \geq 0\right) d\mathbb{P}(\boldsymbol{w}) - \int \boldsymbol{\beta}_{\boldsymbol{w}}^2 \boldsymbol{w}^\top \hat{\boldsymbol{v}} \cdot \mathbb{I}\left(\boldsymbol{w}_{d+1} \geq 0\right) d\mathbb{P}(\boldsymbol{w}).
\end{aligned}
$$

But more interestingly, we take $z = 2$ and simplify equation (75) for the second derivative:

$$
\begin{aligned}
&D_{\boldsymbol{v_0}}^2 f_{NTK}(\hat{\mathbf{0}}) \\
&= -\langle \check{\boldsymbol{w}}, \boldsymbol{v} \rangle^2 \left[\frac{\partial}{\partial \boldsymbol{w}_{d+1}} (\boldsymbol{\beta}_{\boldsymbol{w}}^1)_{d+1} \right]_{\boldsymbol{w}_{d+1}=0} - 2\langle \check{\boldsymbol{w}}, \boldsymbol{v} \rangle \left[\langle \boldsymbol{\beta}_{\boldsymbol{w}}^1 \hat{\boldsymbol{v}} \rangle \right]_{\boldsymbol{w}_{d+1}=0} \\
&- \langle \check{\boldsymbol{w}}, \boldsymbol{v} \rangle^2 \left[\frac{\partial}{\partial \boldsymbol{w}_{d+1}} \boldsymbol{\beta}_{\boldsymbol{w}}^2 \boldsymbol{w}_{d+1} \right]_{\boldsymbol{w}_{d+1}=0} - 2\langle \check{\boldsymbol{w}}, \boldsymbol{v} \rangle^2 \left[\boldsymbol{\beta}_{\boldsymbol{w}}^2 \right]_{\boldsymbol{w}_{d+1}=0} \\
&= -2\langle \check{\boldsymbol{w}}, \boldsymbol{v} \rangle \left[\langle \boldsymbol{\beta}_w^1 \hat{\boldsymbol{v}} \rangle \right]_{\boldsymbol{w}_{d+1}=0} - \langle \check{\boldsymbol{w}}, \boldsymbol{v} \rangle^2 \left[\boldsymbol{\beta}_{\boldsymbol{w}}^2 \right]_{\boldsymbol{w}_{d+1}=0} - 2\langle \check{\boldsymbol{w}}, \boldsymbol{v} \rangle^2 \left[\boldsymbol{\beta}_{\boldsymbol{w}}^2 \right]_{\boldsymbol{w}_{d+1}=0} \\
&= -2\langle \check{\boldsymbol{w}}, \boldsymbol{v} \rangle \left[\langle \boldsymbol{\beta}_{\boldsymbol{w}}^1 \hat{\boldsymbol{v}} \rangle \right]_{\boldsymbol{w}_{d+1}=0} - 3\langle \check{\boldsymbol{w}}, \boldsymbol{v} \rangle^2 \left[\boldsymbol{\beta}_{\boldsymbol{w}}^2 \right]_{\boldsymbol{w}_{d+1}=0},
\end{aligned}
$$

to see a great dependence in the final equality on the beta components and dot product between any particular $\boldsymbol{w}$ and direction of evaluation $\boldsymbol{v_0}$. Thus, for the special case of a training input set φ^∞ whose members are located far from the origin, the regressor becomes a quadratic extrapolator when evaluated near the origin. □