ON THE STABILITY OF THE NEURAL TANGENT KERNEL IN CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In wide neural networks, the Neural Tangent Kernel (NTK) remains approximately constant during training, providing a powerful theoretical tool for studying training dynamics, generalization, and connections to kernel methods. However, this theory is largely restricted to regression losses. It was previously thought that training on a classification loss, or more generally losses involving nonlinear output transformations, breaks this property, leading to divergent logits and a breakdown of the linearization. In this paper, we extend NTK theory to classification by identifying conditions under which wide neural networks remain in the lazy training regime. We show that parameter-space regularization ensures a constant NTK during training for cross-entropy loss, while in the absence of regularization the regime is recovered when targets are non-degenerate, i.e. when all classes have strictly positive probability. Under these conditions, training is well-approximated by the linearized model, yielding an explicit characterization of the solution in terms of the NTK. We further analyze the distribution of trained predictors induced by random initialization and relate this notion of model uncertainty to Bayesian methods.

1 INTRODUCTION

The theoretical understanding of deep learning has significantly advanced through the study of wide neural networks via the neural tangent kernel (NTK) (Jacot et al., 2018). In the infinite-width limit, the NTK remains approximately constant during training, and the network is well-approximated by its parameter linearization (Lee et al., 2019). This perspective enables a precise analysis of training dynamics and generalization (Jacot et al., 2018; Arora et al., 2019; Du et al., 2019), and establishes a close connection between neural networks, kernel methods (Cao & Gu, 2019), and Gaussian processes (Lee et al., 2019; Calvo-Ordoñez et al., 2026b). In particular, for mean-squared error (MSE) loss, the trained network exactly fits the training data and admits a tractable Gaussian distribution on test inputs, determined by the random initialization (He et al., 2020). While this framework has been extensively developed for regression, the application beyond this setting is often limited to finite training times.

For classification objectives such as cross-entropy, the logits that perfectly fit the training data diverge, causing the parameters to leave the lazy training regime (Yu et al., 2025). As a result, the NTK is no longer constant and the linearization breaks down, preventing a direct extension of NTK theory to classification. Calvo-Ordoñez et al. (2026b) showed for MSE loss that parameter-space regularization around initialization preserves the NTK regime and admits a Bayesian interpretation in terms of observation noise. Whether a similar mechanism can recover NTK behavior for classification losses remains unclear, since cross-entropy training is known to produce divergent logits and a breakdown of linearization. In this work, we show that parameter-space regularization is sufficient to recover lazy training for cross-entropy and more general function-space losses, yielding approximation by the linearized model under mild assumptions.

Liu et al. (2020) show that when a nonlinearity is placed in the last layer, the NTK is no longer constant during training, an example being a network with softmax outputs. Our key contribution here is to, instead, treat the softmax as part of the function-space loss, which allows us to recover the NTK regime at the level of the pre-activation outputs. In addition to the regularized setting, we consider the case without parameter-space regularization. We show that when the target probabilities

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

are non-degenerate, i.e. assign strictly positive probability to every class, the pre-activation outputs remain in the NTK regime and are well-approximated by their linearization.

Building on these theoretical results, we study the distribution over trained predictors induced by random initialization. This notion of model uncertainty has been previously analyzed in the NTK regime (He et al., 2020) for MSE with extra label noise, where it is shown to be more conservative than Bayesian last-layer methods, but less conservative than the NTK-GP posterior. For general loss functions, the training dynamics of the linearized network remain nonlinear. Nevertheless, we show that the stationary point admits an explicit characterization via an inverse map in function-space. This characterization enables efficient sampling from the trained network, which we use to approximate the distribution over predictors. Empirically, we find that this closely matches the ensemble distribution obtained by retraining the network from different initializations.

Our main contributions can be summarized as follows:

- We show that parameter-space regularization preserves the NTK regime for cross-entropy and more general losses, extending NTK theory beyond regression.
- We show that, even without regularization, the NTK regime is recovered for classification when target probabilities are non-degenerate, covering settings such as label smoothing.
- We characterize the distribution over trained predictors induced by random initialization, and relate this notion of model uncertainty to Bayesian methods.
- We provide empirical validation supporting the theoretical predictions.

2 Preliminaries

2.1 Wide neural networks in NTK parametrization

Consider a standard feedforward neural network f_θ with inputs x in a compact space $M^d \subseteq \mathbb{R}^d$, parameters $\theta \in \mathbb{R}^p$, and outputs $f_\theta(x) \in \mathbb{R}^K$. For smooth activations, $\theta \mapsto f_\theta(x)$ is twice continuously differentiable. For any points $\mathbf{x}_1, \ldots, \mathbf{x}_N \in M^d$ define $f_\theta(\mathbf{x}) \in \mathbb{R}^{NK}$ by element-wise application, and similarly the parameter-Jacobian $J_\theta(\mathbf{x}) := \nabla_\theta f_\theta(\mathbf{x})^\top \in \mathbb{R}^{NK \times p}$ and parameter-Hessian $H_\theta(\mathbf{x}) := \nabla^2_{\theta\theta} f_\theta(\mathbf{x}) \in \mathbb{R}^{NK \times p \times p}$. If the last layer is linear, Liu et al. (2020) show that the parameter-Jacobian is locally bounded and $\mathcal{O}\left((\log n)^c/\sqrt{n}\right)$-Lipschitz for large layer width n. We restate this in the following lemma, and provide more explanations in Appendix A.

Lemma 2.1. *For any $\delta_0 > 0$ there are $K_1', K_2' > 0$ such that: For every radius $R > 0$ there is large enough layer width n such that with probability $1 - \delta_0$ over random initialization θ_0: For any input $x \in M^d$:*

$$\forall \theta \in B(\theta_0, R) : \|J_\theta(x)\|_{2,2} \leq K_1', \tag{1}$$

$$\forall \theta \in B(\theta_0, R) : \|H_\theta(x)\|_{\mathrm{op},2} \leq \frac{(\log n)^c}{\sqrt{n}} K_2'. \tag{2}$$

Recall that $J_\theta(x) \in \mathbb{R}^{K \times p}$ and $H_\theta(x) \in \mathbb{R}^{K \times p \times p}$. We are using the Euclidean norm on the K-dimensional output dimension, after applying the Euclidean and spectral norm on the parameter dimensions p and $p \times p$ respectively. The constants $K_1', K_2' > 0$ do not depend on R. The constant $c > 0$ only depends on the network architecture.

Lee et al. (2018); Matthews et al. (2018) show that the network output at initialization converges in distribution for infinite width:

Lemma 2.2. *Consider random initialization θ_0. Then $f_{\theta_0}(\cdot)$ converges in distribution to a Gaussian process with zero mean and covariance given by the NNGP Kernel $\mathcal{K}$: For inputs $\mathbf{x}_1, \ldots, \mathbf{x}_N \in M^d$,*

$$f_{\theta_0}(\mathbf{x}) \xrightarrow{d} \mathcal{N}(0, \mathcal{K}(\mathbf{x}, \mathbf{x})). \tag{3}$$

Our results apply to any parametric model with the properties of these two lemmas, which have been proven for more general architectures in (Yang, 2020).

2.2 THE TRAINING LOSS IN FUNCTION-SPACE

Consider training points $\mathbf{x}_1, \dots, \mathbf{x}_N \in M^d$. Consider a function-space loss

$$\mathcal{C} : \mathbb{R}^{NK} \to \mathbb{R}, \quad \mathbf{z} \mapsto \mathcal{C}(\mathbf{z}), \tag{4}$$

which we will evaluate in the function outputs

$$\mathbf{z} := \mathbf{f} := f(\mathbf{x}) := (f(\mathbf{x}_1), \dots, f(\mathbf{x}_N)) \in \mathbb{R}^{NK}. \tag{5}$$

We assume $\mathcal{C}$ is twice continuously differentiable, and denote the function-space gradient by $\nabla_{\mathbf{z}}\mathcal{C}(\mathbf{z}) \in \mathbb{R}^{NK}$ and the function-space Hessian by $\nabla^2_{\mathbf{zz}}\mathcal{C}(\mathbf{z}) \in \mathbb{R}^{NK \times NK}$.

Consider a classification task with K classes. Denote the target probabilities for N training points by $\mathbf{p}_1, \dots, \mathbf{p}_N \in \Delta^{K-1}$. One-hot encodings correspond to $\mathbf{p}_i = e_{\mathbf{y}_i}$ for training labels $\mathbf{y}_i \in \{1, \dots, K\}$. The function-space categorical cross-entropy loss (Appendix C.2) is

$$\mathcal{C}_{\mathrm{CE}}(\mathbf{z}) := \sum_{i=1}^{N} \mathrm{CE}[\mathbf{p}_i, \mathrm{smx}(\mathbf{z}_i)] = -\sum_{i=1}^{N}\sum_{k=1}^{K} \mathbf{p}_{ik} \log \mathrm{smx}(\mathbf{z}_i)_k. \tag{6}$$

Here $\mathrm{smx} : \mathbb{R}^K \to \mathrm{int}(\Delta^{K-1})$ denotes the $\mathrm{softmax}$-function, whose outputs lie in the relative interior of the probability simplex. To avoid redundancies from $\mathrm{smx}(z) = \mathrm{smx}(z + \alpha 1)$ we can alternatively use $\mathrm{softmax}$ with a fixed reference class (Appendix C.3).

2.3 REGULARIZED GRADIENT FLOW

Consider a regularization strength $\beta \geq 0$. Define the regularized parameter-space loss

$$\mathcal{L}^\beta : \mathbb{R}^p \to \mathbb{R}, \quad \mathcal{L}^\beta(\theta) := \mathcal{C}(f_\theta(\mathbf{x})) + \frac{1}{2}\beta \left\|\theta - \theta_0\right\|_2^2. \tag{7}$$

If $\beta = 0$ we assume $\mathcal{C}$ is lower bounded, i.e. $\inf \mathcal{C} > -\infty$. Starting from random initial parameters $\theta_0 \in \mathbb{R}^p$ this objective is minimized using the gradient flow[1]

$$\frac{d}{dt}\theta_t = -\eta_0 \nabla_\theta \mathcal{L}^\beta(\theta_t) = -\eta_0 \left(\nabla_\theta \mathcal{C}(f_{\theta_t}(\mathbf{x})) + \beta(\theta_t - \theta_0)\right). \tag{8}$$

By the chain rule,

$$\nabla_\theta \mathcal{C}(f_\theta(\mathbf{x})) = J_\theta(\mathbf{x})^\top \nabla_{\mathbf{z}} \mathcal{C}(f_\theta(\mathbf{x})) \in \mathbb{R}^p. \tag{9}$$

Lemma 2.3. *Let $\delta_0 > 0$ be arbitrarily small. There is $K_0 > 0$, such that for large enough layer width n, with probability $1 - \delta_0$ over random initialization θ_0:*

$$\mathcal{C}(f_{\theta_t}(\mathbf{x})) \leq \mathcal{L}^\beta(\theta_t) \leq \mathcal{L}^\beta(\theta_0) = \mathcal{C}(f_{\theta_0}(\mathbf{x})) \leq K_0 \quad \textit{for all } t \geq 0. \tag{10}$$

The proof is in Appendix B. Given K_0, we define the function-space loss sublevel set

$$\mathcal{S}_0 := \{\mathbf{z} \in \mathbb{R}^{NK} : \mathcal{C}(\mathbf{z}) \leq K_0\}. \tag{11}$$

2.4 PARAMETER-LINEARIZATION OF A NEURAL NETWORK

The parameter-linearization of a neural network around its initialization $\theta_0 \in \mathbb{R}^p$ is

$$f_\theta^{\mathrm{lin}}(x) := f_{\theta_0}(x) + J_{\theta_0}(x)(\theta - \theta_0). \tag{12}$$

This model remains nonlinear in the input x, but is linear in the parameters θ. Thus, its parameter-Jacobian $J_\theta^{\mathrm{lin}}(x)$ is constant, i.e. $J_\theta^{\mathrm{lin}}(x) = J_{\theta_0}(x)$. Denote by θ_t^{lin} the parameters for gradient flow with the linearized network. They follow the dynamics

$$\frac{d}{dt}\theta_t^{\mathrm{lin}} = -\eta_0 \left(J_{\theta_0}(\mathbf{x})^\top \nabla_{\mathbf{z}} \mathcal{C}\left(f^{\mathrm{lin}}_{\theta_t^{\mathrm{lin}}}(\mathbf{x})\right) + \beta(\theta_t^{\mathrm{lin}} - \theta_0)\right), \quad \theta_0^{\mathrm{lin}} = \theta_0. \tag{13}$$

[1] Picard-Lindelöf gives the existence of a unique solution on $[0, T_{\max})$ for some $T_{\max} > 0$ with $\lim_{t \uparrow T_{\max}} \|\theta_t\|_2 = \infty$ if $T_{\max} < \infty$. We will prove $T_{\max} = \infty$. With $t \geq 0$ we refer to $t \in [0, T_{\max})$.

For any input $x \in M^d$, multiplying both sides with $J_{\theta_0}(x)$ gives the function-space ODE of the linearized network

$$\frac{d}{dt} f^{\text{lin}}_{\theta^{\text{lin}}_t}(x) = -\eta_0 \left(J_{\theta_0}(x) J_{\theta_0}(\mathbf{x})^\top \nabla_{\mathbf{z}} \mathcal{C} \left(f^{\text{lin}}_{\theta^{\text{lin}}_t}(\mathbf{x}) \right) + \beta \left(f^{\text{lin}}_{\theta^{\text{lin}}_t}(x) - f_{\theta_0}(x) \right) \right). \tag{14}$$

We see that the dynamics of the linearized network only depend on the parameters θ^{lin}_t through its function values $\hat{g}_t := f^{\text{lin}}_{\theta^{\text{lin}}_t}$. They are driven by the empirical NTK at initialization $\hat{\Theta}_{\theta_0}(x, \mathbf{x}) = J_{\theta_0}(x) J_{\theta_0}(\mathbf{x})^\top \in \mathbb{R}^{K \times NK}$. We define the function-space values g driven by the analytical NTK Θ through the ODEs

$$\frac{d}{dt} g_t(x) = -\eta_0 \left(\Theta_{x,\mathbf{x}} \nabla_{\mathbf{z}} \mathcal{C} \left(g_t(\mathbf{x}) \right) + \beta \left(g_t(x) - g_0(x) \right) \right), \quad g_0(x) = f_{\theta_0}(x). \tag{15}$$

3 Theoretical results for general loss functions

3.1 Approximation by parameter-linearization

In this section, we establish conditions under which wide neural networks can be approximated by their parameter linearization along the training path. The approximation error vanishes as the width increases, at a rate of order $O((\log n)^c/\sqrt{n})$. We will prove this as a consequence of lazy training.

Theorem 3.1. *Let $\delta_0 > 0$ be arbitrarily small. For $\beta > 0$ assume that $\mathcal{C}$ is convex, and that there is $K_1 > 0$ such that*

$$\forall \mathbf{z} \in \mathcal{S}_0 : \quad \|\nabla_{\mathbf{z}} \mathcal{C}(\mathbf{z})\|_2 \leq K_1. \tag{16}$$

For $\beta = 0$ assume that the Hessian of $\mathcal{C}$ is strictly positive definite and that there are $K_2, \mu_{\mathcal{C}} > 0$, such that

$$\forall \mathbf{z} \in \mathcal{S}_0 : \quad 2\mu_{\mathcal{C}} \left(\mathcal{C}(\mathbf{z}) - \inf \mathcal{C} \right) \leq \|\nabla_{\mathbf{z}} \mathcal{C}(\mathbf{z})\|_2^2 \leq 2K_2 \left(\mathcal{C}(\mathbf{z}) - \inf \mathcal{C} \right). \tag{17}$$

Then we can uniformly approximate the network by its linearization during training. This means that there is $C_2 > 0$ such that for large enough layer width n, with probability $1 - \delta_0$ over random initialization θ_0,

$$\forall x \in M^d, t \geq 0 : \quad \left\| f_{\theta_t}(x) - f^{\text{lin}}_{\theta^{\text{lin}}_t}(x) \right\|_2 \leq C_2 \frac{(\log n)^c}{\sqrt{n}}. \tag{18}$$

We discuss the assumptions in Appendix B, and prove them for various function-space losses in Appendix C.

The proof for $\beta > 0$ is in Appendix D. The assumptions are satisfied by the categorical cross-entropy loss $\mathcal{C}_{\text{CE}}$ (with and without a reference class, see Appendix C.2 and C.3) for general target distributions $\mathbf{p}_i$, including one-hot targets. This contrasts with the unregularized case, where divergence of the logits prevents the network from remaining in the NTK regime (Yu et al., 2025). We will see in the experimental section how even a small parameter regularizer ensures the NTK regime.

The proof for $\beta = 0$ is in Appendix E. When target distributions $\mathbf{p}_i$ have full support, the assumptions of the theorem are satisfied by the categorical cross-entropy loss $\mathcal{C}_{\text{CEr}}$ with a reference class (see Appendix C.3). This corresponds to label smoothing (Szegedy et al., 2016; Müller et al., 2019).

The (standard) categorical cross-entropy loss $\mathcal{C}_{\text{CE}}$ without a reference class is convex, but its Hessian is nowhere strictly positive definite as it is invariant with respect to shifts. In Appendix E.3 we extend Theorem 3.1 for $\beta = 0$ to show that still (when target distributions have full support), the centered version of the network[2] is well approximated by the centered version of its linearization during training. This means that

$$\forall x \in M^d, t \geq 0 : \quad \left\| P \left(f_{\theta_t}(x) - f^{\text{lin}}_{\theta^{\text{lin}}_t}(x) \right) \right\|_2 \leq C_2 \frac{(\log n)^c}{\sqrt{n}}. \tag{19}$$

Due to $\text{smx}(z) = \text{smx}(Pz)$, this implies that

$$\forall x \in M^d, t \geq 0 : \quad \left\| \text{smx}\left(f_{\theta_t}(x) \right) - \text{smx}\left(f^{\text{lin}}_{\theta^{\text{lin}}_t}(x) \right) \right\|_2 \leq \frac{1}{2} C_2 \frac{(\log n)^c}{\sqrt{n}}. \tag{20}$$

[2]Let $P := I_K - \frac{1}{K} 1_K 1_K^\top \in \mathbb{R}^{K \times K}$ denote the orthogonal projection onto $1_K^\perp$. The centered network output is obtained by applying P to the logits, i.e. $P f_\theta(x)$.

The closeness to the linearization is particularly novel in this setting. Liu et al. (2020) studied networks with a nonlinear last layer,

$$\tilde{f}_\theta(x) := \mathrm{smx}(f_\theta(x)), \tag{21}$$

and showed that the NTK of $\tilde{f}$ is no longer constant during training, preventing approximation by direct linearization of $\tilde{f}$. In contrast, we treat the last nonlinearity as part of the loss, while keeping the network architecture with a linear last layer. Equation 20 shows that $\tilde{f}$ can be approximated by applying the last nonlinearity to the linearized pre-activation.

3.2 The function-space training dynamics of the linearized network

In the previous section we characterized the conditions under which wide neural networks are well approximated by their linearized version during training. We now study the gradient flow dynamics of the linearized network, whose stationary point provides an explicit characterization of the trained predictor in function-space.

Consider the kernel regression problem

$$\min_{g \in f_{\theta_0} + \mathcal{H}_\Theta} \mathcal{C}(g(\mathbf{x})) + \frac{1}{2}\beta \left\| g - f_{\theta_0} \right\|^2_{\Theta^{-1}} \quad \text{or} \quad \min_{\mathbf{z} \in \mathbb{R}^{NK}} \mathcal{C}(\mathbf{z}) + \frac{1}{2}\beta (\mathbf{z} - \mathbf{f}_{\theta_0})^\top \Theta^{-1}_{\mathbf{x},\mathbf{x}} (\mathbf{z} - \mathbf{f}_{\theta_0}), \tag{22}$$

via the kernel representation $g(x) = f_{\theta_0}(x) + \Theta_{x,\mathbf{x}}\Theta^{-1}_{\mathbf{x},\mathbf{x}}(\mathbf{z} - \mathbf{f}_{\theta_0})$. The optimal function is equal to the stationary point g_∞ of the function-space ODE (15). It satisfies

$$\Theta_{x,\mathbf{x}} \nabla_{\mathbf{z}} \mathcal{C}\left(g_\infty(\mathbf{x})\right) + \beta g_\infty(x) = \beta f_{\theta_0}(x). \tag{23}$$

Restricting to the training points $x = \mathbf{x}_1, \ldots, \mathbf{x}_N$, define the function-space operator

$$\Phi : \mathbb{R}^{NK} \to \mathbb{R}^{NK}, \quad \Phi(\mathbf{z}) := \Theta_{\mathbf{x},\mathbf{x}} \nabla_{\mathbf{z}} \mathcal{C}(\mathbf{z}) + \beta \mathbf{z}. \tag{24}$$

The stationary point equation (23) implies:

Lemma 3.2. *The predictions of the linearized network in the training points $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are*

$$g_\infty(\mathbf{x}) = \Phi^{-1}(\beta f_{\theta_0}(\mathbf{x})). \tag{25}$$

Using this, we can compute the trained values at any test point x:

Lemma 3.3. *Let $\beta \geq 0$ and assume $\Theta_{\mathbf{x},\mathbf{x}}$ is invertible. Then the prediction of the linearized network at any test point x is*

$$g_\infty(x) = f_{\theta_0}(x) + \Theta_{x,\mathbf{x}}\Theta^{-1}_{\mathbf{x},\mathbf{x}} \left(\underbrace{\Phi^{-1}(\beta f_{\theta_0}(\mathbf{x}))}_{= g_\infty(\mathbf{x})} - f_{\theta_0}(\mathbf{x}) \right). \tag{26}$$

Moreover for $\beta > 0$ the prediction can also be written as

$$g_\infty(x) = f_{\theta_0}(x) - \frac{1}{\beta}\Theta_{x,\mathbf{x}} \nabla_{\mathbf{z}} \mathcal{C}\left(\Phi^{-1}(\beta f_{\theta_0}(\mathbf{x}))\right). \tag{27}$$

The proof is in Appendix F.

For $\beta = 0$, the predictions in the training points are deterministic: $\mathbf{g}_\infty(\mathbf{x}) = \Phi^{-1}(0)$. Thus, $g_\infty(x)$ is normally distributed. For $\beta > 0$ we will show how to approximate it with a normal distribution.

These expressions allow us to sample from the trained neural network $f_{\theta_\infty}(x) \approx g_\infty(x)$ by sampling from the random initialization $f_{\theta_0}((x, \mathbf{x})) \sim \mathcal{N}(0, \mathcal{K}_{(x,\mathbf{x}),(x,\mathbf{x})})$. This provides the distribution of an infinite-width ensemble of predictors induced by initialization, see Figure 1. We analyze this notion of model uncertainty in the next section.

4 Connection of the infinite-width ensemble to Bayesian methods

The previous section characterized the trained linearized predictor through the inverse map Φ^{-1}. We now use this characterization to study the distribution over trained predictors induced by random initialization.

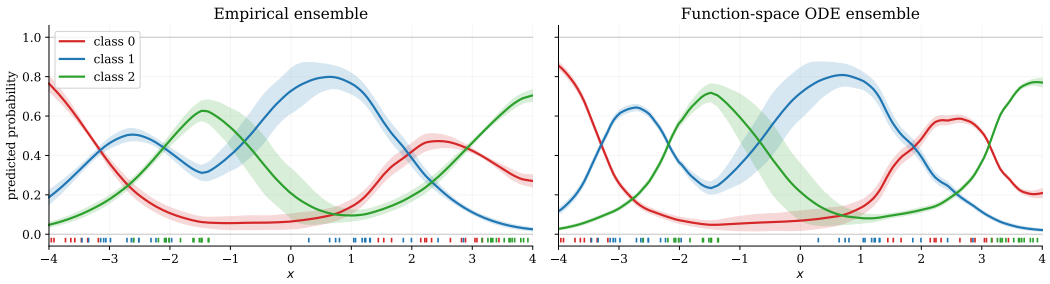

Figure 1: 1d-classification with 3 classes. Left: An ensemble over wide networks, starting from different parameter initializations. Right: The infinite-width limit of the ensemble, using the function-space ODE.

4.1 Connection to MAP inference for classification

Consider parameter regularizer $\beta > 0$ and target probabilities $\mathbf{p}_1, \ldots, \mathbf{p}_N \in \Delta^{K-1}$. In this section we focus on one-hot targets $\mathbf{p}_i = e_{\mathbf{y}_i}$ for labels $\mathbf{y}_i \in \{1, \ldots, K\}$.

For categorical cross-entropy loss $\mathcal{C}_{\mathrm{CE}}$, the function-space operator $\Phi : \mathbb{R}^{NK} \to \mathbb{R}^{NK}$ is given by

$$\Phi(\mathbf{z}) = \sum_{i=1}^{N} \Theta_{\mathbf{x},\mathbf{x}_i} \left(\mathrm{smx}(\mathbf{z}_i) - \mathbf{p}_i\right) + \beta \mathbf{z}. \tag{28}$$

For large layer width, Lemma 3.3 (or solving the function-space ODE (15)) allows us to sample from f_{θ_∞} by sampling from f_{θ_0}. The resulting random variable $\mathrm{smx}(f_{\theta_\infty}(x))$ is a sample probability vector from the simplex of discrete probability distributions on $\{1, \ldots, K\}$. This represents epistemic uncertainty for classification coming from initialization variance. While this differs from Bayesian uncertainty (see Section 4.2), we can still draw parallels by viewing each sample as a MAP estimate for a different prior mean. Consider the probabilistic likelihood model

$$y(x) = \underset{k=1,\ldots,K}{\arg\max} \left(F(x)_k + \tau \epsilon_k\right), \quad \epsilon_k \sim \mathrm{Gumbel}(0, 1). \tag{29}$$

This has conditional likelihood (Maddison et al., 2014)

$$\mathrm{p}\left(y(x) = k | F(x)\right) = \mathrm{smx}\left(\frac{1}{\tau} F(x)\right)_k \propto \exp\left(-\frac{1}{\tau} \mathrm{CE}(e_k, F(x))\right). \tag{30}$$

Moreover, place the GP prior $F(x) \sim \mathcal{N}(F_0(x), \Theta_{x,x})$. The MAP in this model minimizes

$$F \mapsto -\log \mathrm{p}(F | (\mathbf{x}_i, \mathbf{y}_i)_{i=1}^N) = -\sum_{i=1}^{N} \log \mathrm{smx}\left(\frac{1}{\tau} F(\mathbf{x}_i)\right)_{\mathbf{y}_i} + \frac{1}{2} \left\| F - F_0 \right\|_{\Theta^{-1}}^2 + \mathrm{const}\,. \tag{31}$$

Reparameterizing $F = \tau g$, $F_0 = \tau f_{\theta_0}$, $\beta = \tau^2$, recovers the objective

$$g \mapsto -\sum_{i=1}^{N} \log \mathrm{smx}\left(g(\mathbf{x}_i)\right)_{\mathbf{y}_i} + \frac{1}{2} \beta \left\| g - f_{\theta_0} \right\|_{\Theta^{-1}}^2 + \mathrm{const}\,. \tag{32}$$

This shows that in the infinite-width limit, the neural network solution converges to the temperature-rescaled MAP estimate in a probabilistic model whose observation noise is determined by β. The equivalence to the NTK-GP has been analyzed in Calvo-Ordoñez et al. (2026b) for MSE loss corresponding to regression tasks. Our results allow the use of their tools for classification NTK-GP inference under arbitrary prior means via a single neural network pass.

4.2 Connection of the ensemble variance to the Laplace approximation

The infinite-width ensemble distribution over $f_{\theta_\infty}(x)$ is intractable for general function-space losses $\mathcal{C}$. By linearizing $\nabla_{\mathbf{z}} \mathcal{C}(\mathbf{z})$ we can approximate it with a normal distribution: Let $\mathbf{g}^* := \Phi^{-1}(\mathbf{0})$ denote the MAP in the training points for prior mean $f_{\theta_0} = 0$, i.e. the solution of

$$\Theta_{\mathbf{x},\mathbf{x}} \nabla_{\mathbf{z}} \mathcal{C}(\mathbf{g}^*) + \beta \mathbf{g}^* = 0. \tag{33}$$

Linearize the function-space gradient around $\mathbf{g}^*$:

$$\nabla_{\mathbf{z}}\mathcal{C}(\mathbf{z}) \approx \nabla_{\mathbf{z}}C(\mathbf{g}^*) + \nabla^2_{\mathbf{z}}\mathcal{C}(\mathbf{g}^*)(\mathbf{z}-\mathbf{g}^*) = -\beta\Theta^{-1}_{\mathbf{x},\mathbf{x}}\mathbf{g}^* + \nabla^2_{\mathbf{z}}\mathcal{C}(\mathbf{g}^*)(\mathbf{z}-\mathbf{g}^*). \tag{34}$$

Define $\mathbf{H}^* = \nabla^2_{\mathbf{zz}}\mathcal{C}(\mathbf{g}^*) \in \mathbb{R}^{NK\times NK}$ and $\mathbf{A} := \Theta_{\mathbf{x},\mathbf{x}}\mathbf{H}^* + \beta I_{NK}$. While $\mathbf{A}$ is not symmetric, $\mathbf{H}^*\mathbf{A}^{-1}$ is symmetric. Using the linearization in equation (23) gives the approximate stationary point in the training points $\mathbf{x}$ as

$$g_\infty(\mathbf{x}) \approx \mathbf{g}^* + \mathbf{A}^{-1}\beta f_{\theta_0}(\mathbf{x}). \tag{35}$$

Equation (23) further implies for test points $\mathbf{x}'$ that

$$g_\infty(\mathbf{x}') \approx f_{\theta_0}(\mathbf{x}') + \Theta_{\mathbf{x}',\mathbf{x}}\Theta^{-1}_{\mathbf{x},\mathbf{x}}\mathbf{g}^* - \Theta_{\mathbf{x}',\mathbf{x}}\mathbf{H}^*\mathbf{A}^{-1}f_{\theta_0}(\mathbf{x}). \tag{36}$$

This is linear in the initial function values $f_{\theta_0}((\mathbf{x}',\mathbf{x})) \sim \mathcal{N}(0, \mathcal{K}_{(\mathbf{x}',\mathbf{x}),(\mathbf{x}',\mathbf{x})})$, and allows us to approximate the infinite-width ensemble distribution by a Gaussian with mean and covariance

$$\mu^{\Theta}_{\mathrm{Ens}}(\mathbf{x}') = \Theta_{\mathbf{x}',\mathbf{x}}\,\Theta^{-1}_{\mathbf{x},\mathbf{x}}\mathbf{g}^*, \tag{37}$$

$$\Sigma^{\mathcal{K},\Theta}_{\mathrm{Ens}}(\mathbf{x}',\mathbf{x}') = \mathcal{K}_{\mathbf{x}',\mathbf{x}'} - K_{\mathbf{x}',\mathbf{x}}\mathbf{A}^{-\top}\mathbf{H}^*\Theta_{\mathbf{x},\mathbf{x}'} - \Theta_{\mathbf{x}',\mathbf{x}}\mathbf{H}^*\mathbf{A}^{-1}\mathcal{K}_{\mathbf{x},\mathbf{x}'} \tag{38}$$

$$+ \Theta_{\mathbf{x}',\mathbf{x}}\mathbf{A}^{-\top}\mathbf{H}^*\mathcal{K}_{\mathbf{x},\mathbf{x}}\mathbf{H}^*\mathbf{A}^{-1}\Theta_{\mathbf{x},\mathbf{x}'}. \tag{39}$$

For $\mathcal{K} = \Theta$ the (approximate) ensemble covariance simplifies to

$$\Sigma^{\Theta,\Theta}_{\mathrm{Ens}}(\mathbf{x}',\mathbf{x}') = \Theta_{\mathbf{x}',\mathbf{x}'} - \Theta_{\mathbf{x}',\mathbf{x}}\left(\mathbf{H}^*\mathbf{A}^{-1} + \beta\mathbf{A}^{-\top}\mathbf{H}^*\mathbf{A}^{-1}\right)\Theta_{\mathbf{x},\mathbf{x}'}. \tag{40}$$

We have $\Sigma^{\mathcal{K},\Theta}_{\mathrm{Ens}}(\mathbf{x}',\mathbf{x}') \preceq \Sigma^{\Theta,\Theta}_{\mathrm{Ens}}(\mathbf{x}',\mathbf{x}')$ for $\mathcal{K} \preceq \Theta$.

The Laplace approximation (Rasmussen & Williams, 2005) is a standard tool for approximate Bayesian inference in classification GPs, in which model uncertainty is captured by a Bayesian posterior, rather than an ensemble covariance. The Laplace posterior mean is the same as our approximate ensemble mean. The Laplace posterior covariance is

$$\Sigma^{\Theta}_{\mathrm{Lap}}(\mathbf{x}',\mathbf{x}') = \Theta_{\mathbf{x}',\mathbf{x}'} - \Theta_{\mathbf{x}',\mathbf{x}}\mathbf{H}^*\mathbf{A}^{-1}\Theta_{\mathbf{x},\mathbf{x}'}. \tag{41}$$

For $\mathcal{K} = \Theta$ the gap to our approximate ensemble covariance is equal to

$$\Sigma^{\Theta}_{\mathrm{Lap}}(\mathbf{x}',\mathbf{x}') - \Sigma^{\Theta,\Theta}_{\mathrm{Ens}}(\mathbf{x}',\mathbf{x}') = \beta\Theta_{\mathbf{x}',\mathbf{x}}\mathbf{A}^{-\top}\mathbf{H}^*\mathbf{A}^{-1}\Theta_{\mathbf{x},\mathbf{x}'} \succeq 0. \tag{42}$$

Hence the Laplace approximate NTK-GP-posterior covariance is always more uncertain than our approximate ensemble covariance. This generalizes the analysis of He et al. (2020), who consider MSE loss where the ensemble covariance and NTK-GP posterior covariance are available in closed form.

5 Experiments

5.1 Classification toy dataset

Following Liu et al. (2020) we consider a one-dimensional toy dataset with 3 classes. As discussed in Section 3, they show that when the last layer of the network is defined as $\mathrm{softmax}$, the NTK does not remain constant, even when applying label smoothing or parameter regularization. In Figure 2 we show that the NTK of the pre-$\mathrm{softmax}$ layer does remain constant, allowing the linearization discussed in equation (20).

5.2 MNIST Classification

Following Yu et al. (2025) we train a four-layer fully connected neural network on MNIST (LeCun et al., 1998) using 2 classes (odd or even). By using $\mathrm{softmax}$ with a reference class, the logit dimension is 1 and thus the kernel is scalar-valued. We plot the evolution of the empirical NTK $t \mapsto \hat{\Theta}_{\theta_t}(x,x)$ during training in three input points $x \in \mathbb{R}^{784}$ in Figure 3. In line with Yu et al. (2025) we see that the empirical NTK diverges for the classification task with one-hot encodings. When using label smoothing and zero parameter regularization, it converges but still moves nontrivially from initialization. This behavior can be expected as the required layer width n in Theorem 3.1 grows inversely to the $\mu_{\mathcal{C}}$-constant of the function-space loss. When applying parameter regularization $\beta = 0.01$, the NTK remains almost constant during training, even for relatively low layer width.

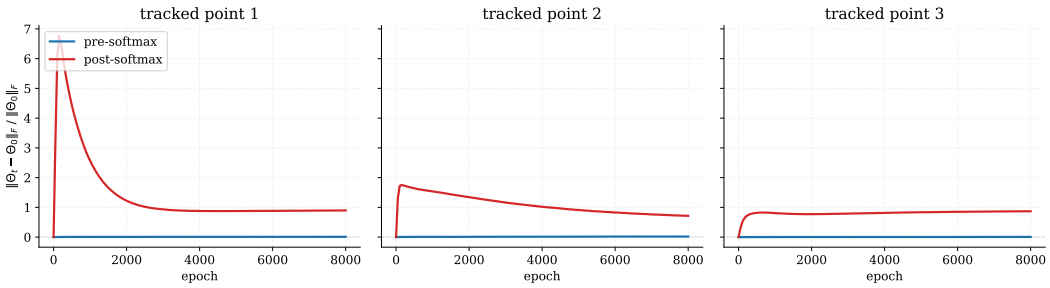

Figure 2: Blue: Pre-softmax NTK (constant). Red: Post-softmax NTK (not constant).

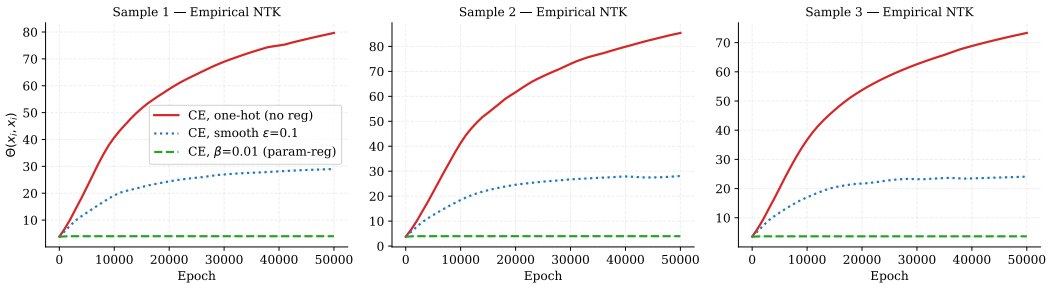

Figure 3: The NTK for MNIST classification does not diverge when using label smoothing or parameter regularization.

6 Related Work

The NTK regime has been studied for classification in Li & Liang (2018); Allen-Zhu et al. (2019); Cao & Gu (2019); Zou et al. (2020); Nitanda et al. (2019); Ji & Telgarsky (2019); Chen et al. (2019), for finite training times. However, when considering infite training times, a central obstruction arises: For cross-entropy or logistic losses on separable data, the optimum is attained only in the limit of diverging logits, and gradient descent drives parameters to infinity while converging only in direction (Soudry et al., 2018; Lyu & Li, 2019; Ji & Telgarsky, 2020). This behavior violates the small-perturbation assumptions underlying NTK linearization and leads to a breakdown of kernel constancy. Recent work formalizes this incompatibility by showing that, under cross-entropy training, the empirical NTK can itself diverge over time rather than remain fixed (Yu et al., 2025). A related limitation arises when nonlinear output layers are incorporated into the model: even at initialization, such architectures need not admit a constant tangent kernel (Liu et al., 2020).

In response to these limitations, prior work has largely moved beyond the NTK regime rather than restoring it. One line of research studies the implicit bias of gradient-based optimization under cross-entropy, characterizing the asymptotic direction of the parameters rather than a finite predictor. This yields max-margin solutions in linear and homogeneous models (Soudry et al., 2018; Lyu & Li, 2019) and extends to deep settings, where it connects to neural collapse geometry (Papyan et al., 2020; Ji et al., 2021; Thrampoulidis et al., 2022). Recent work provides more detailed descriptions of cross-entropy dynamics beyond asymptotics, including tractable path-level analyses in non-convex models (Garrod et al., 2025). A complementary line develops alternative infinite-width limits, such as mean-field and feature-learning regimes, which allow representations to evolve and depart from fixed-kernel behavior (Mei et al., 2018; Yang & Hu, 2021). While these approaches capture phenomena beyond the lazy regime, they do not yield a finite-predictor NTK description of classification. More recently, NTK-based constructions have been used for uncertainty estimation in classification by leveraging empirical NTK features within scalable approximations (Calvo-Ordoñez et al., 2026a), but without characterizing classification training dynamics within the NTK regime.

The present work addresses this gap by identifying conditions under which classification remains in a valid NTK regime. Rather than replacing cross-entropy or analyzing asymptotic directions, we

show that the failure modes of NTK classification can be avoided by modifying the optimization landscape. In particular, regularization around initialization controls parameter drift, while full-support target distributions—such as those induced by label smoothing (Szegedy et al., 2016; Müller et al., 2019)—eliminate the infinite-logit optimum and yield finite solutions. Under these conditions, the relevant object is the pre-activation predictor, whose tangent kernel remains effectively constant throughout training. This yields an NTK characterization of the classification dynamics in function-space.

7 Conclusion

We extend the Neural Tangent Kernel (NTK) theory to classification and more general losses by identifying conditions under which wide neural networks remain in the lazy training regime. In particular, we show that parameter-space regularization and full-support target distributions prevent the divergence of logits and ensure that training is well-approximated by the linearized model. This yields an explicit NTK characterization of classification dynamics at the level of finite predictors. We further analyze the distribution over trained networks induced by random initialization, connecting this notion of uncertainty to Bayesian interpretations. Finally, we empirically validate our theoretical predictions through controlled experimental setups.

AI use statement

In this work, we used generative AI tools to generate synthetic datasets, clean and reformat datasets, write code, assist in the writing of proofs, and cleaning and reformatting datasets. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, interpreting results. Assisting with translation, and supporting qualitative and thematic data analysis is not applicable to this work. Additionally, we used generative AI tools for proofreading, finding related work, and rewriting sections to improve readability. We have reviewed all AI-assisted work, and in particular we verified and tested LLM-generated code for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

The appendix includes clear explanations of any assumptions and a complete proof of the claims.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International conference on machine learning*, pp. 242–252. PMLR, 2019.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *Advances in neural information processing systems*, 32, 2019.

Sergio Calvo-Ordoñez, Jonathan Plenk, Richard Bergna, Álvaro Cartea, Yarin Gal, José Miguel Hernández-Lobato, and Kamil Ciosek. Richer bayesian last layers with subsampled ntk features. *arXiv preprint arXiv:2602.01279*, 2026a.

Sergio Calvo-Ordoñez, Jonathan Plenk, Richard Bergna, Álvaro Cartea, José Miguel Hernández-Lobato, Konstantina Palla, and Kamil Ciosek. A gaussian process view on observation noise and initialization in wide neural networks. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026b.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Luís Carvalho, João L Costa, José Mourão, and Gonçalo Oliveira. The positivity of the neural tangent kernel. *SIAM Journal on Mathematics of Data Science*, 7(2):495–515, 2025.

Zixiang Chen, Yuan Cao, Difan Zou, and Quanquan Gu. How much over-parameterization is sufficient to learn deep relu networks? *arXiv preprint arXiv:1911.12360*, 2019.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pp. 1675–1685. PMLR, 2019.

Connall Garrod, Jonathan P Keating, and Christos Thrampoulidis. Diagonalizing the softmax: Hadamard initialization for tractable cross-entropy dynamics. *arXiv preprint arXiv:2512.04006*, 2025.

Bobby He, Balaji Lakshminarayanan, and Yee Whye Teh. Bayesian deep ensembles via the neural tangent kernel. *Advances in neural information processing systems*, 33:1010–1022, 2020.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Wenlong Ji, Yiping Lu, Yiliang Zhang, Zhun Deng, and Weijie J Su. An unconstrained layer-peeled perspective on neural collapse. *arXiv preprint arXiv:2110.02796*, 2021.

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. *arXiv preprint arXiv:1909.12292*, 2019.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33:17176–17186, 2020.

Y LeCun, L Bottou, Y Bengio, and P Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. *International Conference on Learning Representations*, 2018.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.

Aitor Lewkowycz and Guy Gur-Ari. On the training dynamics of deep networks with l_2 regularization. *Advances in Neural Information Processing Systems*, 33:4790–4799, 2020.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. *Advances in neural information processing systems*, 31, 2018.

Chaoyue Liu, Libin Zhu, and Misha Belkin. On the linearity of large non-linear models: when and why the tangent kernel is constant. *Advances in Neural Information Processing Systems*, 33: 15954–15964, 2020.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

Chris J Maddison, Daniel Tarlow, and Tom Minka. A* sampling. *Advances in neural information processing systems*, 27, 2014.

Alexander G de G Matthews, Mark Rowland, Jiri Hron, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*, 2018.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

Atsushi Nitanda, Geoffrey Chinot, and Taiji Suzuki. Gradient descent can learn less over-parameterized two-layer neural networks on classification problems. *arXiv preprint arXiv:1905.09870*, 2019.

Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, pp. 4951–4960. PMLR, 2019.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2005.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70): 1–57, 2018.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Christos Thrampoulidis, Ganesh Ramachandra Kini, Vala Vakilian, and Tina Behnia. Imbalance trouble: Revisiting neural-collapse geometry. *Advances in Neural Information Processing Systems*, 35:27225–27238, 2022.

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

Greg Yang and Edward J Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pp. 11727–11737. PMLR, 2021.

Zixiong Yu, Songtao Tian, and Guhan Chen. Divergence of neural tangent kernel in classification problems. In *The Thirteenth International Conference on Learning Representations*, 2025.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine learning*, 109(3):467–492, 2020.

A Recap of NNGP and NTK theory for wide neural networks

A.1 Feedforward neural networks in NTK parametrization

Consider an input $x^0 := x \in \mathbb{R}^d$. A neural network with L hidden layers and linear last layer is defined through

$$x^{(l)} := \phi_l\left(x^{(l-1)}, \theta^{(l)}\right) \quad \text{for} \quad l = 1, \dots, L, \tag{43}$$

$$f_\theta(x) := \frac{1}{\sqrt{n_L}} W^{(L+1)} x^{(L)} + b^{(L+1)}. \tag{44}$$

Each hidden layer map $\phi_l : \mathbb{R}^{n_{l-1}} \times \mathbb{R}^{p_l} \to \mathbb{R}^{n_l}$ takes the previous layer $x^{(l-1)} \in \mathbb{R}^{n_{l-1}}$ and current parameters $\theta^{(l)} \in \mathbb{R}^{p_l}$ as input. The final layer parameters are $\theta^{L+1} = (W^{(L+1)}, b^{(L+1)}) \in \mathbb{R}^{p_{L+1}}$, where $W^{(L+1)} \in \mathbb{R}^{K \times n_L}$, $b^{(L+1)} \in \mathbb{R}^K$, $p_{L+1} = K(n_L + 1)$. The full parameter vector is $\theta = \left(\theta^{(l)}\right)_{l=1}^{L+1} \in \mathbb{R}^p$ and the total number of parameters is $p = \sum_{l=1}^{L+1} p_l$.

Consider an activation function $\phi : \mathbb{R} \to \mathbb{R}$ that is applied element-wise. A standard feedforward layer in NTK parameterization is defined by

$$\phi_l(x^{(l-1)}, \theta^{(l)}) = \phi\left(\frac{1}{\sqrt{n_{l-1}}} W^{(l)} x^{(l-1)} + b^{(l)}\right). \tag{45}$$

Here, the parameters are $\theta^{(l)} = (W^{(l)}, b^{(l)}) \in \mathbb{R}^{n_l(n_{l-1}+1)}$, where $W^{(l)} \in \mathbb{R}^{n_l \times n_{l-1}}$, $b^{(l)} \in \mathbb{R}^{n_l}$. They are randomly initialized through $W^{(l)}_{0,ij} \sim \mathcal{N}(0, \sigma^2_{w_l})$, $b^{(l)}_{0,i} \sim \mathcal{N}(0, \sigma^2_{b_l})$ iid. Similarly one can define a residual layer and a convolutional layer.

In the following, f_θ is a standard feedforward neural network whose non-polynomial activation function ϕ is twice continuously differentiable, and ϕ and ϕ' are globally Lipschitz.

A.2 Local boundedness and Lipschitzness of the parameter-Jacobian

Liu et al. (2020) prove for a standard feedforward neural network (as well as ResNets and CNNs with large number of channels):

Lemma 2.1. *For any $\delta_0 > 0$ there are $K_1', K_2' > 0$ such that: For every radius $R > 0$ there is large enough layer width n such that with probability $1 - \delta_0$ over random initialization θ_0: For any input $x \in M^d$:*

$$\forall \theta \in B(\theta_0, R) : \|J_\theta(x)\|_{2,2} \leq K_1', \tag{1}$$

$$\forall \theta \in B(\theta_0, R) : \|H_\theta(x)\|_{\mathrm{op},2} \leq \frac{(\log n)^c}{\sqrt{n}} K_2'. \tag{2}$$

Recall that $J_\theta(x) \in \mathbb{R}^{K \times p}$ and $H_\theta(x) \in \mathbb{R}^{K \times p \times p}$. We are using the Euclidean norm on the K-dimensional output dimension, after applying the Euclidean and spectral norm on the parameter dimensions p and $p \times p$ respectively. The constants $K_1', K_2' > 0$ do not depend on R. The constant $c > 0$ only depends on the network architecture.

To gain intuition consider a one-hidden-layer neural network with one-dimensional inputs and outputs and no biases:

$$f : \mathbb{R} \times \mathbb{R}^{2n} \to \mathbb{R}, \quad f_{a,w}(x) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} a_j \phi(w_j x). \tag{46}$$

As it is linear and therefore easy to analyze in the last-layer parameters a, we only look at the derivatives with respect to hidden weights w. The Jacobian and Hessian entries are

$$\frac{\partial f_\theta(x)}{\partial w_j} = \frac{1}{\sqrt{n}} a_j \phi'(w_j x) x, \quad \frac{\partial^2 f_\theta(x)}{\partial w_j \partial w_k} = \frac{1}{\sqrt{n}} a_j \phi''(w_j x) x^2 1_{\{j=k\}}. \tag{47}$$

Consider the random initialization $a_j, w_j \sim \mathcal{N}(0,1)$ iid. Recall that $|\phi'|$ and $|\phi''|$ are globally bounded. The Euclidean norm of the Jacobian at initialization is

$$\left\| \frac{\partial f_\theta(x)}{\partial w} \right\|_2 = \frac{1}{\sqrt{n}} |x| \sqrt{\sum_{j=1}^{n} a_j^2 \phi'(w_j x)^2} = O(1). \tag{48}$$

The spectral norm of the Hessian at initialization is

$$\left\| \frac{\partial^2 f_\theta(x)}{\partial w^2} \right\|_{\mathrm{op}} = \frac{1}{\sqrt{n}} x^2 \max_{j=1,\dots,n} |a_j \phi''(w_j x)| = O\left(\frac{\sqrt{\log n}}{\sqrt{n}} \right). \tag{49}$$

This shows that the network is $O(1)$-Lipschitz in the parameters, while its parameter-Jacobian is $O((\log n)^c/\sqrt{n})$-Lipschitz.

A.3 Convergence of the neural network at initialization

Lee et al. (2018); Matthews et al. (2018) prove:

Lemma 2.2. *Consider random initialization θ_0. Then $f_{\theta_0}(\cdot)$ converges in distribution to a Gaussian process with zero mean and covariance given by the NNGP Kernel $\mathcal{K}$: For inputs $\mathbf{x}_1, \dots, \mathbf{x}_N \in M^d$,*

$$f_{\theta_0}(\mathbf{x}) \xrightarrow{d} \mathcal{N}(0, \mathcal{K}(\mathbf{x}, \mathbf{x})). \tag{3}$$

This directly implies that the network values are in a compact set at initialization:

Lemma A.1. *For any $\delta_0 > 0$, there is $K_0' > 0$ such that for large enough layer width n, with probability $1 - \delta_0$ over random initialization θ_0: For any input $x \in M^d$:*

$$\|f_{\theta_0}(x)\|_2 \leq K_0'. \tag{50}$$

A.4 Convergence of the empirical NTK at initialization

The following result first appeared in Jacot et al. (2018), and was subsequently extended to more general architectures in Yang (2020). Arora et al. (2019) show convergence uniformly over input vectors in a compact set, which we restate in the following.

Lemma A.2. *Consider random initialization θ_0. Then, $\hat{\Theta}_{\theta_0}$ converges in probability (uniformly over inputs $x, x' \in M^d$) to a deterministic kernel:*

$$\hat{\Theta}_{\theta_0}(x, x') = J_{\theta_0}(x) J_{\theta_0}(x')^\top \xrightarrow{p} \Theta(x, x') \in \mathbb{R}^{K \times K}. \tag{51}$$

For multidimensional outputs, i.e. $K > 1$, $\Theta(x, x')$ is a diagonal matrix. Thus $\Theta(\mathbf{x}, \mathbf{x}') \in \mathbb{R}^{NK \times N'K}$ consists of K blocks of dimension $N \times N'$.

Carvalho et al. (2025) show that the NTK is strictly positive definite for all non-polynomial activations, which applies to our setting. We will use this for zero parameter regularizer $\beta = 0$. For positive parameter regularizer $\beta > 0$ this is not required, which would allow extending our analysis to more general loss functionals that involve infinitely many training points (e.g. PINNs).

B Properties of the function-space loss

In this section we introduce various assumptions on the function-space loss and discuss its properties. Table 1 provides an overview.

Table 1: Summary of assumptions on the function-space loss $\mathcal{C} : \mathbb{R}^{NK} \to \mathbb{R}$.

#	Description	Summary	
(1)	Bounded function-space gradient	$\forall \mathbf{z} \in \mathcal{S}_0 :$	$\lVert \nabla_{\mathbf{z}} \mathcal{C}(\mathbf{z}) \rVert_2 \leq K_1.$
(2)	Bounded gradient growth	$\forall \mathbf{z} \in \mathcal{S}_0 :$	$\lVert \nabla_{\mathbf{z}} \mathcal{C}(\mathbf{z}) \rVert_2^2 \leq 2K_2 \left(\mathcal{C}(\mathbf{z}) - \inf \mathcal{C}\right).$
(3)	Function-space PL	$\forall \mathbf{z} \in \mathcal{S}_0 :$	$\lVert \nabla_{\mathbf{z}} \mathcal{C}(\mathbf{z}) \rVert_2^2 \geq 2\mu_{\mathcal{C}} \left(\mathcal{C}(\mathbf{z}) - \inf \mathcal{C}\right).$

We write for $\mathbf{x}_1, \ldots, \mathbf{x}_N \in M^d$:

$$\mathbf{f}_\theta := f_\theta(\mathbf{x}) := (f_\theta(\mathbf{x}_i))_i \in \mathbb{R}^{NK} \tag{52}$$

and for the parameter-Jacobian and Hessian of the network,

$$\mathbf{J}_\theta := \frac{\partial}{\partial \theta} f_\theta(\mathbf{x}) \in \mathbb{R}^{NK \times p} \quad \text{and} \quad \mathbf{H}_\theta := \frac{\partial^2}{\partial \theta^2} f_\theta(\mathbf{x}) \in \mathbb{R}^{NK \times p \times p}. \tag{53}$$

The function-space loss gradient and Hessian are

$$\nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}_\theta) := (\nabla_{\mathbf{z}_i} \mathcal{C}(\mathbf{f}_\theta))_i \in \mathbb{R}^{NK} \quad \text{and} \quad \nabla^2_{\mathbf{zz}} \mathcal{C}(\mathbf{f}_\theta) := \left(\nabla^2_{\mathbf{z}_i \mathbf{z}_j} \mathcal{C}(\mathbf{f}_\theta)\right)_{ij} \in \mathbb{R}^{NK \times NK}. \tag{54}$$

For losses summing over observations (such as MSE and categorical cross-entropy) the function-space Hessian will be a block-diagonal consisting of N blocks of shape $K \times K$. The chain rule gives the parameter-space loss gradient and Hessian

$$\nabla_\theta \mathcal{C}(\mathbf{f}_\theta) = \mathbf{J}_\theta^\top \nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}_\theta) \in \mathbb{R}^p \quad \text{and} \quad \nabla^2_{\theta\theta} \mathcal{C}(\mathbf{f}_\theta) = \nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}_\theta)^\top \mathbf{H}_\theta + \mathbf{J}_\theta^\top \nabla^2_{\mathbf{zz}} \mathcal{C}(\mathbf{f}_\theta) \mathbf{J}_\theta \in \mathbb{R}^{p \times p}. \tag{55}$$

For $\beta = 0$ we assume that the function-space loss $\mathcal{C}$ is lower bounded, which implies that $\lim_{t \to T_{\max}} \mathcal{C}(\mathbf{f}_{\theta_t})$ exists and is finite.

For general $\beta \geq 0$, we have the following high-probability bound on the loss at initialization and during training:

Lemma 2.3. *Let $\delta_0 > 0$ be arbitrarily small. There is $K_0 > 0$, such that for large enough layer width n, with probability $1 - \delta_0$ over random initialization θ_0:*

$$\mathcal{C}(f_{\theta_t}(\mathbf{x})) \leq \mathcal{L}^\beta(\theta_t) \leq \mathcal{L}^\beta(\theta_0) = \mathcal{C}(f_{\theta_0}(\mathbf{x})) \leq K_0 \quad \textit{for all } t \geq 0. \tag{10}$$

Proof. Using the chain rule and the definition of the gradient flow, we have

$$\frac{d}{dt}\mathcal{L}^\beta(\theta_t) = \nabla_\theta \mathcal{L}^\beta(\theta_t)^\top \frac{d\theta_t}{dt} = -\eta_0 \nabla_\theta \mathcal{L}^\beta(\theta_t)^\top \nabla_\theta \mathcal{L}^\beta(\theta_t) = -\eta_0 \left\|\nabla_\theta \mathcal{L}^\beta(\theta_t)\right\|_2^2 \leq 0. \tag{56}$$

Thus $\mathcal{L}^\beta(\theta_t) \leq \mathcal{L}^\beta(\theta_0)$ and therefore

$$\mathcal{C}(\mathbf{f}_{\theta_t}) \leq \mathcal{C}(\mathbf{f}_{\theta_t}) + \frac{1}{2}\beta \left\|\theta_t - \theta_0\right\|_2^2 = \mathcal{L}^\beta(\theta_t) \leq \mathcal{L}^\beta(\theta_0) = \mathcal{C}(\mathbf{f}_{\theta_0}) \leq K_0. \tag{57}$$

The last inequality follows by Lemma A.1. □

Remark B.1. This Lemma does not apply to the weight-decay regularizer $\frac{1}{2}\left\|\theta\right\|_2^2$ analyzed in Lewkowycz & Gur-Ari (2020): In general, $\frac{1}{2}\left\|\theta_0\right\|_2^2 \neq 0$, and thus for $\beta > 0$ we cannot use $\mathcal{L}^\beta(\theta_0) = \mathcal{C}(f_{\theta_0})$. They show that using gradient flow with this regularizer leads to $f_{\theta_\infty}(\cdot) = 0$ in the infinite-width limit, and we are thus not in the NTK regime.

B.1 Bounded function-space loss gradient

Assumption 1 (Bounded function-space loss gradient)**.** There is $K_1 > 0$ such that the norm of the gradient of the function-space loss is bounded on the loss sublevel set:

$$\left\|\nabla_{\mathbf{z}} \mathcal{C}(\mathbf{z})\right\|_2 \leq K_1 \quad \text{for all } \mathbf{z} \in \mathcal{S}_0. \tag{58}$$

B.2 Bounded gradient growth

Assumption 2 (Bounded gradient growth)**.** There is $K_2 > 0$ such that

$$\left\|\nabla_{\mathbf{z}} \mathcal{C}(\mathbf{z})\right\|_2^2 \leq 2K_2 \left(\mathcal{C}(\mathbf{z}) - \inf \mathcal{C}\right) \quad \text{for all } \mathbf{z} \in \mathcal{S}_0. \tag{59}$$

This assumption is satisfied if the eigenvalues of the function-space Hessian are globally bounded, as the following lemma shows.

Lemma B.2. *Assume that* $\left\|\nabla_{\mathbf{zz}}^2 \mathcal{C}(\mathbf{z})\right\|_{\text{op}} \leq K_2$ *for all* $\mathbf{z} \in \mathbb{R}^{NK}$*. Then,*

$$\left\|\nabla_{\mathbf{z}} \mathcal{C}(\mathbf{z})\right\|_2^2 \leq 2K_2 \left(\mathcal{C}(\mathbf{z}) - \inf \mathcal{C}\right) \quad \textit{for all } \mathbf{z} \in \mathbb{R}^{NK}. \tag{60}$$

Proof. For $\mathbf{y}, \mathbf{z} \in \mathbb{R}^{NK}$,

$$\inf C \leq \mathcal{C}(\mathbf{y}) \tag{61}$$

$$= \mathcal{C}(\mathbf{z}) + \nabla_{\mathbf{z}} \mathcal{C}(\mathbf{z})^\top (\mathbf{y} - \mathbf{z}) + \int_0^1 (1-s)(\mathbf{y} - \mathbf{z})^\top \nabla_{\mathbf{zz}}^2 \mathcal{C}\left(\mathbf{z} + s(\mathbf{y} - \mathbf{z})\right)(\mathbf{y} - \mathbf{z}) ds \tag{62}$$

$$\leq \mathcal{C}(\mathbf{z}) + \nabla_{\mathbf{z}} \mathcal{C}(\mathbf{z})^\top (\mathbf{y} - \mathbf{z}) + \frac{K_2}{2} \|\mathbf{y} - \mathbf{z}\|_2^2. \tag{63}$$

Thus using $\mathbf{y} = \mathbf{z} - \frac{1}{K_2}\nabla_{\mathbf{z}} \mathcal{C}(\mathbf{z})$,

$$\inf \mathcal{C} \leq \mathcal{C}(\mathbf{z}) - \frac{1}{2K_2} \|\nabla_{\mathbf{z}} \mathcal{C}(\mathbf{z})\|_2^2. \tag{64}$$

□

B.3 Function-space PL-condition

This section is only relevant for zero parameter regularizer $\beta = 0$.

Assumption 3 (Function-space PL)**.** There is $\mu_{\mathcal{C}} > 0$, such that

$$\left\|\nabla_{\mathbf{z}} \mathcal{C}(\mathbf{z})\right\|_2^2 \geq 2\mu_{\mathcal{C}} \left(\mathcal{C}(\mathbf{z}) - \inf \mathcal{C}\right) \quad \text{for all } \mathbf{z} \in \mathcal{S}_0. \tag{65}$$

Oymak & Soltanolkotabi (2019); Liu et al. (2022) instead defined the PL-condition in parameter-space for all θ in a ball around θ_0, whose radius is required to be large enough depending on $\mu_{\mathcal{L}}$. Along the trajectory, this means that there is $\mu_{\mathcal{L}} > 0$ such that

$$\|\nabla_\theta \mathcal{C}(\mathbf{f}_{\theta_t})\|_2^2 \geq 2\mu_{\mathcal{L}} \left(\mathcal{C}(\mathbf{f}_{\theta_t}) - \inf \mathcal{C}\right) \quad \text{for all } t \geq 0. \tag{66}$$

As we can lower bound the parameter-space loss-gradient norm by

$$\|\nabla_\theta \mathcal{C}(\mathbf{f}_{\theta_t})\|_2^2 = \left\|\mathbf{J}_{\theta_t}^\top \nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}_{\theta_t})\right\|_2^2 \geq \lambda_{\min}(\hat{\Theta}_{\theta_t}(\mathbf{x}, \mathbf{x})) \left\|\nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}_{\theta_t})\right\|_2^2, \tag{67}$$

it is natural to only assume the $\mu_{\mathcal{C}}$-PL-condition in function-space along the training trajectory if one can additionally show that the empirical neural tangent kernel is sufficiently well-conditioned during training.

A sufficient and easy-to-verify condition that implies the function-space PL condition is strong convexity. It is however not a necessary condition, as e.g. the cross-entropy loss (without a reference class) satisfies the PL-condition but is not strongly convex. For completeness, we prove this in the following lemma, under a slightly more general condition than strong convexity.

Lemma B.3. *Assume that $\mathcal{S}_0$ is convex, and there is $\mu_{\mathcal{C}} > 0$ such that*

$$\lambda_{\min}(\nabla^2_{\mathbf{zz}} \mathcal{C}(\mathbf{z})) \geq \mu_{\mathcal{C}} > 0 \quad \textit{for all } \mathbf{z} \in \mathcal{S}_0. \tag{68}$$

Then it satisfies the function-space PL-condition (Assumption 3).

Proof. We can write for $\mathbf{y}, \mathbf{z} \in \mathcal{S}_0$,

$$\mathcal{C}(\mathbf{y}) = \mathcal{C}(\mathbf{z}) + \nabla_{\mathbf{z}} \mathcal{C}(\mathbf{z})^\top (\mathbf{y} - \mathbf{z}) + \int_0^1 (1-s)(\mathbf{y} - \mathbf{z})^\top \nabla^2_{\mathbf{zz}} \mathcal{C}\left(\mathbf{z} + s(\mathbf{y} - \mathbf{z})\right)(\mathbf{y} - \mathbf{z}) ds \tag{69}$$

$$\geq \mathcal{C}(\mathbf{z}) + \nabla_{\mathbf{z}} \mathcal{C}(\mathbf{z})^\top (\mathbf{y} - \mathbf{z}) + \frac{\mu_{\mathcal{C}}}{2} \|\mathbf{y} - \mathbf{z}\|_2^2 \tag{70}$$

$$\geq \mathcal{C}(\mathbf{z}) - \frac{1}{2\mu_{\mathcal{C}}} \|\nabla_{\mathbf{z}} \mathcal{C}(\mathbf{z})\|_2^2. \tag{71}$$

The function-space PL-condition follows by using $\inf \mathcal{C} = \inf_{\mathbf{y} \in \mathcal{S}_0} \mathcal{C}(\mathbf{y})$ and taking $\mathbf{z} = \mathbf{f}_{\theta_t}$. □

B.4 Positive (semi-)definite function-space Hessian

For $\beta > 0$, we need to assume that the function-space loss $\mathcal{C} : \mathbb{R}^{NK} \to \mathbb{R}$ is convex, as shown by the following example. Consider two classes and a single training point (x, y) with target probability $(0.5, 0.5)$. The regularized function-space objective using the Brier score with a reference class is

$$\mathcal{C}^\beta_{\mathrm{BS}}(z) = \frac{1}{2}\left(\frac{e^z}{1+e^z} - y\right)^2 + \frac{1}{2}\frac{\beta}{\Theta_{x,x}}(z - z_0)^2. \tag{72}$$

Take $\Theta_{x,x} = 1$ and $y = 0.5$. Then, for $\beta = 0.01$, and $z_0 = 5.5$, $\mathcal{C}^\beta_{\mathrm{BS}}$ has 3 stationary points (see Figure 4): A local minimum at $z \approx 0.98$, a local maximum at $z \approx 2.47$, and a global minimum at $z \approx 5.24$.

This is fundamentally different from $\beta = 0$, where the function-space loss is nonconvex but has a unique stationary point $z^* = \mathrm{sxr}^{-1}(0.5) = 0$ (around which it is strictly convex).

For $\beta = 0$, we do not need convexity to establish lazy training/exponential convergence, but can instead rely on the weaker function-space PL condition (Assumption 3). However, to show the closeness to the linearized training dynamics, we need to assume that the Hessian is strictly positive definite. We will relax this for the cross-entropy loss in Appendix E.3.

C Examples of function-space losses

C.1 MSE loss

Consider the function-space MSE loss

$$\mathcal{C}_{\mathrm{MSE}} : \mathbb{R}^{NK} \to \mathbb{R}, \quad \mathcal{C}_{\mathrm{MSE}}(\mathbf{z}) := \sum_{i=1}^N \frac{1}{2} \|\mathbf{z}_i - \mathbf{y}_i\|_2^2. \tag{73}$$

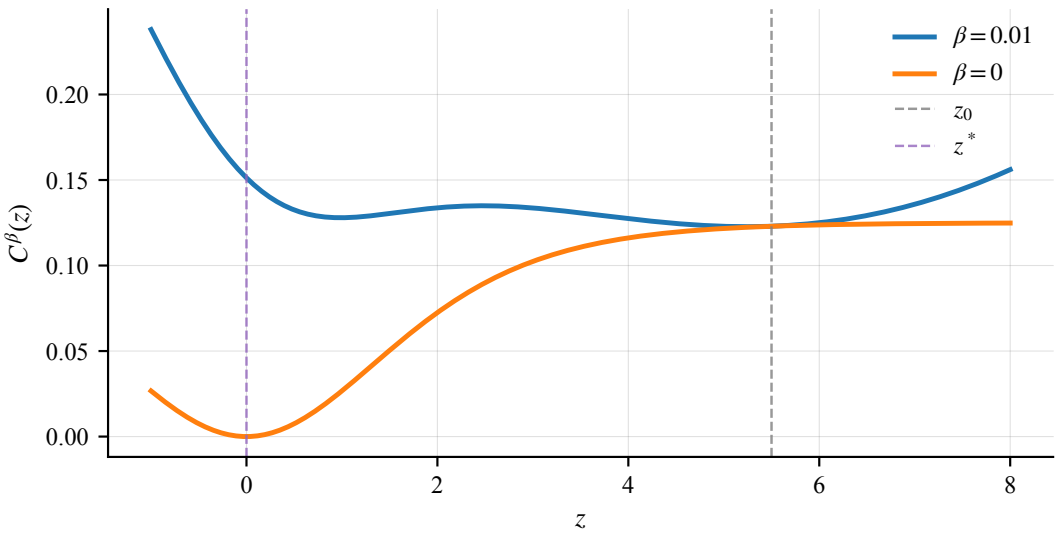

Figure 4: The function-space Brier score with a regularizer can have multiple stationary points.

It is twice continuously differentiable with

$$\nabla_{\mathbf{z}} \mathcal{C}_{\mathrm{MSE}}(\mathbf{z}) = \mathbf{z} - \mathbf{y} \in \mathbb{R}^{NK} \quad \text{and} \quad \nabla^2_{\mathbf{zz}} \mathcal{C}_{\mathrm{MSE}}(\mathbf{z}) = I_{NK}. \tag{74}$$

Lemma C.1. *$\mathcal{C}_{\mathrm{MSE}}$ satisfies Assumption 1.*

Proof. For MSE loss we conveniently have

$$\|\nabla_{\mathbf{z}} \mathcal{C}_{\mathrm{MSE}}(\mathbf{z})\|_2^2 = \sum_{i=1}^{N} \|\mathbf{z}_i - \mathbf{y}_i\|_2^2 = 2\mathcal{C}_{\mathrm{MSE}}(\mathbf{z}) \leq 2K_0. \tag{75}$$

□

As $\nabla^2_{\mathbf{zz}} \mathcal{C}_{\mathrm{MSE}}(\mathbf{z}) = I_{NK}$, the MSE loss satisfies Assumptions 2, 3 and its Hessian is strictly positive definite.

C.2 Categorical cross-entropy loss

Consider a classification task with K classes. The probability simplex in $\mathbb{R}^K$ is

$$\Delta^{K-1} := \left\{ p \in [0,1]^K : \sum_{k=1}^{K} p_k = 1 \right\} \subseteq \mathbb{R}^K. \tag{76}$$

The softmax-function is defined as

$$\mathrm{smx} : \mathbb{R}^K \to \mathrm{int}(\Delta^{K-1}), \quad \mathrm{smx}(z)_k := \frac{e^{z_k}}{\sum_{k'=1}^{K} e^{z_{k'}}}. \tag{77}$$

For N training points consider target probabilities $\mathbf{p}_i \in \Delta^{K-1}$. The function-space categorical cross-entropy loss is

$$\mathcal{C}_{\mathrm{CE}} : \mathbb{R}^{NK} \to \mathbb{R}, \quad \mathcal{C}_{\mathrm{CE}}(\mathbf{z}) := -\sum_{i=1}^{N} \sum_{k=1}^{K} \mathbf{p}_{ik} \log(\mathrm{smx}(\mathbf{z}_i)_k). \tag{78}$$

This simplifies to

$$\mathcal{C}_{\mathrm{CE}}(\mathbf{z}) = \sum_{i=1}^{N} \left(-\sum_{k=1}^{K} \mathbf{p}_{ik} \mathbf{z}_{ik} + \log\left(\sum_{k=1}^{K} e^{\mathbf{z}_{ik}} \right) \right). \tag{79}$$

$\mathcal{C}_{\mathrm{CE}}$ is twice continuously differentiable with

$$\nabla_{\mathbf{z}_i} \mathcal{C}_{\mathrm{CE}}(\mathbf{z}) = \mathrm{smx}(\mathbf{z}_i) - \mathbf{p}_i \in \mathbb{R}^K, \tag{80}$$

$$\nabla^2_{\mathbf{z}_i \mathbf{z}_i} \mathcal{C}_{\mathrm{CE}}(\mathbf{z}) = \mathrm{diag}\left(\mathrm{smx}(\mathbf{z}_i)\right) - \mathrm{smx}(\mathbf{z}_i)\,\mathrm{smx}(\mathbf{z}_i)^\top \in \mathbb{R}^{K \times K}. \tag{81}$$

Lemma C.2. *$\mathcal{C}_{\mathrm{CE}}$ satisfies Assumption 1.*

Proof. For any $\mathbf{z} \in \mathbb{R}^{NK}$,

$$\|\nabla_{\mathbf{z}} \mathcal{C}_{\mathrm{CE}}(\mathbf{z})\|_2^2 = \sum_{i=1}^N \|\nabla_{\mathbf{z}_i} \mathcal{C}_{\mathrm{CE}}(\mathbf{z})\|_2^2 = \sum_{i=1}^N \|\mathrm{smx}(\mathbf{z}_i) - \mathbf{p}_i\|_2^2 \leq 2N. \tag{82}$$

□

Lemma C.3. *$\mathcal{C}_{\mathrm{CE}}$ satisfies Assumption 2.*

Proof. For any $\mathbf{z} \in \mathbb{R}^{NK}$ the function-space Hessian $\nabla^2_{\mathbf{zz}} \mathcal{C}_{\mathrm{CE}}(\mathbf{z}) \in \mathbb{R}^{NK \times NK}$ is a block-diagonal matrix with blocks $\nabla^2_{\mathbf{z}_i \mathbf{z}_i} \mathcal{C}_{\mathrm{CE}}(\mathbf{z}) \in \mathbb{R}^{K \times K}$. Thus its spectral norm is

$$\left\|\nabla^2_{\mathbf{zz}} \mathcal{C}_{\mathrm{CE}}(\mathbf{z})\right\|_{\mathrm{op}} = \max_{i=1,\ldots,N} \left\|\nabla^2_{\mathbf{z}_i \mathbf{z}_i} \mathcal{C}_{\mathrm{CE}}(\mathbf{z})\right\|_{\mathrm{op}} \tag{83}$$

$$= \max_{i=1,\ldots,N} \left\|\mathrm{diag}(\mathrm{smx}(\mathbf{z}_i)) - \mathrm{smx}(\mathbf{z}_i)\,\mathrm{smx}(\mathbf{z}_i)^\top\right\|_{\mathrm{op}} \leq \frac{1}{2}. \tag{84}$$

□

Lemma C.4. *Assume the targets are a discrete probability distribution with full support, i.e. $\mathbf{p}_{ik} \in (0, 1)$. Then $\mathcal{C}_{\mathrm{CE}}$ satisfies Assumption 3.*

Proof. Let $\mathbf{z} \in \mathcal{S}_0$. As cross-entropy is nonnegative, we have for $i = 1, \ldots, N$,

$$-\sum_{k=1}^K \mathbf{p}_{ik} \log(\mathrm{smx}(\mathbf{z}_i)_k) \leq -\sum_{j=1}^N \sum_{k=1}^K \mathbf{p}_{jk} \log(\mathrm{smx}(\mathbf{z}_j)_k) = \mathcal{C}_{\mathrm{CE}}(\mathbf{z}) \leq K_0. \tag{85}$$

This gives a lower bound on the probits on the loss sublevel set:

$$\mathrm{smx}(\mathbf{z}_i)_k \geq e^{-K_0/\mathbf{p}_{ik}} \geq \min_{i,k} e^{-K_0/\mathbf{p}_{ik}} =: 2\mu_{\mathcal{C}} > 0. \tag{86}$$

Using the $KL - \chi^2$ inequality it follows that

$$\mathcal{C}_{\mathrm{CE}}(\mathbf{z}) - \inf \mathcal{C}_{\mathrm{CE}} = \mathcal{C}_{\mathrm{CE}}(\mathbf{z}) - \sum_{i=1}^N \mathrm{H}(\mathbf{p}_i) \tag{87}$$

$$= \sum_{i=1}^N \sum_{k=1}^K \mathbf{p}_{ik} \log\left(\frac{\mathbf{p}_{ik}}{\mathrm{smx}(\mathbf{z}_i)_k}\right) \tag{88}$$

$$\leq \sum_{i=1}^N \sum_{k=1}^K \frac{(\mathbf{p}_{ik} - \mathrm{smx}(\mathbf{z}_i)_k)^2}{\mathrm{smx}(\mathbf{z}_i)_k} \tag{89}$$

$$\leq \frac{1}{2\mu_{\mathcal{C}}} \sum_{i=1}^N \sum_{k=1}^K (\mathbf{p}_{ik} - \mathrm{smx}(\mathbf{z}_i)_k)^2 = \frac{1}{2\mu_{\mathcal{C}}} \|\nabla_{\mathbf{z}} \mathcal{C}_{\mathrm{CE}}(\mathbf{z})\|_2^2. \tag{90}$$

□

$\mathcal{C}_{\mathrm{CE}}$ is convex, but its Hessian has eigenvalue zero due to the redundancy $\mathrm{smx}(z) = \mathrm{smx}(z + \alpha 1)$.

C.3 Categorical cross-entropy loss with a reference class

Consider $K+1$ classes with target probabilities $(\mathbf{p}_{i0}, \mathbf{p}_i) \in \Delta^K$.[3] We use the 0-th class with target probabilities $\mathbf{p}_{i0} = 1 - \sum_{k=1}^K \mathbf{p}_{ik}$ as a reference class, by setting $z_0 = 0$. We define the last K coordinates of the $\mathrm{softmax}$-function with a reference class by

$$\mathrm{sxr} : \mathbb{R}^K \to \mathbb{R}^K, \quad \mathrm{sxr}(z)_k := \frac{e^{z_k}}{1 + \sum_{k'=1}^K e^{z_{k'}}}. \tag{91}$$

The probability of the reference class is $\mathrm{sxr}(z)_0 = (1 + \sum_{k=1}^K e^{z_k})^{-1}$. The resulting $z \mapsto (\mathrm{sxr}(z)_0, \mathrm{sxr}(z))$ is a bijection between K-dimensional vectors and discrete probability distributions on $\{0, \ldots, K\}$ with full support, i.e. $\mathrm{int}(\Delta^K)$.

The function-space categorical cross-entropy loss with a reference class is

$$\mathcal{C}_{\mathrm{CEr}} : \mathbb{R}^{NK} \to \mathbb{R}, \quad \mathcal{C}_{\mathrm{CEr}}(\mathbf{z}) := -\sum_{i=1}^N \sum_{k=0}^K \mathbf{p}_{ik} \log(\mathrm{sxr}(\mathbf{z}_i)_k). \tag{92}$$

This is equal to

$$\mathcal{C}_{\mathrm{CEr}}(\mathbf{z}) = \sum_{i=1}^N \left(-\sum_{k=1}^K \mathbf{p}_{ik} \mathbf{z}_{ik} + \log\left(1 + \sum_{k=1}^K e^{\mathbf{z}_{ik}}\right)\right). \tag{93}$$

For 2 classes (logit dimension $K = 1$) we get the logit loss, which for one-hot labels $\mathbf{p}_{i1} \in \{0, 1\}$ and $\mathbf{y}_i := 1_{\{\mathbf{p}_{i1}=1\}} - 1_{\{\mathbf{p}_{i1}=0\}}$ simplifies to

$$\mathcal{C}_{CEr}(\mathbf{z}) = \sum_{i=1}^N \left(-\mathbf{p}_{i1}\mathbf{z}_i + \log\left(1 + e^{\mathbf{z}_i}\right)\right) = \sum_{i=1}^N \log\left(1 + e^{-\mathbf{y}_i \mathbf{z}_i}\right). \tag{94}$$

$\mathcal{C}_{\mathrm{CEr}}$ is twice continuously differentiable with

$$\nabla_{\mathbf{z}_i} \mathcal{C}_{\mathrm{CEr}}(\mathbf{z}) = \mathrm{sxr}(\mathbf{z}_i) - \mathbf{p}_i \in \mathbb{R}^K, \tag{95}$$

$$\nabla^2_{\mathbf{z}_i \mathbf{z}_i} \mathcal{C}_{\mathrm{CEr}}(\mathbf{z}) = \mathrm{diag}\left(\mathrm{sxr}(\mathbf{z}_i)\right) - \mathrm{sxr}(\mathbf{z}_i)\,\mathrm{sxr}(\mathbf{z}_i)^\top \in \mathbb{R}^{K \times K}. \tag{96}$$

We can show that the categorical cross-entropy loss with a reference class satisfies Assumptions 1, 2, 3 just like for the standard categorical cross-entropy loss. Moreover, removing the redundancy along $\alpha \mapsto \mathbf{z} + \alpha 1$ makes its Hessian strictly positive definite.

Lemma C.5. *The Hessian of $\mathcal{C}_{\mathrm{CEr}}$ is strictly positive definite.*

Proof. Write $\mathbf{s}_i := \mathrm{sxr}(\mathbf{z}_i)$, then $\nabla^2_{\mathbf{z}_i \mathbf{z}_i} \mathcal{C}_{\mathrm{CEr}}(\mathbf{z}) = \mathrm{diag}(\mathbf{s}_i) - \mathbf{s}_i \mathbf{s}_i^\top$. For any $v \in \mathbb{R}^K$ with $v \neq 0$ the Cauchy-Schwarz inequality gives

$$v^T \nabla^2_{\mathbf{z}_i \mathbf{z}_i} \mathcal{C}_{\mathrm{CEr}}(\mathbf{z}) v = \sum_{k=1}^K \mathbf{s}_{ik} v_k^2 - \left(\sum_{k=1}^K \mathbf{s}_{ik} v_k\right)^2 \tag{97}$$

$$\geq \sum_{k=1}^K \mathbf{s}_{ik} v_k^2 - \left(\sum_{k=1}^K \mathbf{s}_{ik}\right)\left(\sum_{k=1}^K \mathbf{s}_{ik} v_k^2\right) \tag{98}$$

$$= \left(1 - \sum_{k=1}^K \mathbf{s}_{ik}\right)\left(\sum_{k=1}^K \mathbf{s}_{ik} v_k^2\right) > 0. \tag{99}$$

□

D Proofs for positive parameter regularizer

The proof consists of two steps. First, we prove that the network is in the lazy training regime. Then, we prove that it can be approximated by its linearization during training.

[3] We consider $K+1$ classes when using smx with a reference class to achieve notational consistency with the function-space loss input dimension NK.

D.1 EXPONENTIAL DECAY OF THE LOSS-GRADIENT NORM

Theorem D.1. *Let $\beta > 0$. Assume that Assumption 1 holds, and that the function-space loss is convex. Then there are constants R, $c_\beta > 0$ such that for large enough layer width n: With probability $1 - \delta_0$ over random initialization θ_0, for all $t \geq 0$,*

$$\left\| \frac{d}{dt}\theta_t \right\|_2 \leq \eta_0 c_\beta R e^{-c_\beta \eta_0 t} \quad \textit{and thus} \quad \|\theta_t - \theta_0\|_2 \leq R. \tag{100}$$

Proof. By Assumption 1 there is $K_1 > 0$, such that for n large enough, with probability $1 - \frac{1}{2}\delta_0$ over random initialization, $\|\nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}_{\theta_t})\|_2 \leq K_1$. Further, using Lemma 2.1, let K_1' be the constant for local Lipschitzness/Boundedness of the Jacobian with probability $1 - \frac{1}{2}\delta_0$ for n large enough.

Define $c_\beta := \frac{1}{2}\beta$ and $R := \frac{K_1 \sqrt{N} K_1'}{c_\beta}$. By Picard-Lindelöf the gradient flow ODE has a unique solution as long as $\theta_t \in B(\theta_0, R)$. Consider such $t \geq 0$. Recall that

$$\frac{d\theta_t}{dt} = -\eta_0 \nabla_\theta \mathcal{L}^\beta(\theta_t) = -\eta_0 \left(\nabla_\theta \mathcal{C}(\mathbf{f}_{\theta_t}) + \beta(\theta_t - \theta_0) \right). \tag{101}$$

The dynamics of the squared norm of the gradient are

$$\frac{d}{dt} \left\| \nabla_\theta \mathcal{L}^\beta(\theta_t) \right\|_2^2 = 2 \left(\nabla_\theta \mathcal{L}^\beta(\theta_t) \right)^\top \nabla^2_{\theta\theta} \mathcal{L}^\beta(\theta_t) \frac{d\theta_t}{dt} \tag{102}$$

$$= -2\eta_0 \left(\nabla_\theta \mathcal{L}^\beta(\theta_t) \right)^\top \nabla^2_{\theta\theta} \mathcal{L}^\beta(\theta_t) \left(\nabla_\theta \mathcal{L}^\beta(\theta_t) \right). \tag{103}$$

The Hessian of the regularized loss is

$$\nabla^2_{\theta\theta} \mathcal{L}^\beta(\theta_t) = \nabla^2_{\theta\theta} \mathcal{C}(\mathbf{f}_{\theta_t}) + \beta I_p = \nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}_{\theta_t})^\top \mathbf{H}_{\theta_t} + \mathbf{J}_{\theta_t}^\top \nabla^2_{\mathbf{z}} \mathcal{C}(\mathbf{f}_{\theta_t}) \mathbf{J}_{\theta_t} + \beta I_p \in \mathbb{R}^{p \times p}. \tag{104}$$

We can bound the matrix norm of the model curvature term $\nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}_{\theta_t})^\top \mathbf{H}_{\theta_t}$ using Assumption 1 and the bound of the parameter-Hessian from Lemma 2.1:

$$\left\| \nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}_{\theta_t})^\top \mathbf{H}_{\theta_t} \right\|_{\text{op}} \leq \left\| \nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}_{\theta_t}) \right\|_2 \left\| \mathbf{H}_{\theta_t} \right\|_{\text{op},2} \leq K_1 \sqrt{N} \frac{(\log n)^c}{\sqrt{n}} K_2'. \tag{105}$$

This implies that its lowest negative eigenvalue is of order $O((\log n)^c / \sqrt{n})$. The Gauss-Newton $\mathbf{J}_{\theta_t}^\top \nabla^2_{\mathbf{z}} \mathcal{C}(\mathbf{f}_{\theta_t}) \mathbf{J}_{\theta_t} \in \mathbb{R}^{p \times p}$ is positive semi-definite. Hence for positive $\beta > 0$ there is large enough layer width n such that

$$\lambda_{\min} \left(\nabla^2_{\theta\theta} \mathcal{L}^\beta(\theta_t) \right) \geq -\frac{(\log n)^c}{\sqrt{n}} K_1 \sqrt{N} K_2' + 0 + \beta \geq \frac{\beta}{2} =: c_\beta. \tag{106}$$

Thus,

$$\frac{d}{dt} \left\| \nabla_\theta \mathcal{L}^\beta(\theta_t) \right\|_2^2 \leq -2\eta_0 c_\beta \left\| \nabla_\theta \mathcal{L}^\beta(\theta_t) \right\|_2^2. \tag{107}$$

By Gronwall's inequality, the norm of the gradient decays exponentially:

$$\left\| \nabla_\theta \mathcal{L}^\beta(\theta_t) \right\|_2^2 \leq e^{-2\eta_0 c_\beta t} \left\| \nabla_\theta \mathcal{L}^\beta(\theta_0) \right\|_2^2. \tag{108}$$

We can bound the norm of the initial gradient by using Assumption 1 and the bound on the parameter-Jacobian from Lemma 2.1:

$$\left\| \nabla_\theta \mathcal{L}^\beta(\theta_0) \right\|_2 = \left\| \mathbf{J}_{\theta_0}^\top \nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}_{\theta_0}) \right\|_2 \leq \left\| \nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}_{\theta_0}) \right\|_2 \left\| \mathbf{J}_{\theta_0} \right\|_{2,2} \leq K_1 \sqrt{N} K_1'. \tag{109}$$

Hence,

$$\left\| \frac{d\theta_t}{dt} \right\|_2 = \eta_0 \left\| \nabla_\theta \mathcal{L}^\beta(\theta_t) \right\|_2 \leq \eta_0 e^{-\eta_0 c_\beta t} \left\| \nabla_\theta \mathcal{L}^\beta(\theta_0) \right\|_2 \leq \eta_0 K_1 \sqrt{N} K_1' e^{-\eta_0 c_\beta t}. \tag{110}$$

For the distance of parameters from initialization,

$$\|\theta_t - \theta_0\|_2 \leq \int_0^t \left\| \frac{d\theta_t}{du} \right\|_2 du \leq \eta_0 K_1 \sqrt{N} K_1' \int_0^t e^{-\eta_0 c_\beta u} du = \frac{K_1 \sqrt{N} K_1'}{c_\beta} (1 - e^{-\eta_0 c_\beta t}). \tag{111}$$

Thus $\|\theta_t - \theta_0\|_2 < R$, and the above holds for all $t \geq 0$. □

D.2 CLOSENESS TO LINEARIZATION

Theorem D.2. *Let $\beta > 0$. Consider a function-space loss that satisfies Assumption 1 and is convex. Then there is a constant $C_2 > 0$ such that for large enough layer width n: With probability $1 - \delta_0$ over random initialization θ_0,*

$$\forall x \in M^d, t \geq 0 : \quad \left\| f_{\theta_t}(x) - f^{\text{lin}}_{\theta^{\text{lin}}_t}(x) \right\|_2 \leq C_2 \frac{(\log n)^c}{\sqrt{n}}. \tag{112}$$

Recall that the linearized network and parameters of training with the linearized network are

$$f^{\text{lin}}_\theta(x) = f_{\theta_0}(x) + J_{\theta_0}(x)(\theta - \theta_0) \quad \text{and} \quad \frac{d}{dt}\theta^{\text{lin}}_t = -\eta_0 \left(\nabla_\theta \mathcal{C}(\mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t}) + \beta(\theta^{\text{lin}}_t - \theta_0) \right). \tag{113}$$

For general function-space losses, θ^{lin}_t does not follow a linear ODE. However, its dynamics are much easier to understand. Using the triangle inequality, we bound

$$\left\| f_{\theta_t}(x) - f^{\text{lin}}_{\theta^{\text{lin}}_t}(x) \right\|_2 \leq \left\| f_{\theta_t}(x) - f^{\text{lin}}_{\theta_t}(x) \right\|_2 + \left\| f^{\text{lin}}_{\theta_t}(x) - f^{\text{lin}}_{\theta^{\text{lin}}_t}(x) \right\|_2 \tag{114}$$

$$= \left\| f_{\theta_t}(x) - f^{\text{lin}}_{\theta_t}(x) \right\|_2 + \left\| J_{\theta_0}(x) \left(\theta_t - \theta^{\text{lin}}_t \right) \right\|_2 \tag{115}$$

We will show that $\left\| f_{\theta_t}(x) - f^{\text{lin}}_{\theta_t}(x) \right\|_2$ and $\left\| \theta_t - \theta^{\text{lin}}_t \right\|_2$ are $O((\log n)^c/\sqrt{n})$.

Lemma D.3. *Let $\delta_0 > 0$. The lazy training property (100) implies that there is large enough layer width n such that with probability $1-\delta_0$ over random initialization θ_0, the trajectories of the training outputs of the network and its linearization are in a compact set:*

$$\mathbf{f}_{\theta_t}, \mathbf{f}^{\text{lin}}_{\theta_t}, \mathbf{f}_{\theta^{\text{lin}}_t}, \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t} \in B(0; \sqrt{N}(K'_1 R + K'_0)) \quad \textit{for all } t \geq 0. \tag{116}$$

Denote by $L_{\mathcal{C}}$ and $L_{\nabla \mathcal{C}}$ the Lipschitz-constants of $\mathcal{C}$ and $\nabla \mathcal{C}$ on this compact set.

Proof. Using Lemma 2.3 and the bound on the parameter-Jacobian from Lemma 2.1, as well as the bound on the initial function values from Lemma A.1:

$$\|\mathbf{f}_{\theta_t}\|_2 \leq \|\mathbf{f}_{\theta_t} - \mathbf{f}_{\theta_0}\|_2 + \|\mathbf{f}_{\theta_0}\|_2 \tag{117}$$

$$\leq \sup_{u \in [0,1]} \|\mathbf{J}_{\theta_0 + u(\theta_t - \theta_0)}\|_{2,2} \|\theta_t - \theta_0\|_2 + \sqrt{N} K'_0 \tag{118}$$

$$\leq \sqrt{N} K'_1 R + \sqrt{N} K'_0. \tag{119}$$

The bound for the linearized trajectories follows equivalently. □

Lemma D.4. *Let $\beta \geq 0$. Assume that equation (100) holds (lazy training). Then,*

$$\forall x \in M^d, t \geq 0 : \quad \left\| f_{\theta_t}(x) - f^{\text{lin}}_{\theta_t}(x) \right\|_2 \leq \frac{(\log n)^c}{\sqrt{n}} K'_2 R^2. \tag{120}$$

Proof. To bound $\left\| f_{\theta_t}(x) - f^{\text{lin}}_{\theta_t}(x) \right\|_2$, we use Lemma 2.1 and equation (100).

$$\left\| \frac{d}{dt} \left(f_{\theta_t}(x) - f^{\text{lin}}_{\theta_t}(x) \right) \right\|_2 = \left\| \left(J_{\theta_t}(x) - J_{\theta_0}(x) \right) \frac{d\theta_t}{dt} \right\|_2 \tag{121}$$

$$\leq \|J_{\theta_t}(x) - J_{\theta_0}(x)\|_2 \left\| \frac{d\theta_t}{dt} \right\|_2 \leq \frac{(\log n)^c}{\sqrt{n}} K'_2 R \eta_0 c_\beta R e^{-c_\beta \eta_0 t}. \tag{122}$$

By integrating it follows that

$$\left\| f_{\theta_t}(x) - f^{\text{lin}}_{\theta_t}(x) \right\|_2 \leq \frac{(\log n)^c}{\sqrt{n}} K'_2 R^2 \eta_0 c_\beta \int_0^t e^{-c_\beta \eta_0 u} du \leq \frac{(\log n)^c}{\sqrt{n}} K'_2 R^2. \tag{123}$$

□

Lemma D.5. *Let $\beta \geq 0$. Assume that equation (100) holds (lazy training). We can write*

$$\frac{d}{dt}(\theta_t - \theta_t^{\text{lin}}) = -\eta_0 \Delta_t - \eta_0 \left(\mathbf{J}_{\theta_0}^\top \overline{\mathbf{H}}_t \mathbf{J}_{\theta_0} + \beta I_p\right)(\theta_t - \theta_t^{\text{lin}}), \tag{124}$$

where we define the averaged function-space Gauss Newton

$$\overline{\mathbf{H}}_t := \int_0^1 \nabla^2_{\mathbf{zz}} \mathcal{C}\left(\mathbf{f}^{\text{lin}}_{\theta_t^{\text{lin}}} + s(\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta_t^{\text{lin}}})\right) ds \in \mathbb{R}^{NK \times NK}. \tag{125}$$

Moreover, the residual term $\Delta_t \in \mathbb{R}^p$ satisfies (for some $K^\Delta > 0$)

$$\|\Delta_t\|_2 \leq \frac{(\log n)^c}{\sqrt{n}} K^\Delta \quad \textit{for all } t \geq 0. \tag{126}$$

Proof. We can write

$$\begin{aligned}
\frac{d}{dt}\left(\theta_t - \theta_t^{\text{lin}}\right) =& -\eta_0 \left(\mathbf{J}_{\theta_t}^\top \nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}_{\theta_t}) - \mathbf{J}_{\theta_0}^\top \nabla_{\mathbf{z}} \mathcal{C}\left(\mathbf{f}^{\text{lin}}_{\theta_t^{\text{lin}}}\right) + \beta(\theta_t - \theta_t^{\text{lin}})\right) && (127)\\
=& -\eta_0 \left(\mathbf{J}_{\theta_t} - \mathbf{J}_{\theta_0}\right)^\top \nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}_{\theta_t}) && (128)\\
& - \eta_0 \mathbf{J}_{\theta_0}^\top \left(\nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}_{\theta_t}) - \nabla_{\mathbf{z}} \mathcal{C}\left(\mathbf{f}^{\text{lin}}_{\theta_t}\right)\right) && (129)\\
& - \eta_0 \mathbf{J}_{\theta_0}^\top \left(\nabla_{\mathbf{z}} \mathcal{C}\left(\mathbf{f}^{\text{lin}}_{\theta_t}\right) - \nabla_{\mathbf{z}} \mathcal{C}\left(\mathbf{f}^{\text{lin}}_{\theta_t^{\text{lin}}}\right)\right) && (130)\\
& - \eta_0 \beta(\theta_t - \theta_t^{\text{lin}}). && (131)
\end{aligned}$$

We summarize the first two terms (128) and (129) as

$$\Delta_t := \left(\mathbf{J}_{\theta_t} - \mathbf{J}_{\theta_0}\right)^\top \nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}_{\theta_t}) + \mathbf{J}_{\theta_0}^\top \left(\nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}_{\theta_t}) - \nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}^{\text{lin}}_{\theta_t})\right) \in \mathbb{R}^p. \tag{132}$$

Using the Fundamental Theorem of Calculus, we write the third term (130) as

$$\begin{aligned}
\mathbf{J}_{\theta_0}^\top \left(\nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}^{\text{lin}}_{\theta_t}) - \nabla_{\mathbf{z}} \mathcal{C}\left(\mathbf{f}^{\text{lin}}_{\theta_t^{\text{lin}}}\right)\right) &= \mathbf{J}_{\theta_0}^\top \int_0^1 \nabla^2_{\mathbf{zz}} \mathcal{C}\left(\mathbf{f}^{\text{lin}}_{\theta_t^{\text{lin}}} + s(\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta_t^{\text{lin}}})\right)\left(\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta_t^{\text{lin}}}\right) ds && (133)\\
&= \mathbf{J}_{\theta_0}^\top \int_0^1 \nabla^2_{\mathbf{zz}} \mathcal{C}\left(\mathbf{f}^{\text{lin}}_{\theta_t^{\text{lin}}} + s(\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta_t^{\text{lin}}})\right) ds \mathbf{J}_{\theta_0}(\theta_t - \theta_t^{\text{lin}}) && (134)\\
&= \mathbf{J}_{\theta_0}^\top \overline{\mathbf{H}}_t \mathbf{J}_{\theta_0}(\theta_t - \theta_t^{\text{lin}}). && (135)
\end{aligned}$$

To bound the first term (128) of Δ_t, we use Lemma D.3 and Lemma 2.1:

$$\left\|\left(\mathbf{J}_{\theta_t} - \mathbf{J}_{\theta_0}\right)^\top \nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}_{\theta_t})\right\|_2 \leq \|\nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}_{\theta_t})\|_2 \|\mathbf{J}_{\theta_t} - \mathbf{J}_{\theta_0}\|_{2,2} \leq L_{\mathcal{C}} \sqrt{N} \frac{(\log n)^c}{\sqrt{n}} K_2' R. \tag{136}$$

To bound the second term (129) of Δ_t, we use Lemma D.3, Lemma D.4, and Lemma 2.1:

$$\begin{aligned}
\left\|\mathbf{J}_{\theta_0}^\top \left(\nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}_{\theta_t}) - \nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}^{\text{lin}}_{\theta_t})\right)\right\|_2 &\leq \left\|\nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}_{\theta_t}) - \nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}^{\text{lin}}_{\theta_t})\right\|_2 \|\mathbf{J}_{\theta_0}\|_{2,2} && (137)\\
&\leq L_{\nabla \mathcal{C}} \left\|\mathbf{f}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta_t}\right\|_2 \sqrt{N} K_1' \leq \frac{(\log n)^c}{\sqrt{n}} L_{\nabla \mathcal{C}} K_1' K_2' N R^2. && (138)
\end{aligned}$$

Defining $K^\Delta := L_{\mathcal{C}} \sqrt{N} K_2' R + L_{\nabla \mathcal{C}} K_1' K_2' N R^2$ we get the $O((\log n)^c/\sqrt{n})$-bound on $\sup_{t \geq 0} \|\Delta_t\|_2$. □

Proof of Theorem D.2. Write

$$\left\|f_{\theta_t}(x) - f^{\text{lin}}_{\theta_t^{\text{lin}}}(x)\right\|_2 \leq \left\|f_{\theta_t}(x) - f^{\text{lin}}_{\theta_t}(x)\right\|_2 + \|J_{\theta_0}(x)\|_2 \left\|\theta_t - \theta_t^{\text{lin}}\right\|_2. \tag{139}$$

Lemma D.4 gives

$$\forall x \in M^d, t \geq 0: \quad \left\|f_{\theta_t}(x) - f^{\text{lin}}_{\theta_t}(x)\right\|_2 \leq \frac{(\log n)^c}{\sqrt{n}} K_2' R^2. \tag{140}$$

Using Lemma D.5 and $\lambda_{\min}\left(\mathbf{J}_{\theta_0}^\top \overline{\mathbf{H}}_t \mathbf{J}_{\theta_0} + \beta I_p\right) \geq \beta$,

$$\frac{d}{dt}\frac{1}{2}\|\theta_t - \theta_t^{\text{lin}}\|_2^2 = (\theta_t - \theta_t^{\text{lin}})^\top \frac{d}{dt}(\theta_t - \theta_t^{\text{lin}}) \tag{141}$$

$$= -\eta_0(\theta_t - \theta_t^{\text{lin}})^\top \Delta_t - \eta_0(\theta_t - \theta_t^{\text{lin}})^\top \left(\mathbf{J}_{\theta_0}^\top \overline{\mathbf{H}}_t \mathbf{J}_{\theta_0} + \beta I_p\right)(\theta_t - \theta_t^{\text{lin}}) \tag{142}$$

$$\leq \eta_0 \|\theta_t - \theta_t^{\text{lin}}\|_2 K^\Delta \frac{(\log n)^c}{\sqrt{n}} - \eta_0 \beta \|\theta_t - \theta_t^{\text{lin}}\|_2^2. \tag{143}$$

As we also have $\frac{d}{dt}\frac{1}{2}\left\|\theta_t - \theta_t^{\text{lin}}\right\|_2^2 = \left\|\theta_t - \theta_t^{\text{lin}}\right\|_2 \frac{d}{dt}\left\|\theta_t - \theta_t^{\text{lin}}\right\|_2$, we get by Gronwall's inequality,

$$\|\theta_t - \theta_t^{\text{lin}}\|_2 \leq \frac{(\log n)^c}{\sqrt{n}} \frac{K^\Delta}{\beta}\left(1 - e^{-\beta\eta_0 t}\right) \leq \frac{(\log n)^c}{\sqrt{n}} \frac{K^\Delta}{\beta}. \tag{144}$$

This concludes the proof with

$$C_2 := K_2' R^2 + K_1' \frac{K^\Delta}{\beta}. \tag{145}$$

□

E Proofs for zero parameter regularizer

E.1 Exponential decay of the loss-gradient norm

Theorem E.1. *Let $\beta = 0$. Consider a function-space loss that satisfies Assumptions 2 and 3. Then there are $R, c_0 > 0$ such that for large enough layer width n: With probability $1 - \delta_0$ over random initialization θ_0, for all $t \geq 0$,*

$$\left\|\frac{d}{dt}\theta_t\right\|_2 \leq \eta_0 c_0 R e^{-c_0\eta_0 t} \quad \textit{and thus} \quad \|\theta_t - \theta_0\|_2 \leq R. \tag{146}$$

The proof closely follows Oymak & Soltanolkotabi (2019); Liu et al. (2022), where it was presented for discrete-time gradient descent.

Proof. Recall that by Lemma 2.3 with high probability over random initialization θ_0,

$$\mathcal{C}(\mathbf{f}_{\theta_0}) \leq K_0. \tag{147}$$

Define $c_0 := \mu_{\mathcal{C}} \lambda_{\min}(\Theta_{\mathbf{x},\mathbf{x}})/2$ and $R := \frac{\sqrt{2K_2 N K_1'^2 (K_0 - \inf \mathcal{C})}}{c_0}$, using the constants $K_2, \mu_{\mathcal{C}}$ from Assumptions 2 and 3. Consider $t \geq 0$ as long as $\|\theta_t - \theta_0\|_2 \leq R$. Due to the convergence of the NTK at initialization, this implies that it is still sufficiently well conditioned at time t, i.e. there is n large enough (only depending on R), such that $\lambda_{\min}(\hat{\Theta}_{\theta_t}(\mathbf{x},\mathbf{x})) \geq \frac{1}{2}\lambda_{\min}(\Theta_{\mathbf{x},\mathbf{x}})$. Now, Assumption 3 implies that the loss gap decays exponentially quick during training:

$$\frac{d}{dt}\mathcal{C}(\mathbf{f}_{\theta_t}) = \nabla_\theta \mathcal{C}(\mathbf{f}_{\theta_t}) \frac{d}{dt}\theta_t = -\eta_0 \left\|\nabla_\theta \mathcal{C}(\mathbf{f}_{\theta_t})\right\|_2^2 \tag{148}$$

$$\leq -\eta_0 \lambda_{\min}(\hat{\Theta}_{\theta_t}(\mathbf{x},\mathbf{x})) \left\|\nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}_{\theta_t})\right\|_2^2 \tag{149}$$

$$\leq -\eta_0 \frac{1}{2}\lambda_{\min}(\Theta_{\mathbf{x},\mathbf{x}}) 2\mu_{\mathcal{C}} \left(\mathcal{C}(\mathbf{f}_{\theta_t}) - \inf \mathcal{C}\right), \tag{150}$$

and thus Gronwall's inequality,

$$\mathcal{C}(\mathbf{f}_{\theta_t}) - \inf \mathcal{C} \leq \left(\mathcal{C}(\mathbf{f}_{\theta_0}) - \inf \mathcal{C}\right) e^{-2c_0\eta_0 t} \leq (K_0 - \inf \mathcal{C}) e^{-2c_0\eta_0 t}. \tag{151}$$

Using Assumption 2, one can now use the exponential decay of the loss gap, to show that the norm of the loss-gradient decays exponentially quick:

$$\left\|\frac{d}{dt}\theta_t\right\|_2 = \eta_0 \left\|\nabla_\theta \mathcal{C}(\mathbf{f}_{\theta_t})\right\|_2 \leq \eta_0 \sqrt{N} K_1' \left\|\nabla_{\mathbf{z}} \mathcal{C}(\mathbf{f}_{\theta_t})\right\|_2 \tag{152}$$

$$\leq \eta_0 \sqrt{N} K_1' \sqrt{2K_2 \left(\mathcal{C}(\mathbf{f}_{\theta_t}) - \inf \mathcal{C}\right)} \tag{153}$$

$$\leq \eta_0 \sqrt{2N K_1'^2 K_2 (K_0 - \inf \mathcal{C})} e^{-c_0\eta_0 t}. \tag{154}$$

This also implies that

$$\|\theta_t - \theta_0\|_2 \leq \int_0^t \left\| \frac{d}{du}\theta_u \right\|_2 du \leq \eta_0 \sqrt{2NK_1'^2 K_2 (K_0 - \inf \mathcal{C})} \int_0^t e^{-c_0 \eta_0 u} du < R. \tag{155}$$

Thus the above holds for all $t \geq 0$. □

E.2 Closeness to linearization

Theorem E.2. *Let $\beta = 0$. Consider a function-space loss that satisfies Assumptions, 2, 3, and has strictly positive definite Hessian. Then there is a constant $C_2 > 0$ such that for large enough layer width n: With probability $1 - \delta_0$ over random initialization θ_0,*

$$\forall x \in M^d, t \geq 0 : \quad \left\| f_{\theta_t}(x) - f^{\text{lin}}_{\theta^{\text{lin}}_t}(x) \right\|_2 \leq C_2 \frac{(\log n)^c}{\sqrt{n}}. \tag{156}$$

Proof. We will follow the proof of Theorem D.2 in function-space instead of parameter-space, using Lemma D.4 and Lemma D.5 for $\beta = 0$. Write

$$\left\| f_{\theta_t}(x) - f^{\text{lin}}_{\theta^{\text{lin}}_t}(x) \right\|_2 \leq \left\| f_{\theta_t}(x) - f^{\text{lin}}_{\theta_t}(x) \right\|_2 + \left\| f^{\text{lin}}_{\theta_t}(x) - f^{\text{lin}}_{\theta^{\text{lin}}_t}(x) \right\|_2. \tag{157}$$

Lemma D.4 gives

$$\forall x \in M^d, t \geq 0 : \quad \left\| f_{\theta_t}(x) - f^{\text{lin}}_{\theta_t}(x) \right\|_2 \leq \frac{(\log n)^c}{\sqrt{n}} K_2' R^2. \tag{158}$$

Using Lemma D.5 gives (after multiplying both sides with $J_{\theta_0}(x)$ to get into function-space)

$$\frac{d}{dt}\left(f^{\text{lin}}_{\theta_t}(x) - f^{\text{lin}}_{\theta^{\text{lin}}_t}(x) \right) = -\eta_0 J_{\theta_0}(x) \Delta_t - \eta_0 J_{\theta_0}(x) \mathbf{J}_{\theta_0}^\top \overline{\mathbf{H}}_t (\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t}). \tag{159}$$

In particular, in the training points:

$$\frac{d}{dt}(\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t}) = -\eta_0 \mathbf{J}_{\theta_0} \Delta_t - \eta_0 \hat{\mathbf{\Theta}}_0 \overline{\mathbf{H}}_t (\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t}). \tag{160}$$

We will now give a lower bound on the lowest eigenvalue of $\overline{\mathbf{H}}_t$, to conclude with a bound on $\left\| \mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t} \right\|_2$ via Gronwall's inequality.

Due to $\|\theta_t - \theta_0\| \leq R$ (and $\left\|\theta^{\text{lin}}_t - \theta_0\right\| \leq R$, which can similarly be shown), the values $\mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t} + s\left(\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t}\right)$ are in a compact set: With high probability over random initialization θ_0,

$$\left\| \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t} + s(\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t}) \right\|_2 = \left\| \mathbf{f}_{\theta_0} + \mathbf{J}_{\theta_0} \left((1-s)(\theta^{\text{lin}}_t - \theta_0) + s(\theta_t - \theta_0) \right) \right\|_2 \tag{161}$$

$$\leq \|\mathbf{f}_{\theta_0}\|_2 + \|\mathbf{J}_{\theta_0}\|_{2,2} R \leq \sqrt{N}(K_0' + K_1' R). \tag{162}$$

Define

$$a_{\mathcal{C}} := \min_{\mathbf{z} \in B_{\sqrt{N}(K_0' + K_1' R)}(0)} \lambda_{\min}(\nabla^2_{\mathbf{zz}} \mathcal{C}(\mathbf{z})) > 0. \tag{163}$$

Then for any $\mathbf{z} \in \mathbb{R}^{NK}$ and $t \geq 0$,

$$\mathbf{z}^\top \overline{\mathbf{H}}_t \mathbf{z} = \int_0^1 \mathbf{z}^\top \nabla^2_{\mathbf{zz}} \mathcal{C}\left(\mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t} + s\left(\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t} \right) \right) \mathbf{z} ds \geq a_{\mathcal{C}} \|\mathbf{z}\|_2^2. \tag{164}$$

Now,

$$\frac{d}{dt}\frac{1}{2}\|\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t}\|^2_{\hat{\mathbf{\Theta}}_0^{-1}} = -\eta_0 \left(\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t} \right)^\top \hat{\mathbf{\Theta}}_0^{-1} \mathbf{J}_{\theta_0} \Delta_t \tag{165}$$

$$- \eta_0 \left(\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t} \right)^\top \hat{\mathbf{\Theta}}_0^{-1} \hat{\mathbf{\Theta}}_0 \overline{\mathbf{H}}_t (\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t}). \tag{166}$$

We can bound the first term using Cauchy-Schwarz and $\|\mathbf{J}_{\theta_0}\Delta_t\|_{\hat{\mathbf{\Theta}}_0^{-1}} \leq \|\Delta_t\|_2$:

$$-\eta_0 \left(\mathbf{f}_{\theta_t}^{\text{lin}} - \mathbf{f}_{\theta_t^{\text{lin}}}^{\text{lin}}\right)^\top \hat{\mathbf{\Theta}}_0^{-1}\mathbf{J}_{\theta_0}\Delta_t \leq \eta_0 \|\mathbf{f}_{\theta_t}^{\text{lin}} - \mathbf{f}_{\theta_t^{\text{lin}}}^{\text{lin}}\|_{\hat{\mathbf{\Theta}}_0^{-1}} \|\mathbf{J}_{\theta_0}\Delta_t\|_{\hat{\mathbf{\Theta}}_0^{-1}} \tag{167}$$

$$\leq \eta_0 \|\mathbf{f}_{\theta_t}^{\text{lin}} - \mathbf{f}_{\theta_t^{\text{lin}}}^{\text{lin}}\|_{\hat{\mathbf{\Theta}}_0^{-1}} K^\Delta \frac{(\log n)^c}{\sqrt{n}}. \tag{168}$$

For the second term, using $a_{\mathcal{C}} > 0$,

$$\left(\mathbf{f}_{\theta_t}^{\text{lin}} - \mathbf{f}_{\theta_t^{\text{lin}}}^{\text{lin}}\right)^\top \overline{\mathbf{H}}_t \left(\mathbf{f}_{\theta_t}^{\text{lin}} - \mathbf{f}_{\theta_t^{\text{lin}}}^{\text{lin}}\right) \geq a_{\mathcal{C}} \|\mathbf{f}_{\theta_t}^{\text{lin}} - \mathbf{f}_{\theta_t^{\text{lin}}}^{\text{lin}}\|_2^2 \geq a_{\mathcal{C}} \frac{\lambda_{\min}(\mathbf{\Theta})}{2} \|\mathbf{f}_{\theta_t}^{\text{lin}} - \mathbf{f}_{\theta_t^{\text{lin}}}^{\text{lin}}\|_{\hat{\mathbf{\Theta}}_0^{-1}}^2. \tag{169}$$

In total we get

$$\frac{d}{dt}\|\mathbf{f}_{\theta_t}^{\text{lin}} - \mathbf{f}_{\theta_t^{\text{lin}}}^{\text{lin}}\|_{\hat{\mathbf{\Theta}}_0^{-1}} \leq \eta_0 K^\Delta \frac{(\log n)^c}{\sqrt{n}} - \eta_0 a_{\mathcal{C}} \frac{\lambda_{\min}(\mathbf{\Theta})}{2} \|\mathbf{f}_{\theta_t}^{\text{lin}} - \mathbf{f}_{\theta_t^{\text{lin}}}^{\text{lin}}\|_{\hat{\mathbf{\Theta}}_0^{-1}}, \tag{170}$$

and thus by Gronwall's inequality,

$$\|\mathbf{f}_{\theta_t}^{\text{lin}} - \mathbf{f}_{\theta_t^{\text{lin}}}^{\text{lin}}\|_{\hat{\mathbf{\Theta}}_0^{-1}} \leq \frac{(\log n)^c}{\sqrt{n}} \frac{2K^\Delta}{a_{\mathcal{C}}\lambda_{\min}(\mathbf{\Theta})}(1 - e^{-\frac{1}{2}a_{\mathcal{C}}\lambda_{\min}(\mathbf{\Theta})\eta_0 t}). \tag{171}$$

Thus,

$$\|\mathbf{f}_{\theta_t}^{\text{lin}} - \mathbf{f}_{\theta_t^{\text{lin}}}^{\text{lin}}\|_2 \leq \sqrt{2\lambda_{\max}(\mathbf{\Theta})} \frac{(\log n)^c}{\sqrt{n}} \frac{2K^\Delta}{a_{\mathcal{C}}\lambda_{\min}(\mathbf{\Theta})}. \tag{172}$$

Now, in the test points: Multiplying the training-point dynamics by $\widehat{\Theta}_0(x, \mathbf{x})\widehat{\mathbf{\Theta}}_0^{-1}$ and subtracting them from the test-point dynamics gives

$$\frac{d}{dt}\left(f_{\theta_t}^{\text{lin}}(x) - f_{\theta_t^{\text{lin}}}^{\text{lin}}(x) - \widehat{\Theta}_0(x, \mathbf{x})\widehat{\mathbf{\Theta}}_0^{-1}\left(\mathbf{f}_{\theta_t}^{\text{lin}} - \mathbf{f}_{\theta_t^{\text{lin}}}^{\text{lin}}\right)\right) = -\eta_0 \left(J_{\theta_0}(x) - \widehat{\Theta}_0(x, \mathbf{x})\widehat{\mathbf{\Theta}}_0^{-1}\mathbf{J}_{\theta_0}\right)\Delta_t, \tag{173}$$

as the $\overline{\mathbf{H}}_t$-term cancels out due to

$$\widehat{\Theta}_0(x, \mathbf{x})\widehat{\mathbf{\Theta}}_0^{-1}\widehat{\mathbf{\Theta}}_0 = \widehat{\Theta}_0(x, \mathbf{x}). \tag{174}$$

Now, note that the second term (129) in the definition of Δ_t vanishes due to

$$\left(J_{\theta_0}(x) - \widehat{\Theta}_0(x, \mathbf{x})\widehat{\mathbf{\Theta}}_0^{-1}\mathbf{J}_{\theta_0}\right)\mathbf{J}_{\theta_0}^\top = \widehat{\Theta}_0(x, \mathbf{x}) - \widehat{\Theta}_0(x, \mathbf{x})\widehat{\mathbf{\Theta}}_0^{-1}\widehat{\mathbf{\Theta}}_0 = 0, \tag{175}$$

Therefore,

$$\left(J_{\theta_0}(x) - \widehat{\Theta}_0(x, \mathbf{x})\widehat{\mathbf{\Theta}}_0^{-1}\mathbf{J}_{\theta_0}\right)\Delta_t = \left(J_{\theta_0}(x) - \widehat{\Theta}_0(x, \mathbf{x})\widehat{\mathbf{\Theta}}_0^{-1}\mathbf{J}_{\theta_0}\right)\left(\mathbf{J}_{\theta_t} - \mathbf{J}_{\theta_0}\right)^\top \nabla_{\mathbf{z}}\mathcal{C}\left(\mathbf{f}_{\theta_t}\right). \tag{176}$$

The first factor is uniformly bounded over $x \in M^d$ by $K_1'\left(1 + \frac{2N(K_1')^2}{\lambda_{\min}(\mathbf{\Theta})}\right)$. For the second factor, by local Lipschitzness of the parameter Jacobian,

$$\|\mathbf{J}_{\theta_t} - \mathbf{J}_{\theta_0}\|_{2,2} \leq \sqrt{N}K_2' R \frac{(\log n)^c}{\sqrt{n}}.$$

For the third factor, we use Assumption 2, together with the exponential decay from Theorem E.1, to get

$$\|\nabla_{\mathbf{z}}\mathcal{C}\left(\mathbf{f}_{\theta_t}\right)\|_2^2 \leq 2K_2(K_0 - \inf \mathcal{C})e^{-2c_0\eta_0 t}.$$

Thus, defining $K^{\text{test}} := \frac{1}{c_0}K_1'\left(1 + \frac{2N(K_1')^2}{\lambda_{\min}(\mathbf{\Theta})}\right)\sqrt{N}K_2' R\sqrt{2K_2(K_0 - \inf \mathcal{C})}$,

$$\left\|\frac{d}{dt}\left(f_{\theta_t}^{\text{lin}}(x) - f_{\theta_t^{\text{lin}}}^{\text{lin}}(x) - \widehat{\Theta}_0(x, \mathbf{x})\widehat{\mathbf{\Theta}}_0^{-1}\left(\mathbf{f}_{\theta_t}^{\text{lin}} - \mathbf{f}_{\theta_t^{\text{lin}}}^{\text{lin}}\right)\right)\right\|_2 \leq \eta_0 c_0 K^{\text{test}} \frac{(\log n)^c}{\sqrt{n}} e^{-c_0\eta_0 t}. \tag{177}$$

At $t = 0$, the expression is zero. Integrating therefore gives, uniformly over $t \geq 0$ and $x \in M^d$,

$$\left\|f_{\theta_t}^{\text{lin}}(x) - f_{\theta_t^{\text{lin}}}^{\text{lin}}(x) - \widehat{\Theta}_0(x, \mathbf{x})\widehat{\mathbf{\Theta}}_0^{-1}\left(\mathbf{f}_{\theta_t}^{\text{lin}} - \mathbf{f}_{\theta_t^{\text{lin}}}^{\text{lin}}\right)\right\|_2 \leq K^{\text{test}} \frac{(\log n)^c}{\sqrt{n}}. \tag{178}$$

Using the triangle inequality and the bound on the training points, this finished the proof.

□

E.3 Extension for cross-entropy without a reference class

In this section we consider $\beta = 0$ and the standard cross-entropy loss without a reference class, $\mathcal{C} = \mathcal{C}_{\mathrm{CE}}$, when the target probabilities have full support. It satisfies Assumptions 1, 2 and 3 (see Appendix C.2). Thus it satisfies the lazy training property from Theorem E.1. However, as the function-space loss Hessian vanishes on the N-dimensional subspace $\mathrm{range}(I_N \otimes 1_K)$ we cannot directly apply Theorem E.2. We can however show that we are close to the linearized version when projecting both to a subspace along which the Hessian does have positive eigenvalues.

Define $U := I_N \otimes 1_K \in \mathbb{R}^{NK \times N}$ and the block-centering projection

$$\mathbf{P} := I_{NK} - U(U^\top U)^{-1}U^\top = I_N \otimes \left(I_K - \frac{1}{K} 1_K 1_K^\top\right) \in \mathbb{R}^{NK \times NK}. \tag{179}$$

This is a block diagonal whose N blocks are the projection $P := I_K - \frac{1}{K} 1_K 1_K^\top \in \mathbb{R}^{K \times K}$ onto $1^\perp$. For standard cross-entropy loss, $\mathcal{C}(\mathbf{z}) = \mathcal{C}(\mathbf{Pz})$, and thus $\nabla_{\mathbf{z}} \mathcal{C}(\mathbf{z}) = \mathbf{P} \nabla_{\mathbf{z}} \mathcal{C}(\mathbf{z}) = \nabla_{\mathbf{z}} \mathcal{C}(\mathbf{Pz})$, and $\nabla^2_{\mathbf{zz}} \mathcal{C}(\mathbf{z}) = \mathbf{P} \nabla^2_{\mathbf{zz}} \mathcal{C}(\mathbf{z}) \mathbf{P}$.

Theorem E.3. *Let $\beta = 0$. Consider the function-space cross-entropy loss without a reference class, and assume the target probabilities have full support, i.e. $\mathbf{p}_{ik} \in (0, 1)$. There is a constant $C_2 > 0$ such that for large enough layer width n: With probability $1 - \delta_0$ over random initialization θ_0,*

$$\forall x \in M^d, t \geq 0 : \quad \left\| P\left(f_{\theta_t}(x) - f^{\mathrm{lin}}_{\theta^{\mathrm{lin}}_t}(x)\right)\right\|_2 \leq C_2 \frac{(\log n)^c}{\sqrt{n}}. \tag{180}$$

Due to $\mathrm{smx}(z) = \mathrm{smx}(Pz)$, this implies that

$$\forall x \in M^d, t \geq 0 : \quad \left\| \mathrm{smx}\left(f_{\theta_t}(x)\right) - \mathrm{smx}\left(f^{\mathrm{lin}}_{\theta^{\mathrm{lin}}_t}(x)\right)\right\|_2 \leq \frac{1}{2} C_2 \frac{(\log n)^c}{\sqrt{n}}. \tag{181}$$

Proof. Intuitively the theorem follows from considering the projected network $\tilde{f} := Pf$, whose gradient flow is driven by the projected NTK $\tilde{\Theta} := P\Theta P$. This is not invertible, but has the pseudo inverse $\tilde{\Theta}^\dagger = \Theta^{-1}(I_{NK} - U(U^\top \Theta^{-1} U)^{-1} U^\top \Theta^{-1})$. For completeness, we present the entire proof. As in the proof of Theorem E.2, use the triangle inequality to write

$$\left\| P\left(f_{\theta_t}(x) - f^{\mathrm{lin}}_{\theta^{\mathrm{lin}}_t}(x)\right)\right\|_2 \leq \left\| P\left(f_{\theta_t}(x) - f^{\mathrm{lin}}_{\theta_t}(x)\right)\right\|_2 + \left\| P\left(f^{\mathrm{lin}}_{\theta_t}(x) - f^{\mathrm{lin}}_{\theta^{\mathrm{lin}}_t}(x)\right)\right\|_2 \tag{182}$$

By Lemma D.4 we can bound the first term:

$$\left\| P\left(f_{\theta_t}(x) - f^{\mathrm{lin}}_{\theta_t}(x)\right)\right\|_2 \leq \left\| f_{\theta_t}(x) - f^{\mathrm{lin}}_{\theta_t}(x)\right\|_2 \leq \frac{(\log n)^c}{\sqrt{n}} K_2' R^2. \tag{183}$$

Using Lemma D.5 gives (after multiplying both sides with $J_{\theta_0}(x)$ to get into function-space)

$$\frac{d}{dt}\left(f^{\mathrm{lin}}_{\theta_t}(x) - f^{\mathrm{lin}}_{\theta^{\mathrm{lin}}_t}(x)\right) = -\eta_0 J_{\theta_0}(x) \Delta_t - \eta_0 J_{\theta_0}(x) \mathbf{J}_{\theta_0}^\top \overline{\mathbf{H}}_t (\mathbf{f}^{\mathrm{lin}}_{\theta_t} - \mathbf{f}^{\mathrm{lin}}_{\theta^{\mathrm{lin}}_t}). \tag{184}$$

In particular, in the training points:

$$\frac{d}{dt}(\mathbf{f}^{\mathrm{lin}}_{\theta_t} - \mathbf{f}^{\mathrm{lin}}_{\theta^{\mathrm{lin}}_t}) = -\eta_0 \mathbf{J}_{\theta_0} \Delta_t - \eta_0 \hat{\mathbf{\Theta}}_0 \overline{\mathbf{H}}_t (\mathbf{f}^{\mathrm{lin}}_{\theta_t} - \mathbf{f}^{\mathrm{lin}}_{\theta^{\mathrm{lin}}_t}). \tag{185}$$

Define the oblique projection

$$\hat{\mathbf{\Pi}} := I_{NK} - U(U^\top \hat{\mathbf{\Theta}}^{-1} U)^{-1} U^\top \hat{\mathbf{\Theta}}^{-1} \in \mathbb{R}^{NK \times NK}. \tag{186}$$

We get:

$$\frac{d}{dt} \frac{1}{2} \|\hat{\mathbf{\Pi}}(\mathbf{f}^{\mathrm{lin}}_{\theta_t} - \mathbf{f}^{\mathrm{lin}}_{\theta^{\mathrm{lin}}_t})\|^2_{\hat{\mathbf{\Theta}}_0^{-1}} = -\eta_0 \left(\hat{\mathbf{\Pi}}(\mathbf{f}^{\mathrm{lin}}_{\theta_t} - \mathbf{f}^{\mathrm{lin}}_{\theta^{\mathrm{lin}}_t})\right)^\top \hat{\mathbf{\Theta}}_0^{-1} \hat{\mathbf{\Pi}} \mathbf{J}_{\theta_0} \Delta_t \tag{187}$$

$$- \eta_0 \left(\hat{\mathbf{\Pi}}(\mathbf{f}^{\mathrm{lin}}_{\theta_t} - \mathbf{f}^{\mathrm{lin}}_{\theta^{\mathrm{lin}}_t})\right)^\top \hat{\mathbf{\Theta}}_0^{-1} \hat{\mathbf{\Pi}} \hat{\mathbf{\Theta}}_0 \overline{\mathbf{H}}_t (\mathbf{f}^{\mathrm{lin}}_{\theta_t} - \mathbf{f}^{\mathrm{lin}}_{\theta^{\mathrm{lin}}_t}). \tag{188}$$

We can bound the first term using Cauchy-Schwarz:

$$-\eta_0 \left(\hat{\mathbf{\Pi}}(\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t})\right)^\top \hat{\mathbf{\Theta}}_0^{-1}\hat{\mathbf{\Pi}}\mathbf{J}_{\theta_0}\Delta_t \le \eta_0 \|\hat{\mathbf{\Pi}}(\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t})\|_{\hat{\mathbf{\Theta}}_0^{-1}} \|\hat{\mathbf{\Pi}}\mathbf{J}_{\theta_0}\Delta_t\|_{\hat{\mathbf{\Theta}}_0^{-1}} \tag{189}$$

$$\le \eta_0 \|\hat{\mathbf{\Pi}}(\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t})\|_{\hat{\mathbf{\Theta}}_0^{-1}} K^\Delta \frac{(\log n)^c}{\sqrt{n}}. \tag{190}$$

For the second term, we note that by Lemma D.3 the trajectories of $\mathbf{f}^{\text{lin}}_{\theta_t}$ and $\mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t}$ are in a compact set (with high probability over initialization). Define $a^+_{\mathcal{C}} > 0$ as the minimum of the minimum eigenvalue of $\nabla^2_{\mathbf{zz}}\mathcal{C}(\mathbf{z})$ on $\text{range}(U)^\perp$ for $\mathbf{z}$ in this compact set. We use $\hat{\mathbf{\Theta}}_0^{-1}\hat{\mathbf{\Pi}}\hat{\mathbf{\Theta}}_0 = \hat{\mathbf{\Pi}}^\top$, $\overline{H}_t = \mathbf{P}\overline{H}_t\mathbf{P}$, and $\mathbf{P}\hat{\mathbf{\Pi}} = \mathbf{P}$ to write

$$\left(\hat{\mathbf{\Pi}}(\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t})\right)^\top \hat{\mathbf{\Theta}}_0^{-1}\hat{\mathbf{\Pi}}\hat{\mathbf{\Theta}}_0\overline{\mathbf{H}}_t(\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t}) = \left(\mathbf{P}(\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t})\right)^\top \overline{\mathbf{H}}_t \left(\mathbf{P}(\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t})\right) \tag{191}$$

$$\ge a^+_{\mathcal{C}} \|\mathbf{P}(\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t})\|_2^2 \tag{192}$$

$$\ge a^+_{\mathcal{C}} \frac{\lambda_{\min}(\mathbf{\Theta})}{2} \|\mathbf{P}(\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t})\|^2_{\hat{\mathbf{\Theta}}_0^{-1}} \tag{193}$$

$$\ge a^+_{\mathcal{C}} \frac{\lambda_{\min}(\mathbf{\Theta})}{2} \|\hat{\mathbf{\Pi}}(\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t})\|^2_{\hat{\mathbf{\Theta}}_0^{-1}}. \tag{194}$$

In total we get

$$\frac{d}{dt}\|\hat{\mathbf{\Pi}}(\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t})\|_{\hat{\mathbf{\Theta}}_0^{-1}} \le \eta_0 K^\Delta \frac{(\log n)^c}{\sqrt{n}} - \eta_0 a^+_{\mathcal{C}} \frac{\lambda_{\min}(\mathbf{\Theta})}{2} \|\hat{\mathbf{\Pi}}(\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t})\|_{\hat{\mathbf{\Theta}}_0^{-1}}, \tag{195}$$

and thus by Gronwall's inequality,

$$\|\hat{\mathbf{\Pi}}(\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t})\|_{\hat{\mathbf{\Theta}}_0^{-1}} \le \frac{(\log n)^c}{\sqrt{n}} \frac{2K^\Delta}{a^+_{\mathcal{C}}\lambda_{\min}(\mathbf{\Theta})}(1 - e^{-\frac{1}{2}a^+_{\mathcal{C}}\lambda_{\min}(\mathbf{\Theta})\eta_0 t}). \tag{196}$$

Thus, using $\mathbf{P} = \mathbf{P}\hat{\mathbf{\Pi}}$

$$\|\mathbf{P}(\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t})\|_2 \le \|\hat{\mathbf{\Pi}}(\mathbf{f}^{\text{lin}}_{\theta_t} - \mathbf{f}^{\text{lin}}_{\theta^{\text{lin}}_t})\|_2 \le \sqrt{2\lambda_{\max}(\mathbf{\Theta})} \frac{(\log n)^c}{\sqrt{n}} \frac{2K^\Delta}{a^+_{\mathcal{C}}\lambda_{\min}(\mathbf{\Theta})}. \tag{197}$$

The bound for arbitrary test points follows as in the previous section. □

F Proof of lemmas for the linearized dynamics

Lemma 3.3. *Let $\beta \ge 0$ and assume $\Theta_{\mathbf{x},\mathbf{x}}$ is invertible. Then the prediction of the linearized network at any test point x is*

$$g_\infty(x) = f_{\theta_0}(x) + \Theta_{x,\mathbf{x}}\Theta^{-1}_{\mathbf{x},\mathbf{x}} \left(\underbrace{\Phi^{-1}(\beta f_{\theta_0}(\mathbf{x}))}_{=g_\infty(\mathbf{x})} - f_{\theta_0}(\mathbf{x}) \right). \tag{26}$$

Moreover for $\beta > 0$ the prediction can also be written as

$$g_\infty(x) = f_{\theta_0}(x) - \frac{1}{\beta}\Theta_{x,\mathbf{x}}\nabla_{\mathbf{z}}\mathcal{C}\left(\Phi^{-1}(\beta f_{\theta_0}(\mathbf{x}))\right). \tag{27}$$

Proof of Lemma 3.3. In the training points

$$\frac{d}{dt}\mathbf{g}_t = -\eta_0 \left(\Theta_{\mathbf{x},\mathbf{x}}\nabla_{\mathbf{z}}\mathcal{C}(\mathbf{g}_t) + \beta(\mathbf{g}_t - \mathbf{g}_0)\right). \tag{198}$$

Thus,

$$-\eta_0 \nabla_z \mathcal{C}(\mathbf{g}_t) = \Theta^{-1}_{\mathbf{x},\mathbf{x}} \left(\frac{d}{dt}\mathbf{g}_t + \eta_0\beta(\mathbf{g}_t - \mathbf{g}_0) \right). \tag{199}$$

Plugging this in for any test point,

$$\frac{d}{dt}(g_t(x) - g_0(x)) = \Theta_{x,\mathbf{x}}\Theta_{\mathbf{x},\mathbf{x}}^{-1}\left(\frac{d}{dt}\mathbf{g}_t + \eta_0\beta(\mathbf{g}_t - \mathbf{g}_0)\right) - \eta_0\beta(g_t(x) - g_0(x)). \tag{200}$$

Hence,

$$\frac{d}{dt}\left(e^{\eta_0\beta t}(g_t(x) - g_0(x))\right) = \Theta_{x,\mathbf{x}}\Theta_{\mathbf{x},\mathbf{x}}^{-1}\frac{d}{dt}\left(e^{\eta_0\beta t}(\mathbf{g}_t - \mathbf{g}_0)\right). \tag{201}$$

Integrating gives

$$g_t(x) - g_0(x) = \Theta_{x,\mathbf{x}}\Theta_{\mathbf{x},\mathbf{x}}^{-1}(\mathbf{g}_t - \mathbf{g}_0). \tag{202}$$

Thus, using $\mathbf{g}_\infty = \Phi^{-1}(\beta f_{\theta_0}(\mathbf{x}))$ gives the first expression.

Moreover, consider $\beta > 0$. Then, rearranging the function-space stationary point equation

$$\Theta_{x,\mathbf{x}}\nabla_{\mathbf{z}}\mathcal{C}\left(\mathbf{g}_\infty\right) + \beta g_\infty(x) = \beta f_{\theta_0}(x), \tag{203}$$

gives

$$f_{\theta_\infty^{\text{lin}}}^{\text{lin}}(x) = f_{\theta_0}(x) - \frac{1}{\beta}\Theta_{x,\mathbf{x}}\nabla_{\mathbf{z}}\mathcal{C}\left(\mathbf{g}_\infty\right). \tag{204}$$

The second expression follows by again plugging in $\mathbf{g}_\infty = \Phi^{-1}(\beta f_{\theta_0}(\mathbf{x}))$. □