Just one LayerNorm

A Preprint

Yet, it is difficult to say how the network will behave outside of the training domain, as the training process does not necessarily constrains the network's output there. One of the most complete theoretical approaches to studying the behaviour of Deep Neural Networks is the Neural Tangent Kernel (NTK) theory. Studies show that as the network width approaches infinity, the trained network's output is equivalent to the posterior mean of a Gaussian Process.

We consider a standard supervised learning setting with a dataset $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^n$, where each input $x_i \in \mathbb{R}^{n_0}$ and corresponding target $y_i \in \mathbb{R}$. We assume that the datapoints are distinct, i.e., $x_i \neq x_j$ for $i \neq j$. Let $X_{\text{train}} = \{x_i\}_{i=1}^n$ and $Y_{\text{train}} = \{y_i\}_{i=1}^n$.

Our goal is to train a neural network $f_\theta : \mathbb{R}^{n_0} \to \mathbb{R}$ with parameters $\theta \in \Omega \subset \mathbb{R}^d$ by minimizing the mean squared error (MSE) loss:

$$\mathcal{L}(\theta; \mathcal{D}_{\text{train}}) = \frac{1}{2}\|f_\theta(x_i) - y_i\|_2^2.$$

Parameters are initialized as described in Appendix A and updated via gradient descent:

$$\theta^{(t)} = \theta^{(t-1)} - \eta \nabla_\theta \mathcal{L}(\theta; \mathcal{D}_{\text{train}}),$$

with learning rate $\eta > 0$.

It has been observed in prior work that, even for finite-width networks, the training dynamics approximate a kernel gradient descent process governed by a data-dependent kernel. Kernel:

$$\Theta(x, x^{'}; \theta) =< \nabla_\theta f_\theta(x), \nabla_\theta f_\theta(x^{'}) >$$

also known as empirical NTK. More over, if learning rate is small enough network converges to

$$f_{\theta_\infty}(x^\star) = \Theta(x^\star, X_{\text{train}})\, \Theta(X_{\text{train}}, X_{\text{train}})^{-1} \left(Y_{\text{train}} - f_{\theta_0}(X_{\text{train}})\right) + f_{\theta_0}(x^\star),$$

Under standard parameter initialization, the network output at initialization satisfies $\mathbb{E}[f_{\theta_0}(x^\star)] = 0$. Substituting into the kernel regression expression, we obtain the expected prediction at convergence:

$$\mathbb{E}[f_{\theta_\infty}(x^\star)] = \Theta(x^\star, X_{\text{train}})\, \Theta(X_{\text{train}}, X_{\text{train}})^{-1} Y_{\text{train}}.$$

This is exactly the posterior mean of a Gaussian Process (GP) with kernel $\Theta(x, x')$ trained on $\mathcal{D}_{\text{train}}$ and evaluated at $x^\star$. Hence, gradient descent in a neural network initialized appropriately can be interpreted as performing kernel regression in a GP framework.

Very important assumptions:

- Activation functions act element-wise and almost everywhere differentiable
- activation function $\phi : \mathbb{R} \to \mathbb{R}$ is positive n-homogeneous for some $n > \frac{1}{2}$; that is $\phi(\lambda x) = \lambda^n \phi(x)$ for all $\lambda > 0$. This class includes common nonlinearities such as ReLU (with $n = 1$)
- Gram matrix over the training set, denoted Θ_{train}, is positive definite. In particular, its minimum eigenvalue satisfies $\lambda_{\min} > 0$. This condition ensures that Θ_{train} is non-singular and thus invertible, since a symmetric matrix is invertible if and only if it is positive definite—i.e., all of its eigenvalues are strictly positive.

For NN with nonlinearities satisfying assumptions and FC:

$$\sup_{x \in \mathbb{R}^{n_0}} |\mathbb{E}[f_\theta(x)]| = \infty.$$

In the framework of kernel methods, the functions learnable by the neural network correspond to elements of the reproducing kernel Hilbert space (RKHS) induced by the NTK. The the RKHS associated with the NTK consists only of functions that grow without bound. This implies that the network, in the infinite-width or kernel regime, is biased toward learning functions with unbounded growth, which may affect generalization unless additional constraints (e.g., normalization) are introduced.

Inclusion of LayerNorm

For a neural network $f_\theta(x)$ composed of fully connected layers and nonlinearities satisfying assumptions, the inclusion of at least one Layer Normalization (LayerNorm) layer guarantees that the Neural Tangent Kernel (NTK) remains pointwise bounded. Specifically, there exists a constant $C > 0$ such that:

$$\forall x \in \mathbb{R}^{n_0}, \quad \Theta(x, x) \leq C.$$

Authors prove that inclusion of a LayerNorm causes the variance NTK to take the form of a ratio of two (strictly positive) functions of input norm.Under the assumptions of bounded NTK diagonal and positive definiteness of the kernel matrix Θ_{train}, the expected output of the network at any input $x \in \mathbb{R}^{n_0}$ is bounded as follows:

$$|\mathbb{E}[f_\theta(x)]| \leq B(\mathcal{D}_{\text{train}}) = s \cdot C \cdot \frac{\max\limits_{y \in Y_{\text{train}}} |y|}{\lambda_{\min} \cdot |\mathcal{D}_{\text{train}}|},$$

where the expectation is over the random initialization of parameters, $\lambda_{\min}$ is the smallest eigenvalue of Θ_{train}, C is a constant depending on the kernel Θ, and s is a scaling factor from initialization or architecture.

This inequality highlights how the expected output remains controlled in the presence of normalization and well-conditioned kernel matrices, even as the network size grows.

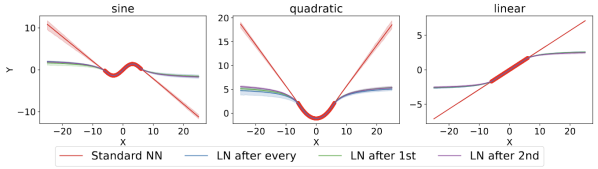

Figure 1: Predictions made by networks with various architectures when trained on synthetic datasets. Red dot show the train set datapoints. The solid lines indicate average values over 5 seeds and shaded areas are 95% confidence intervals of the mean estimator.

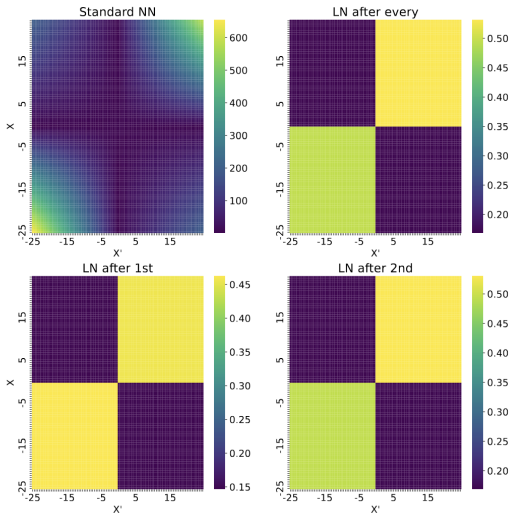

Figure 2: Heatmaps showing the values of empirical NTK values $\Theta(x, x')$ plotted on domain $x, x' \in [-25, 25]$ with brighter colours indicating higher values. Note that the scales are different for each heatplot, with the values range for the NTK of Standard NN being orders of magnitude higher than others. The displayed values are averages over 5 seeds.

The UCI Protein Tertiary Structure dataset, comprising 45,730 samples with nine physicochemical features, is used to assess the impact of LayerNorm on out-of-distribution (OOD) generalization. The task is to predict the RMSD of atomic distances.

Models are trained on 90% of proteins with surface area below 20,000 Å^2. The remaining small-area proteins form the in-distribution (ID) set; proteins with larger surface area constitute the OOD set.

Fully connected networks with two hidden layers (width 128) are tested with and without LayerNorm (LN). Variants include LN before each hidden layer or before selected layers. XGBoost is included as a non-neural baseline.

Results show similar ID performance across models. However, LN-equipped networks significantly outperform both the standard neural net and XGBoost in the OOD setting. The exact placement of LN has little effect—R^2 scores remain within overlapping confidence intervals. Histograms of OOD predictions indicate that models without LN produce heavier-tailed, more extreme outputs. Error increases with surface area for all models, but most severely without normalization.

Overall, LayerNorm improves stability and accuracy under distributional shift, likely due to bounded NTK behavior and better-controlled output scaling.

A second real-world experiment examines age prediction from facial images using the UTKFace dataset [?], which contains over 20,000 cropped and aligned faces labeled with age, gender, and ethnicity. Ages range from 0 to 116 years, and ethnicity includes five categories: White, Black, Asian, Indian, and Others.

To evaluate robustness under demographic shift, models are trained on images from four ethnicities (White, Black, Asian, Indian). Validation is split into an in-distribution (ID) set—samples from these same groups—and an out-of-distribution (OOD) set containing the remaining ethnicity category.

All models use a frozen ResNet-16 backbone followed by two fully connected layers (size 128, ReLU). Variants differ in the application of LayerNorm: no LN, LN after each layer, or LN after only the first or second layer.

Results indicate that, as in previous experiments, LayerNorm improves OOD generalization. All LN variants outperform the unnormalized model under distributional shift, with minimal sensitivity to exact LN placement.