On the Equivalence between Implicit and Explicit Neural Networks: A High-dimensional Viewpoint

Zenan Ling   LINGZENAN@HUST.EDU.CN
Zhenyu Liao   ZHENYU_LIAO@HUST.EDU.CN
Robert C. Qiu   CAIMING@HUST.EDU.CN
Huazhong University of Science and Technology, Wuhan, China

Abstract

Implicit neural networks have demonstrated remarkable success in various tasks. However, there is a lack of theoretical analysis of the connections and differences between implicit and explicit networks. In this paper, we study high-dimensional implicit neural networks and provide the high dimensional equivalents for the corresponding conjugate kernels and neural tangent kernels. Built upon this, we establish the equivalence between implicit and explicit networks in high dimensions.

1. Introduction

Implicit neural networks (NNs) [2] have recently emerged as a new paradigm in neural network design. An implicit NN is equivalent to an infinite-depth weight-shared explicit NN with input-injection. Unlike explicit NNs, implicit NNs generate features by directly solving for the fixed point, rather than through layer-by-layer forward propagation. Moreover, implicit NNs have the remarkable advantage that gradients can be computed analytically only through the fixed point with *implicit differentiation*. Therefore, training implicit NNs only requires constant memory.

Despite the empirical success achieved by implicit NNs [3, 11], our theoretical understanding of these models is still limited. In particular, there is a lack of theoretical analysis of the training dynamics and generalization performance of implicit NNs, and possibly more importantly, whether these properties can be connected to those of explicit NNs. [2] demonstrates that any deep NN can be reformulated as a special implicit NN. However, it remains unknown whether general implicit NNs have advantages over explicit NNs. [6] extends previous neural tangent kernel (NTK) studies to implicit NNs and give the exact expression of the NTK of the ReLU implicit NNs. However, the differences between implicit and explicit NTKs are not analyzed. Moreover, previous works [9, 10] have proved the global convergence of gradient descent for training implicit NNs. However, it is still unclear what distinguishes the training dynamic of implicit NNs and that of explicit NNs.

In this paper, we investigate implicit NNs from a high-dimensional view. Specifically, we perform a fine-grained asymptotic analysis on the eigenspectra of conjugate kernel (CKs) and NTKs of implicit NNs, which play a fundamental role in the convergence and generalization high dimensional NNs [8]. By considering input data uniformly drawn from the unit sphere, we derive, with recent advances in random matrix theory, high-dimensional (spectral) equivalents for the CKs and NTKs of implicit NNs, and establish the equivalence between implicit and explicit NNs by matching the coefficients of the corresponding asymptotic spectral equivalents. Surprisingly, our results reveal that a *single-layer* explicit NN with carefully designed activations has the same CK or NTK eigenspectra as a ReLU implicit NN, whose depth is essentially *infinite*.

2. Preliminaries

2.1. Implicit and Explicit NNs

Implicit NNs. In this paper, we study a typical implicit neural network, the deep equilibrium model (DEQ) [2]. Let $X = [x_1, \cdots, x_n] \in \mathbb{R}^{d \times n}$ denote the input data. We define a vanilla DEQ with the transform at the l-th layer as

$$h_i^{(l)} = \sqrt{\frac{\sigma_a^2}{m}} A z_i^{(l-1)} + + \sqrt{\frac{\sigma_b^2}{m}} B x_i, \quad z_i^{(l)} = \phi(h_i^{(l)}) \tag{1}$$

where $A \in \mathbb{R}^{m \times m}$ and $B \in \mathbb{R}^{m \times d}$ are weight matrices, $\sigma_a, \sigma_b \in \mathbb{R}$ are constants, ϕ is an element-wise activation, $h_i^{(l)}$ is the pre-activation and $z_i^{(l)} \in \mathbb{R}^m$ is the output feature of the l-th hidden layer corresponding to the input data x_i. The output of the last hidden layer is defined by $z_i^* \triangleq \lim_{l \to \infty} z_i^{(l)}$ and we denote the corresponding pre-activation by h_i^*. Note that z_i^* can be calculated by directly solving for the equilibrium point of the following equation

$$z_i^* = \phi\left(\sqrt{\frac{\sigma_a^2}{m}} A z_i^* + \sqrt{\frac{\sigma_b^2}{m}} B x_i\right). \tag{2}$$

We are interested in the conjugate kernel and neural tangent kernel (Implicit-CK and Implicit-NTK, for short) of implicit neural networks defined in Eq. (2). Following [6], we denote the corresponding Implicit-CK by $G^* = \lim_{l \to \infty} G^{(l)}$ where the (i, j)-th entry of $G^{(l)}$ is defined recursively as

$$G_{ij}^{(0)} = x_i^\top x_j, \quad \Lambda_{ij}^{(l)} = \begin{bmatrix} G_{ii}^{(l-1)} & G_{ij}^{(l-1)} \\ G_{jj}^{(l-1)} & G_{jj}^{(l-1)} \end{bmatrix},$$

$$G_{ij}^{(l)} = \sigma_a^2 \mathbb{E}_{(\mathsf{u},\mathsf{v}) \sim \mathcal{N}(0, \Lambda_{ij}^{(l)})}[\phi(\mathsf{u})\phi(\mathsf{v})] + \sigma_b^2 x_i^\top x_j, \tag{3}$$

$$\dot{G}_{ij}^{(l)} = \sigma_a^2 \mathbb{E}_{(\mathsf{u},\mathsf{v}) \sim \mathcal{N}(0, \Lambda_{ij}^{(l)})}[\phi'(\mathsf{u})\phi'(\mathsf{v})].$$

And the Implicit-NTK is defined as $K^* = \lim_{l \to \infty} K^{(l)}$ whose the (i, j)-th entry is defined as

$$K_{ij}^{(l)} = \sum_{h=1}^{l+1} \left(G_{ij}^{(h-1)} \prod_{h'=h}^{l+1} \dot{G}_{ij}^{(h')}\right). \tag{4}$$

Explicit Neural Networks. We consider a single-layer fully-connected NN model defined as $Y = \sqrt{\frac{1}{p}}\sigma(WX)$ where $W \in \mathbb{R}^{p \times d}$ is the weight matrix and σ is an element-wise activation function. Let $w \sim \mathcal{N}(0, I_d)$, the corresponding Explicit-CK matrix Σ and Explicit-NTK matrix Θ are defined as follows:

$$\Sigma = \mathbb{E}_w[\sigma(w^\top X)^\top \sigma(w^\top X)], \quad \Theta = \Sigma + \left(X^\top X\right) \odot \mathbb{E}_w[\sigma'(w^\top X)^\top \sigma'(w^\top X)]. \tag{5}$$

2.2. CKs and NTKs of ReLU Implicit NNs

We make the following assumptions on the random initialization, the input data, and activations.

2

Assumption 1 *(i) As $n \to \infty$, $d/n \to c \in (0, \infty)$. All data points $\boldsymbol{x}_i$, $i \in [n]$, are independent and uniformly sampled from $\mathbb{S}^{d-1}$. (ii) $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{W}$ are independent and have i.i.d entries of zero mean, unit variance, and finite fourth kurtosis. Moreover, we require $\sigma_a^2 + \sigma_b^2 = 1$. (iii) The activation ϕ of the implicit NN is the normalized ReLU, i.e., $\phi(x) = \sqrt{2}\max(x, 0)$. The activation σ of the explicit NN is a C^3 function.*

Remark 1 *(i) Despite derived here for uniform distribution on the unit sphere, we conjecture that our results extend the result to more general distributions by using the technique developed in [5, 7]. (ii) The additional requirement on the variance is to ensure the existence and uniqueness of the fixed point of the NTK and to keep the diagonal entries of the CK matrix at 1, see examples in [6]. (iii) It is possible to extend our results to implicit NNs with general activations by using the technique proposed in [10]. We defer the extension to more general data distributions and activation functions to future work.*

Under Assumptions 1, the limits of Implicit-CK and Implicit-NTK exist, and one can have precise expressions of $\boldsymbol{G}^*$ and $\boldsymbol{K}^*$ as follows [6, 9].

Lemma 1 *Let $f(x) = \frac{\sqrt{1-x^2}+(\pi-\arccos x)x}{\pi}$. Under Assumptions 1, the fixed point of Implicit-CK $\boldsymbol{G}_{ij}^*$ is the root of*

$$\boldsymbol{G}_{ij}^* = \sigma_a^2 f(\boldsymbol{G}_{ij}^*) + (1 - \sigma_a^2)\boldsymbol{x}_i^\top \boldsymbol{x}_j. \tag{6}$$

The limit of Implicit-NTK is

$$\boldsymbol{K}_{ij}^* = h(\boldsymbol{G}_{ij}^*) \triangleq \frac{\boldsymbol{G}_{ij}^*}{1 - \dot{\boldsymbol{G}}_{ij}^*} \quad \text{where} \quad \dot{\boldsymbol{G}}_{ij}^* \triangleq \sigma_a^2 \pi^{-1}(\pi - \arccos(\boldsymbol{G}_{ij}^*)). \tag{7}$$

3. Main Results

In this section, we prove the high-dimensional equivalents for CKs and NTKs of implicit and explicit NNs. As a result, by matching the coefficients of the asymptotic spectral equivalents, we establish the equivalence between implicit and explicit NNs in high dimensions.

3.1. Asymptotic Approximations

CKs. We begin by defining several quantities that are crucial to our results. Note that the unique fixed point of Eq. (6) exists as long as $\sigma_a^2 < 1$. We define the implicit map induced from Eq. (6) as $\boldsymbol{G}_{ij}^* \triangleq g(\boldsymbol{x}_i^\top \boldsymbol{x}_j)$. Let $\angle^* = g(0)$ be the solution of $\angle^* = \sigma_a^2 f(\angle^*)$ when $\boldsymbol{x}_i^\top \boldsymbol{x}_j = 0$. Using implicit differentiation, one can obtain that

$$g'(0) = \frac{1 - \sigma_a^2}{1 - \sigma_a^2 f'(\angle^*)}, \quad g''(0) = \frac{\sigma_a^2(1 - \sigma_a^2)^2 f''(\angle^*)}{(1 - \sigma_a^2 f'(\angle^*))^3}.$$

Now we are ready to present the asymptotic equivalent of the Implicit-CK matrix.

Theorem 1 (Asymptotic approximation of Implicit-CKs) *Let Assumptions 1 hold. As $n, d \to \infty$, the Implicit-CK matrix $\boldsymbol{G}^*$ defined in Eq. (6) can be approximated consistently in operator norm, by the matrix $\overline{\boldsymbol{G}}$, that is $\|\boldsymbol{G}^* - \overline{\boldsymbol{G}}\|_2 \to 0$, where*

$$\overline{\boldsymbol{G}} = \alpha \mathbf{1}\mathbf{1}^\top + \beta \boldsymbol{X}^\top \boldsymbol{X} + \mu \boldsymbol{I}_n,$$

with $\alpha = g(0) + \frac{g''(0)}{2d}$, $\beta = g'(0)$, and $\mu = g(1) - g(0) - g'(0)$.

Theorem 2 (Asymptotic approximation for Explicit-CKs) *Let Assumptions 1 hold. As $n, d \rightarrow \infty$, the Explicit-CK matrix Σ defined in Eq. (5) can be approximated consistently in operator norm, by the matrix $\overline{\Sigma}$, that is $\|\Sigma - \overline{\Sigma}\|_2 \rightarrow 0$, where*

$$\overline{\Sigma} = \alpha_1 \mathbf{1}\mathbf{1}^\top + \beta_1 \mathbf{X}^\top \mathbf{X} + \mu_1 \mathbf{I}_n,$$

with $\alpha_1 = \mathbb{E}[\sigma(z)]^2 + \frac{\mathbb{E}[\sigma''(z)]^2}{2d}$, $\beta_1 = \mathbb{E}[\sigma'(z)]^2$, and $\mu_1 = \mathbb{E}[\sigma^2(z)] - \mathbb{E}[\sigma(z)]^2 - \mathbb{E}[\sigma'(z)]^2$, for $z \sim \mathcal{N}(0, 1)$.

NTKs. For the Implicit-NTK, we define $\mathbf{K}_{ij}^* = k(\mathbf{x}_i^\top \mathbf{x}_j)$, i.e., $k(\mathbf{x}_i^\top \mathbf{x}_j) = h(g(\mathbf{x}_i^\top \mathbf{x}_j))$, for $i, j \in [n]$. It is easy to check that $k(0) = h(\angle^*)$ and $k(1) = h(g(1))$. Using implicit differentiation again, we have

$$k'(0) = \frac{(1 - \sigma_a^2)h'(\angle^*)}{\sigma_a^2 f'(\angle^*) - 1}, \ k''(0) = \frac{(1 - \sigma_a^2)^2(h''(\angle^*) - \sigma_a^2 f'(\angle^*)h''(\angle^*) + \sigma_a^2 h'(\angle^*)f''(\angle^*))}{(1 - \sigma_a^2 f'(\angle^*))^3}.$$

Now we are ready to present the asymptotic equivalent of the Implicit-NTK matrix.

Theorem 3 (Asymptotic approximation for Implicit-NTKs) *Let Assumptions 1 hold. As $n, d \rightarrow \infty$, the Implicit-NTK matrix $\mathbf{K}^*$ defined Eq. (7) in can be approximated consistently in operator norm, by the matrix $\overline{\mathbf{K}}$, that is $\|\mathbf{K}^* - \overline{\mathbf{K}}\|_2 \rightarrow 0$, where*

$$\overline{\mathbf{K}} = \dot{\alpha}\mathbf{1}\mathbf{1}^\top + \dot{\beta}\mathbf{X}^\top \mathbf{X} + \dot{\mu}\mathbf{I}_n,$$

with $\dot{\alpha} = k(0) + \frac{k''(0)}{2d}$, $\dot{\beta} = k'(0)$, and $\dot{\mu} = k(1) - k(0) - k'(0)$.

Theorem 4 (Asymptotic approximation for Explicit-NTKs) *Let Assumptions 1 hold. As $n, d \rightarrow \infty$, the Explicit-NTK matrix Θ defined in Eq. (5) can be approximated consistently in operator norm, by the matrix $\overline{\Theta}$, that is $\|\Theta^* - \overline{\Theta}\|_2 \rightarrow 0$, where*

$$\overline{\Theta} = \dot{\alpha}_1 \mathbf{1}\mathbf{1}^\top + \dot{\beta}_1 \mathbf{X}^\top \mathbf{X} + \dot{\mu}_1 \mathbf{I}_n,$$

with $\dot{\alpha}_1 = \mathbb{E}[\sigma(z)]^2 + \frac{3\mathbb{E}[\sigma''(z)]^2}{2d}$, $\dot{\beta}_1 = 2\mathbb{E}[\sigma'(z)]^2$, and $\dot{\mu}_1 = \mathbb{E}[\sigma^2(z)] + \mathbb{E}[\sigma'(z)^2] - \mathbb{E}[\sigma(z)]^2 - 2\mathbb{E}[\sigma'(z)]^2$ for $z \sim \mathcal{N}(0, 1)$.

Remark 2 *(i) Due to the homogeneity of the ReLU function, the Implicit-CK and the Implicit-NTK are essentially inner product kernel random matrices. Consequently, Theorem 1 and 3 can be built upon the results in [4]. We postpone the study on general activations to future work. (ii) The results in Theorem 2 and 4 generalize those of [1, 7] to the cases of "non-centred" activations, i.e., we do not require $\mathbb{E}[\sigma(z)] = 0$ for $z \sim \mathcal{N}(0, 1)$.*

3.2. The Equivalence between Implicit and Explicit NNs

In the following corollary, we show a concrete case of a single-layer explicit NN with an quadratic activation, that matches the CK or NTK eigenspectra of a ReLU implicit NN. The idea is to utilize the results of Theorems 1-4 to match the coefficients of the asymptotic equivalents such that $\alpha_1 = \alpha, \beta_1 = \beta, \mu_1 = \mu$, or $\dot{\alpha}_1 = \dot{\alpha}, \dot{\beta}_1 = \dot{\beta}, \dot{\mu}_1 = \dot{\mu}$. We implement numerical simulations to verify our theory. The numerical results are shown in Figure 1.
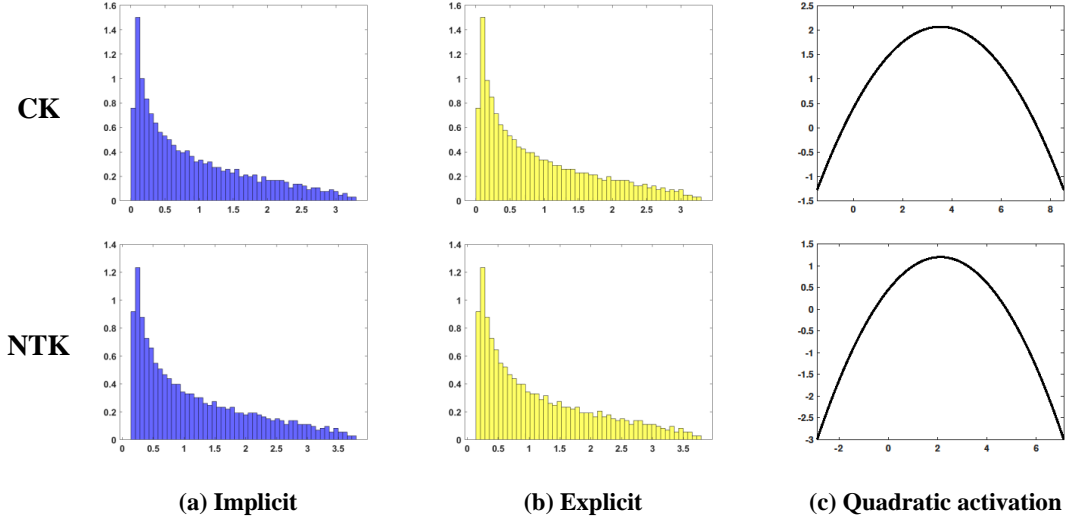
(a) Implicit   **(b) Explicit**   **(c) Quadratic activation**

Figure 1: We independently generate $n = 1\,000$ data points from the $d = 1\,200$-dimensional unit sphere. We use Gaussian initialization and σ_a^2 is set as 0.2. Upper: the CK results. Bottom: the NTK results. (a) spectral densities of implicit kernels, (b) spectral densities of explicit kernels, (c) quadratic activations.

Corollary 1 *We consider a quadratic polynomial activation $\sigma(t) = a_2 t^2 + a_1 t + a_0$. Let Assumptions 1 hold. As $n, d \to \infty$, the Implicit-CK matrix $\mathbf{G}^*$ defined in Eq. (6) can be approximated consistently in operator norm, by the Explicit-CK matrix $\mathbf{\Sigma}$ defined in Eq. (5), i.e., $\|\mathbf{G}^* - \mathbf{\Sigma}\|_2 \to 0$, as long as*

$$a_2 = \pm\sqrt{\frac{\mu}{2}} \quad a_1 = \pm\sqrt{\beta}, \quad a_0 = \pm\sqrt{\alpha - \frac{\mu}{d}} - a_2,$$

and the Implicit-NTK matrix $\mathbf{K}^*$ defined in Eq. (7) can be approximated consistently in operator norm, by the Explicit-NTK matrix $\mathbf{\Theta}$ defined in Eq. (5), i.e., $\|\mathbf{K}^* - \mathbf{\Theta}\|_2 \to 0$, as long as

$$a_2 = \pm\sqrt{\frac{\dot{\mu}}{6}}, \quad a_1 = \pm\sqrt{\frac{\dot{\beta}}{2}}, \quad a_0 = \pm\sqrt{\dot{\alpha} - \frac{\dot{\mu}}{d}} - a_2.$$

4. Conclusion

In this paper, we study the CKs and NTKs of high-dimensional ReLU implicit NNs. We prove the asymptotic spectral equivalents for Implicit-CKs and Implicit-NTKs. Moreover, we establish the equivalence between implicit and explicit NNs by matching the coefficients of the asymptotic spectral equivalents. In particular, we show that a single-layer explicit NN with carefully designed activations has the same CK or NTK eigenspectra as a ReLU implicit NN. For future work, it would be interesting to extend our analysis to more general data distributions and activation functions.

5

References

[1] Hafiz Tiomoko Ali, Zhenyu Liao, and Romain Couillet. Random matrices in service of ml footprint: ternary random features with no performance loss. *ICLR*, 2022.

[2] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[3] Shaojie Bai, Vladlen Koltun, and J Zico Kolter. Multiscale deep equilibrium models. *Advances in Neural Information Processing Systems*, 2020.

[4] Noureddine El Karoui. The spectrum of kernel random matrices. *The Annuals of Statistics*, 2010.

[5] Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. *Advances in neural information processing systems*, 33:7710–7721, 2020.

[6] Zhili Feng and J Zico Kolter. On the neural tangent kernel of equilibrium models. *arxiv*, 2020.

[7] Lingyu Gu, Yongqi Du, Zhang Yuan, Di Xie, Shiliang Pu, Robert Qiu, and Zhenyu Liao. " lossless" compression of deep neural networks: A high-dimensional neural tangent kernel approach. *Advances in Neural Information Processing Systems*, 35:3774–3787, 2022.

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

[9] Zenan Ling, Xingyu Xie, Qiuhao Wang, Zongpeng Zhang, and Zhouchen Lin. Global convergence of over-parameterized deep equilibrium models. In *International Conference on Artificial Intelligence and Statistics*, pages 767–787. PMLR, 2023.

[10] Lan V Truong. Global convergence rate of deep equilibrium models with general activations. *arXiv preprint arXiv:2302.05797*, 2023.

[11] Xingyu Xie, Qiuhao Wang, Zenan Ling, Xia Li, Guangcan Liu, and Zhouchen Lin. Optimization induced equilibrium networks: An explicit optimization perspective for understanding equilibrium models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.