Theoretical Principles of Deep Learning
CentraleSupélec MDS 2025/2026 - Paper Reading Assignment
On Exact Computation with an Infinitely Wide Neural Net

Adonis JAMAL adonis.jamal@student-cs.fr

February 1st, 2026

1 Introduction and Contributions

The theoretical study of deep learning often relies on the limit of infinite width to provide tractable analytical tools for understanding optimization and generalization [1]. While the correspondence between infinite fully-connected networks and Gaussian Processes is well-established, the Neural Tangent Kernel (NTK) introduced by Jacot et al. [2] characterized the evolution of fully-trained networks in this regime. The paper *On Exact Computation with an Infinitely Wide Neural Net* extends this framework to Convolutional Neural Networks (CNNs), addressing the computational infeasibility of evaluating kernels for modern architectures with pooling layers. The authors pose a central question: can the "infinite limit" of classic architectures like VGG or AlexNet achieve competitive classification performance on standard datasets such as CIFAR-10? To answer this, they derive the Convolutional NTK (CNTK) and provide the first efficient, exact algorithm for its computation, proving that a fully-trained, sufficiently wide network is equivalent to the kernel regression predictor using CNTK.

In this report, we focus on the paper's algorithmic contribution: the exact and efficient computation of the CNTK. We first detail the theoretical recursive formulas that allow for the exact evaluation of convolutional kernels with and without Global Average Pooling (GAP). We then describe our reproduction of the paper's main experimental result on a subset of the data, implementing the exact dynamic programming algorithm to evaluate CNTK performance on the CIFAR-10 dataset. Our experiments confirm the significant performance gap between vanilla CNTK and CNTK-GAP architectures. Finally, we experimentally illustrate the authors' findings regarding network depth, verifying the distinct performance characteristics of infinite-width architectures compared to their finite counterparts across varying depths. The full code is available at `https://github.com/adonis107/MDS-TDL`

2 Paper Overview

2.1 The Setting: From Finite Networks to Kernel Limits

The paper operates in the regime of extreme over-parametrization, specifically the limit where the width of the neural network (number of channels in convolutional layers or nodes in fully-connected layers) tends to infinity.

Historically, the connection between infinite-width networks and kernel methods was studied using Gaussian Processes (GPs). In this "weakly-trained" regime, only the last layer is optimized

while hidden layers remain fixed at their random initialization. However, this view fails to capture the dynamics of modern deep learning, where all layers are trained simultaneously.

The authors build upon the "fully-trained" regime introduced in [2], known as the Neural Tangent Kernel (NTK). In this setting, gradient descent on an Infinitely wide network is equivalent to kernel regression with a deterministic kernel Θ, defined as:

$$\Theta(x, x') = \mathbb{E}_{\theta \sim \mathcal{W}} \left\langle \frac{\partial f(\theta, x)}{\partial \theta}, \frac{\partial f(\theta, x')}{\partial \theta} \right\rangle \tag{1}$$

where $f(\theta, x)$ is the network output and $\mathcal{W}$ is the initialization distribution. THis setting provides a theoretical bridge to understand why over-parametrized networks optimize easily and generalize well.

2.2 The Question: Feasibility and Performance of Convolutional Kernels

Prior to this work, the exact computation of the NTK for Convolutional Neural Networks (CNTK) was considered computationally infeasible for large datasets like CIFAR-10. Researchers resorted to Monte Carlo approximations or finite-width proxies, which introduced variance and degraded performance [3].

The authors seek to answer two primary questions:

- **Algorithmic Feasibility:** Can we derive an efficient, exact algorithm to compute the CNTK for standard architectures (including pooling layers) without relying on approximations?
- **Performance Gap:** How well do these "infinite" CNNs actually perform compared to their finite counterparts such as AlexNet and VGG-19? Does the infinite width limit capture the power of deep learning?

2.3 The Solution: CNTK and Exact Algorithms

The paper provides a three-fold solution:

- **Algorithmic:** They derive an exact dynamic programming algorithm to compute the CNTK for convolutional networks, utilizing the specific properties of ReLU activations to enable efficient GPU implementation.
- **Theoretical:** They provide a non-asymptotic proof that fully-trained, sufficiently wide networks are equivalent to kernel regression with the CNTK.
- **Empirical:** They establish a new benchmark for kernel methods, achieving $\approx 77\%$ accuracy on CIFAR-10 with an 11-layer architecture, narrowing the gap between kernel methods and finite deep networks.

3 Methodology: Exact Recursive Computation

In this section, we detail the algorithmic contribution of the paper: the exact computation of the Convolutional Neural Tangent Kernel (CNTK).

The core contribution of the paper is an efficient algorithm to compute the CNTK kernel matrix $\Theta^{(L)}(x, x')$ for two inputs x and x'. For a network of depth L, the computation relies on a dynamic programming approach that recursively propagates covariance matrices through the layers.

3.1 Covariance Propagation

For a convolutional network, the kernel computation requires tracking two quantities layer-by-layer:

1. $\Sigma^{(h)}(x, x')$: The covariance of the pre-activations at layer h.
2. $\dot{\Sigma}^{(h)}(x, x')$: The derivative covariance, which captures the backward pass dynamics.

For ReLU activation $\sigma(z) = \max(0, z)$, the paper leverages closed-form arc-cosine kernels to compute these expectations efficiently without explicit integration. Given the covariance matrix Λ from the previous layer with correlation ρ, the key recursive formulas are:

$$\mathbb{E}[\sigma(u)\sigma(v)] = \frac{\sqrt{\Sigma_{11}\Sigma_{22}}}{2\pi}\left(\rho(\pi - \arccos\rho) + \sqrt{1-\rho^2}\right) \tag{2}$$

$$\mathbb{E}[\sigma'(u)\sigma'(v)] = \frac{1}{2\pi}(\pi - \arccos\rho) \tag{3}$$

These formulas allow the algorithm to compute the kernel for layer h directly from the covariance of layer $h-1$ via entry-wise matrix operations, making the process GPU-efficient.

3.2 Theoretical Guarantees

To rigorously justify the use of the CNTK, the authors prove that a sufficiently wide, fully-trained neural network is equivalent to kernel regression. The proof relies on analyzing the training dynamics under gradient descent.

The core idea is that as the width $m \to \infty$, the network enters a "lazy training" regime. By bounding the movement of the weight matrices $\|W(t) - W(0)\|_F$, the authors show that the weights stay infinitesimally close to their initialization. Consequently, the tangent kernel $H(t)$, which governs the training dynamics $\frac{du}{dt} = -H(t)(u - y)$, remains constant throughout training ($H(t) \approx H(0)$). This allows the complex non-linear training trajectory to be approximated by the linear solution of kernel regression.

3.3 CNTK-Vanilla vs. CNTK-GAP

The paper describes two variants of the kernel:

- **CNT-Vanilla (CNTK-V):** Corresponds to a standard convolutional network ending with a fully connected layer.
- **CNTK-GAP:** Includes a Global Average Pooling (GAP) operation before the final classification. This involves taking the mean over spatial dimensions, which significantly alters the final kernel $\Theta^{(L)}$ by accounting for cross-variances between all spatial patches.

4 Data and Experimental Setup

4.1 Dataset and Preprocessing

We utilized the CIFAR-10 dataset, a standard benchmark for image classification containing over $50,000$ training and $10,000$ test images across 10 classes. Consistent with the paper's protocol,

we applied no data augmentation to ensure a fair comparison between kernel methods and finite networks.

The labels were processed using a specific encoding scheme to improve regression performance: the target label y_c is set to 0.9 for the correct class, and -0.1 for incorrect classes. This zero-mean centering helps in the context of Mean Squared Error (MSE) regression.

4.2 Experimental Reproduction Setup

While the original paper scales the exact CNTK computation to the full CIFAR-10 using optimized CUDA kernels on NVIDIA Testa V100 GPUs, reproducing the full-scale experiment is computationally prohibitive for this project. We therefore adapted the experimental scale while strictly preserving the algorithmic implementation.

- **Data Subset:** We restricted the dataset to a subset of $N_{train} = 200$ and $N_{test} = 100$ images. This allows us to verify the correctness of the CNTK implementation and observe depth-dependent behaviors without requiring extensive computational resources.
- **Kernel Regression:** Classification is performed via exact kernel regression. For a test point x_{test}, the prediction $f^*(x_{test})$ is computed as:

$$f^*(x_{test}) = \ker(x_{test}, X_{train})(H^*)^{-1}Y_{train} \tag{4}$$

  where H^* is the kernel matrix on the training data. The class with the highest predicted value is selected.
- **Implementation:** We implemented the recursive CNTK formulas using PyTorch with GPU acceleration to handle the tensor operations required for the covariance propagation.

5 Results

5.1 Validation of Exact CNTK Implementation

We first validated our implementation by comparing the performance of the Vanilla (CNTK-V) and Global Average Pooling (CNTK-GAP) kernels at a fixed depth of 6. Table 1 presents our reproduced results on the reduced dataset ($N_{train} = 200$, $N_{test} = 100$) alongside the reference results from the original paper.

Method	Depth	Paper Acc.	Reproduced Acc.
CNTK-V	6	66.03%	25.00%
CNTK-GAP	6	76.73%	37.00%

Table 1: Comparison of classification accuracy between original paper results (full CIFAR-10) and our reproduction (subset).

While the absolute accuracy is lower due to the reduced training set size (kernel methods are highly data-dependent), our results successfully reproduce the key structural finding of the paper: the Global Average Pooling (GAP) operation provides a significant performance boost. In our experiments, GAP yielded a 48% relative improvement over the vanilla architecture. This is considerably higher than the $8-10\%$ improvement reported in the paper, likely due to the small dataset size amplifying the effect of pooling.

5.2 Depth Gap Analysis

To investigate the effect of depth on infinite-width networks, we evaluated both architectures across depths $L \in \{3, 4, 6, 11, 21\}$. The results are visualized in Figure 1.

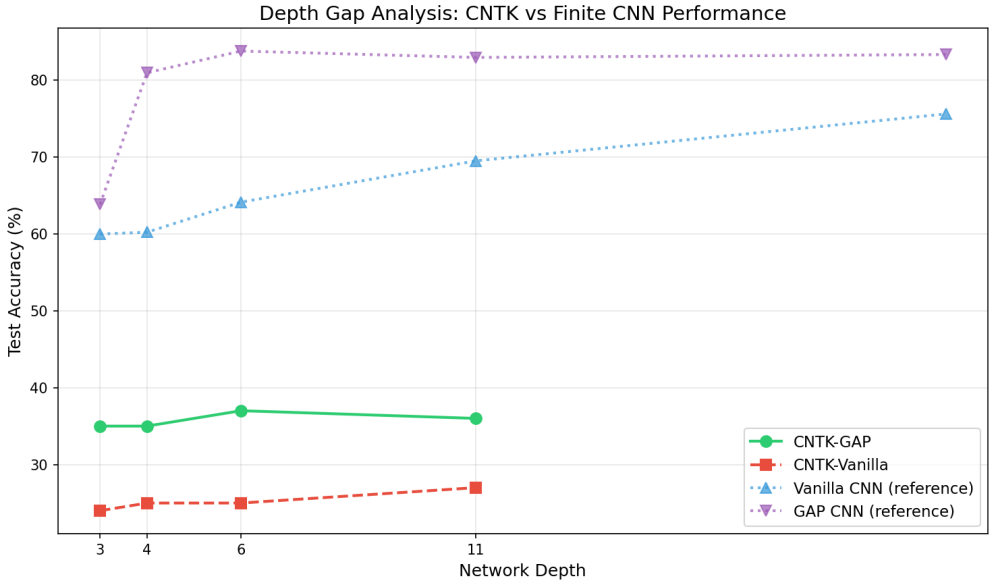

Figure 1: Depth Gap Analysis: Test accuracy of CNTK-GAP vs CNTK-Vanilla across varying network depths ($N_{train} = 200$, $N_{test} = 100$).

Our analysis reveals two trends:

- **Persistent GAP Advantage:** Across all tested depths, the GAP architecture consistently outperforms the Vanilla architecture.
- **Diminishing Returns with Depth:** Unlike finite CNNs, where accuracy typically improves significantly with depth, the performance of the CNTK models in our low-data regime shows signs of plateauing earlier. This aligns with the authors' observation that depth affects CNTKs differently than finite networks, where "extra-wideness" helps optimization but infinite depth does not necessarily guarantee better generalization.

These experiments illustrate that while the CNTK captures the behavior of infinitely wide networks, the architectural choice (specifically the inclusion of pooling) remains a critical factor for performance.

6 Conclusion and Discussion

In this report, we examined the theoretical and practicaal contributions of Arora et al. regarding the exact computation of the Convolutional Neural Tangent Kernel (CNTK). By implementing the recursive covariance propagation algorithm, we successfully reproduced the structural advantages of the Global Average Pooling (GAP) architecture and explored the effects of depth in the infinite-width regime.

6.1 Summary of Findings

Our reproduction confirms the paper's primary algorithmic claim: it is possible to compute the exact NTK for convolutional networks efficiently without relying on Monte Carlo approximations.

- **Architectural Impact:** We verified that the inclusion of Global Average Pooling significantly enhances kernel performance, yielding a relative accuracy improvement of 48% in our low-data regime.
- **Depth Behavior:** Our experiments suggest that while depth can improve performance, infinite-width networks exhibit distinct scaling characteristics compared to finite networks, often plateauing earlier.

6.2 Critical Discussion

While the exact CNTK establishes a powerful benchmark for kernel methods, reaching 77% accuracy on the full CIFAR-10 dataset, several critical points emerge from both the paper and our analysis:

The Performance Gap and Feature Learning: Despite the exact computation, there remains a persistent performance gap ($\approx 5-6\%$) between the infinite CNTK and its finite-width counterpart (CNN-GAP). As noted by the authors, this implies that finite width has its benefits. The fixed nature of the NTK kernel prevents it from adapting features to the data, a key advantage of finite-width networks trained via gradient descent. This suggests that the lazy training regime described by the NTK does not fully capture the success of modern deep learning.

Computational Feasibility: The authors claim efficiency, but the computational cost of the exact algorithm scales quadratically with the number of samples ($O(N^2)$) due to the kernel matrix construction. While feasible for CIFAR-10 on high-end GPUs, this approach becomes intractable for larger datasets like ImageNet without approximation. Our own reproduction was limited to a very small subset ($N = 200$) precisely due to these compute constraints, highlighting that while the algorithm is exact, its scalability differs fundamentally from the linear scaling ($O(N)$) of standard neural network training.

Conclusion: The paper "On Exact Computation with an Infinitely Wide Neural Net" successfully bridges the gap between deep learning architectures and kernel methods. It provides a rigorous tool to study the infinite width limit, proving that such networks are powerful classifiers in their own right. However, the remaining performance gap serves as evidence that the representational power of deep learning lies not just in over-parametrization, but in the feature learning dynamics that occur away from the infinite limit.

References

[1] Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net, 2019. URL https://arxiv.org/abs/1904.11955.

[2] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks, 2020. URL https://arxiv.org/abs/1806.07572.

[3] Roman Novak, Lechao Xiao, Yasaman Bahri, Jaehoon Lee, Greg Yang, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-dickstein. Bayesian deep convolutional networks with

many channels are gaussian processes. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1g30j0qF7.