On the Relationship Between Neural Tangent Kernel Frobenius Distance and Distillation Sample Complexity

Arnav Sharma
CortexPD Labs
arnavsharma.0914@gmail.com

Ahmed Wez
CortexPD Labs
ahmed.moh.wez@gmail.com

Karthik Srikumar
CortexPD Labs
Karthiksrikumar83@gmail.com

Abstract

Knowledge distillation is a popular method for compressing large neural networks, from large language models to computer vision models, into smaller, more efficient models. However, predicting the effectiveness of a distillation for any given teacher-student pair without incurring expensive training costs is a significant challenge. This concept is also relevant when designing models intended to resist distillation, a case common when developers try to protect their intellectual property. To address this, we propose a theoretical framework that connects the properties of a teacher model to the inherent difficulty of distillation. Our work is centered on the conjecture that, under Neural Tangent Kernel (NTK) assumptions, this difficulty is lower bounded by the distance between the teacher and student kernel matrices. We then propose Centered Kernel Alignment (CKA) as a computable proxy for this conjectured bound, based on the heuristic assumption that representation similarity reflects the similarity of the models' learning dynamics. This framework offers mathematical tools to estimate the feasibility of distillation prior to experimentation.

1 Introduction

Knowledge distillation is a foundational approach in machine learning for the dual purpose of model compression and knowledge transfer (4; 2). The process consists of training a smaller, "student" model to mimic the behavior of a previously trained, larger-capacity "teacher" model by having the student learn from the teacher's full, soft probability distributions (softened logits) over the outputs. This is important as the soft targets provide valuable information about class-to-class relationships and generalizing capabilities not present when using the simple ground-truth labels. This is critical for placing state-of-the-art foundation models into deployment scenarios on devices with limited resources, such as mobile devices and edge computing.

Though there has been significant empirical testing across many architectures (10; 11; 3), we still have a limited theoretical understanding of why knowledge transfer works. Mathematically, we can describe distillation as a problem in statistical learning theory, where the goal of the student network is to create an approximation of the function learned by the teacher (6). We need to look at the functional properties of both networks to assess their compatibility. The central theoretical questions concern both the sample complexity of the transfer—namely what data is needed for the student to successfully reproduce the teacher—and whether the student's ability to approximate a teacher is always capped by the teacher they chose.

A large divide between empirical practice and theoretical comprehension creates a substantial bottleneck: the selection of compatible teacher-student pairs relies on heuristics and expensive, iterative experimentation. In the face of computational cost present in today's architectures, this is becoming unreasonable. This leads us to the central research question of this work: can distillation difficulty be formalized and predicted a priori? More specifically, can we provide a rigorous bound on the sample complexity required for a successful transfer with respect to an intrinsic, computable measure of functional and geometric divergence between the teacher and student models?

1.1 Related Work

Model compression via knowledge transfer was formalized in influential studies on knowledge distillation (4; 2). Since then, there has been extensive empirical support for its efficacy across a variety of domains and architectures, including convolutional networks (10) and large-scale transformers (11), as recently surveyed by (3). However, the theoretical principles underlying the conditions under which distillation is effective are less understood. Recent work has attempted to address this gap within the NTK framework by analyzing distillation in the infinite-width scenario, which provides a mathematically tractable setting for analyzing the learning dynamics of deep neural networks (6). This paper builds directly on this theoretical aspect of the literature.

Framework and Contributions

This paper provides a theoretical framework that relates the sample complexity of distillation to the geometric distance between teacher and student models as measured by their respective kernel matrices. Specifically, our main contributions are as follows:

- **A conjecture for a lower bound on distillation error**, positing that a student's minimum achievable error is governed by the Frobenius distance between the teacher and student Neural Tangent Kernels (NTKs) (Theorem 4.1).
- **A proof establishing an exact identity between Centered Kernel Alignment (CKA) and the geometric distance of normalized Gram matrices** (Theorem 4.2). This provides a practical proxy for our main conjecture, under a stated heuristic assumption.
- **A set of precise theoretical conjectures** that extend the framework by linking distillation difficulty to information-theoretic measures and architecture-specific properties, such as the entropy of transformer attention patterns (Theorem 4.4, Theorem 4.6).

2 Main Definitions

Definition 2.1 (Teacher and Student Networks)**.** Let $f_T : \mathcal{X} \to \mathcal{Y}$ and $f_S : \mathcal{X} \to \mathcal{Y}$ denote the teacher and student networks, respectively, where $\mathcal{X} \subseteq \mathbb{R}^d$ is the input space.

Definition 2.2 (Sample Complexity Gap)**.** For a target accuracy $\epsilon > 0$ and confidence $\delta > 0$, the sample complexity gap is $\Delta(\epsilon, \delta) = n_{\text{distill}}(\epsilon, \delta) - n_{\text{scratch}}(\epsilon, \delta)$, where n_{distill} is the number of samples required for the student to match the teacher's performance via distillation, and n_{scratch} is the sample complexity for training the student from scratch on true labels.

Definition 2.3 (Neural Tangent Kernel (NTK))**.** For a neural network $f(x; \theta)$ with parameters θ, the NTK at initialization θ_0 is the $n \times n$ matrix K with entries $K_{ij} = \nabla_\theta f(x_i; \theta_0) \cdot \nabla_\theta f(x_j; \theta_0)$. We denote the teacher and student NTKs as K_T and K_S, respectively.

Definition 2.4 (Centered Kernel Alignment (CKA))**.** For representation matrices $X_T, X_S \in \mathbb{R}^{n \times d}$ from a teacher and student, CKA measures the similarity of their Gram matrices $K_T = X_T X_T^T$ and $K_S = X_S X_S^T$. Let $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ be the centering matrix. Then,

$$\text{CKA}(X_T, X_S) = \frac{\langle HK_TH, HK_SH \rangle_F}{\|HK_TH\|_F \|HK_SH\|_F}$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product.

3 Core Assumptions

Our theoretical results are based on the following standard assumptions from statistical learning theory. Here we give only a brief justification; more extended discussions can be found in Supplementary Section B.

Assumption 3.1 (Neural Tangent Kernel Regime)**.** Both teacher and student networks are sufficiently broad to function in the NTK regime in which the kernels of both networks vary approximately constant for the duration of the training (5). This allows analysis using kernel regression methods.

Assumption 3.2 (Feature Kernel as a Heuristic Proxy for NTK)**.** Let K^{NTK} be the Neural Tangent Kernel of a network at initialization, and let $K^{\text{Gram}} = XX^T$ be the Gram matrix derived from the feature representations of a subnetwork. This manuscript operates under a **working heuristic** that the geometric similarity of these two types of kernels is correlated. Specifically, we assume that $1 - \text{CKA}(X_T, X_S)$ serves as a practical, computable proxy for the normalized NTK distance.

Justification: This is a broad, non-standard assumption, and we are not proving its validity here, but we find it plausible based on existing empirical research on overparameterized models. The neural tangent kernel (NTK) describes the learning dynamics of the function being learned, under gradient descent with the framework of initialization (5; 8). The representations learned by neural networks, which correspond to the output of the softmax layer ln p(x), that are represented by the matrix X in this work, is determined by the learning dynamics governed by the NTK. Therefore, it is reasonable to think that networks that learn with fundamentally different learning dynamics (i.e., different NTKs) will produce representations with different geometric structures (i.e., different Gram matrices). Previous empirical studies have proposed the idea of representational similarity based on CKA, as a measure of representational similarity relates to behavior and generalization (7). We formalize this idea by proposing that tractable similarity of representations will be useful as a proxy for the intractable similarity of the underlying function spaces that could be accessed through gradient descent learning. The verification of the regimes in which this heuristic holds is an important future work direction.

Assumption 3.3 (Bounded Outputs & Lipschitz Continuity)**.** Teacher and student outputs are bounded, $\|f(x)\| \leq B$. The teacher function f_T is L-Lipschitz. These are standard regularity conditions ensuring that generalization bounds can be applied.

Assumption 3.4 (I.I.D. Data)**.** Training data are drawn i.i.d. from a distribution $\mathcal{D}$ with bounded support.

4 Theoretical Framework and Main Results

Our framework establishes a direct link from the similarity of teacher and student models, as measured by their kernel matrices, to the sample complexity of distillation.

Conjecture 4.1 (Distillation Infeasibility from Kernel Distance)**.** Under Assumptions 3.1, 3.3, and 3.4, we conjecture that the approximation error of the student f_S with respect to the teacher f_T is lower-bounded by the Frobenius distance between their respective NTK matrices. Specifically, there may exist a constant $C_A > 0$ dependent on the model architectures and data distribution such that:

$$\|f_S^* - f_T\|_{L^2(\mathcal{D})}^2 \geq C_A \|K_T - K_S\|_F^2,$$

where $f_S^* = \arg\min_{f \in \mathcal{H}_S} \|f - f_T\|_{L^2(\mathcal{D})}$ is the best-in-class student approximation. An immediate consequence is that for any target error ϵ, distillation is infeasible if the teacher and student are sufficiently dissimilar, i.e., if $\|K_T - K_S\|_F^2 > \epsilon / C_A$, no amount of data can bridge the performance gap.

Heuristic Argument. The total expected squared error of the student decomposes into approximation error and estimation error:

$$\mathbb{E}[\|f_S - f_T\|_{L^2(\mathcal{D})}^2] = \underbrace{\|f_S - f_S^*\|_{L^2(\mathcal{D})}^2}_{\text{Estimation Error}} + \underbrace{\|f_S^* - f_T\|_{L^2(\mathcal{D})}^2}_{\text{Approximation Error}} + \text{cross-term}.$$

The estimation error arises solely due to finite sampling and vanishes as the number of samples $n \to \infty$. The approximation error is a real limitation based on the student's hypothesis space $\mathcal{H}_S$, as

it is a property of the set of hypotheses we consider to approximate the additive model. Because the estimation error can only be non-negative, the total error will always be lower-bounded below the approximation error:

$$\mathbb{E}[\|f_S - f_T\|^2_{L^2(\mathcal{D})}] \geq \|f^*_S - f_T\|^2_{L^2(\mathcal{D})}.$$

The core of this conjecture is the posited inequality linking this function-space approximation error to the matrix-space distance between kernels. This link is highly non-trivial. It suggests that a large distance between the kernels, which define the geometry of the respective function spaces, implies that the teacher function $f_T \in \mathcal{H}_T$ is "far" from any function in the student space $\mathcal{H}_S$. If this holds, then to achieve a total error $\mathbb{E}[\|f_S - f_T\|^2] \leq \epsilon$, it is necessary that the irreducible approximation error is also less than ϵ. This gives the infeasibility condition stated in the conjecture, establishing a hard limit on distillability governed by architectural (kernel) mismatch. A rigorous proof of the core inequality remains an open problem. □

Proposition 4.2 (CKA as a Proxy for Kernel Distance)**.** *Let $K_T = X_T X_T^T$ and $K_S = X_S X_S^T$ be empirical Gram matrices from hidden representations. The squared Frobenius distance between the centered and normalized Gram matrices is exactly related to their CKA value:*

$$\left\| \frac{HK_TH}{\|HK_TH\|_F} - \frac{HK_SH}{\|HK_SH\|_F} \right\|^2_F = 2\left(1 - \mathit{CKA}(X_T, X_S)\right)$$

Proof. This result follows directly from the definition of the Frobenius norm and CKA. A full proof is provided in Supplementary Section A.1. □

Remark: In the NTK regime, empirical Gram matrices approximate the true NTK matrices. Therefore, high CKA (a value approaching 1) indicates high similarity between the normalized kernels, suggesting a smaller NTK distance and thus easier distillation, as per Theorem 4.1.

4.1 Conjectures and Future Directions

Beyond the NTK regime and for specific architectures like transformers, we propose several conjectures that connect more fine-grained properties of the teacher to distillation difficulty.

Conjecture 4.3 (Approximation Error Bound from CKA)**.** We conjecture that under certain regularity conditions on the teacher and student network architectures, the true approximation error is directly bounded by a function of CKA. Specifically, there may exist a constant C such that for CKA sufficiently close to 1:

$$\|f_S - f_T\|_{L^2(\mathcal{D})} \leq C\sqrt{1 - \text{CKA}(X_T, X_S)}$$

Heuristic Argument. Theorem 4.2 provides an exact identity: $2(1 - \text{CKA}) = \|\hat{K}_T - \hat{K}_S\|^2_F$, where $\hat{K}$ denotes a normalized, centered Gram matrix. There are two aspects associated with the main gap in proving this hypothesis. First is the formal justification that the similarity of Gram matrices guarantees similarity of the true NTKs (as described in Assumption 3.2). The second outcome is to show that proximity for *normalized* kernels implies that the *unnormalized* kernels that govern function approximation are also boundedly close. This step would probably require significant assumptions regarding the kernels' spectra since normalization loses information about scale (distance). If these substantial gaps are crossed, the result follows. □

Conjecture 4.4 (Attention Complexity and Distillation)**.** For transformer networks, define the teacher's average attention entropy as $\mathcal{C}_{\text{attn}} = \frac{1}{LH}\sum_{l,h} H(A^{(l,h)})$, where $H(\cdot)$ is the Shannon entropy of an attention head's probability distribution. We conjecture that higher attention complexity correlates with increased distillation sample complexity: $\Delta(\epsilon, \delta) \geq f(\mathcal{C}_{\text{attn}})$ for some monotonically increasing function f.

Remark 4.5 (Reasoning). This conjecture is based on the notion that high-entropy attention patterns are less sparse, yielding more complex, distributed functional mappings over the input sequence. These complex mappings may be more inherently difficult for a small-capacity student model to approximate, which would in turn contribute to the approximation error, and the number of samples needed for a student to learn such complex behavior. Conversely, teachers exhibiting "peaked" or low-entropy attention may produce simpler more transferable input-output associations.

Conjecture 4.6 (Information-Theoretic Bound)**.** The distillation sample complexity gap may be lower-bounded by the gap in mutual information between the teacher's/student's representations and the task labels Y:

$$\Delta(\epsilon, \delta) \geq C \cdot \frac{I(X_T; Y) - I(X_S; Y)}{\log(1/\epsilon)}$$

for some constant C.

Proof Roadmap. This conjecture reconceptualizes distillation as a problem of information transfer. The reasoning is that if the representations X_S of the student agent are poorer in information-theoretic terms (e.g., capture less information about the labels Y than the representations X_T of the teacher agent), a performance gap must happen. Providing a rigorous proof would be tedious and likely require enabling information-theoretic generalization bounds (e.g., using Russo & Zou, Xu & Raginsky). One would need to link the available upper bound on the generalization gap resulting from the difference in mutual information to a corresponding sample complexity gap. This is more speculative than other possible theoretical approaches, but is nevertheless a key angle to examine. □

5 Limitations and Future Work

5.1 Limitations

NTK Regime Assumption: The results that we have established (Theorem 4.1) are limited to the infinite-width NTK setting, and their relevance to real-world, finite-width networks in which the kernel changes during training is an approximation. In the future, we should hope to extend these bounds past the NTK context.

We have established results only for the infinite-width NTK regime. In finite-width networks, where the kernel evolves during training, our results offer an approximation. Ultimately, next steps should be to push these bounds outside the NTK setting.

Proxy Assumption: The capability of our framework to use CKA as a practical metric overly relies on Assumption 3.2, which connects the geometry of learned representations to the NTK at initialization. This is reasonable in overparameterized models, as features learned late in training can remain stable, but this connection is non-trivial and not formally proven. A significant area for future work will be to confirm the regimes where such assumption holds.

Computability: Although CKA is much more computationally efficient than computing the exact NTK, both can still be extremely computationally expensive for large datasets ($n > 10^5$). Therefore, we require scalable approximation methods, such as sketching, to be useful in practice. We give the details on computational complexity and scalable algorithms in Supplementary Section A.2.

Constants in Bounds: The constant C in our main proposition is tied to characteristics of the data distribution and model architecture, which may be challenging to estimate. Therefore, the highlight of the bound is the relationship it reflects with respect to kernel distance, not getting an exact sample number.

5.2 Future Work

This work lays a theoretical foundation for understanding distillation. Key directions for future research include:

- **Theoretical Extensions:** Proving the conjectures presented (4.3, 4.4, 4.6), or finding counterexamples. Developing non-NTK-based bounds for practical, finite-width networks is a critical next step.
- **Architecture-Specific Analysis:** Extending our framework to explicitly account for architectural components like convolutional layers, normalization, and residual connections.
- **Empirical Validation:** Rigorously testing our theoretical predictions on a wide range of teacher-student pairs across diverse datasets. A brief plan for such experiments is outlined in Supplementary Section C.

6 Conclusion

In this work, we have introduced a mathematically grounded framework that allows us to reason about the hardness of knowledge distillation. By formalizing distillation hardness as a gap in sample complexity, we derived an actionable, provable bound rooted in the Frobenius distance between teacher and student Neural Tangent Kernels (NTKs). By assuming a link between representation similarity and NTK similarity, we propose Centered Kernel Alignment as a useful, computable measure for estimating distillation difficulty.

Although our best results only hold in the NTK regime, we have conjectured a series of questions that lead to a fuller theory that captures information-theoretic and architecture-based properties. This framework allows us to transition the field away from purely empirical determinants towards a scientific, predictive principle for knowledge transfer in deep learning.

References

[1] Arora, S., Du, S. S., Hu, W., Li, Z., and Wang, R. (2019). On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[2] Bucilua, C., Caruana, R., and Niculescu-Mizil, A. (2006). Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[3] Gou, J., Yu, B., Maybank, S. J., and Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6), 1789-1819.

[4] Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

[5] Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[6] Ji, S. and Zhu, Z. (2020). Knowledge distillation in wide neural networks: Risk bound, data efficiency and imperfect teacher. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[7] Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. (2019). Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*.

[8] Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Sohl-Dickstein, J., and Pennington, J. (2019). Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[9] Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. (2017). SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[10] Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. (2014). FitNets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

[11] Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Supplementary Materials

A Proofs and Computational Details

A.1 Full Proof of Proposition 4.2

Proposition A.1 (CKA as a Proxy for Kernel Distance)**.** *Let $K_T = X_T X_T^T$ and $K_S = X_S X_S^T$ be empirical Gram matrices from hidden representations. Let H be the centering matrix. The squared Frobenius distance between the centered and normalized Gram matrices is exactly related to their CKA value by:*

$$\left\| \frac{HK_TH}{\|HK_TH\|_F} - \frac{HK_SH}{\|HK_SH\|_F} \right\|_F^2 = 2\left(1 - \mathit{CKA}(X_T, X_S)\right)$$

Proof. This result follows directly from the definitions of the Frobenius norm and CKA. Let $\hat{K}_T = \frac{HK_TH}{\|HK_TH\|_F}$ and $\hat{K}_S = \frac{HK_SH}{\|HK_SH\|_F}$ be the normalized, centered Gram matrices. By construction, both $\hat{K}_T$ and $\hat{K}_S$ have a unit Frobenius norm, i.e., $\|\hat{K}_T\|_F = 1$ and $\|\hat{K}_S\|_F = 1$.

By the definition of the Frobenius norm:

$$\begin{aligned}
\|\hat{K}_T - \hat{K}_S\|_F^2 &= \langle \hat{K}_T - \hat{K}_S, \hat{K}_T - \hat{K}_S \rangle_F \\
&= \langle \hat{K}_T, \hat{K}_T \rangle_F - 2\langle \hat{K}_T, \hat{K}_S \rangle_F + \langle \hat{K}_S, \hat{K}_S \rangle_F \\
&= \|\hat{K}_T\|_F^2 - 2\langle \hat{K}_T, \hat{K}_S \rangle_F + \|\hat{K}_S\|_F^2
\end{aligned}$$

Since $\|\hat{K}_T\|_F = 1$ and $\|\hat{K}_S\|_F = 1$:

$$\|\hat{K}_T - \hat{K}_S\|_F^2 = 1 - 2\langle \hat{K}_T, \hat{K}_S \rangle_F + 1 = 2 - 2\langle \hat{K}_T, \hat{K}_S \rangle_F$$

By the definition of Centered Kernel Alignment (Theorem 2.4):

$$\text{CKA}(X_T, X_S) = \langle \hat{K}_T, \hat{K}_S \rangle_F$$

Substituting this back into the previous equation gives the stated identity:

$$\|\hat{K}_T - \hat{K}_S\|_F^2 = 2(1 - \text{CKA}(X_T, X_S))$$

This completes the proof. □

Remark A.2 (Connecting CKA to the NTK Bound). Proposition 4.1 bounds distillation difficulty using the NTK distance, while Proposition 4.2 relates CKA to the Gram matrix distance. To connect these two results, we rely on Assumption 3.2, which posits that the similarity of activation-based Gram matrices is a valid proxy for the similarity of gradient-based NTKs. Thus, if a value of high CKA is close to 1 (meaning that the normalized Gram matrices are close together), it is reasonable to assume that the NTK distance between those definitions is also small and thus distillation may be easier. This assumption leads to a practically estimable theoretical bound.

A.2 Computational Efficiency and Complexity

The theoretical quantities that are very important in our framework (NTK distance $\|K_T - K_S\|_F$ and CKA) are not computationally efficient for large datasets. In this section we provide a formal analysis of these costs and offer scalable approximation methods.

Exact Computation Complexity. The direct computation of our proposed diagnostics is often infeasible. Let n be the number of samples, P the number of model parameters, and d the dimension of hidden representations.

- **NTK Matrix Construction:** Computing a single entry $K(x_i, x_j) = \nabla_\theta f(x_i) \cdot \nabla_\theta f(x_j)$ requires two Jacobian computations, each costing $O(P)$. To construct the full $n \times n$ NTK matrix, the total time complexity is $O(n^2 P)$ and space complexity is $O(n^2 + nP)$.
- **CKA Computation:** For representations $X \in \mathbb{R}^{n \times d}$, computing the Gram matrix $K = XX^T$ requires $O(n^2 d)$ time and $O(n^2)$ space. CKA computation is dominated by this step.

Scalable Approximation Algorithms. For large n, we must resort to approximations. We propose using sketching based on random projections.

- **Random Projections (Sketching):** Instead of materializing the full Jacobian matrix $J \in \mathbb{R}^{n \times P}$, we can compute a sketch $\tilde{J} = JS \in \mathbb{R}^{n \times r}$, where $S \in \mathbb{R}^{P \times r}$ is a random projection matrix (e.g., Gaussian) with $r \ll n, P$. The approximate kernel is then $\tilde{K} = \tilde{J}\tilde{J}^T$. The time complexity is dominated by computing the projected Jacobians, which costs $O(nPr)$, and forming $\tilde{K}$ in $O(nr^2)$. The approximation error for $\|K - \tilde{K}\|_F$ is bounded with high probability and scales with $1/\sqrt{r}$.

Algorithm 1 Efficient NTK Distance and CKA Estimation via Sketching

Input: Teacher f_T, Student f_S, Data $\{x_i\}_{i=1}^n$, sketch size r.
Output: Approx. NTK distance d_F, Approx. CKA score c.
// NTK Distance Approximation
Sample random projection matrix $S_P \in \mathbb{R}^{P \times r}$.
for $i = 1$ **to** n **do**
  Compute Jacobians $\mathbf{j}_{T,i} = \nabla_\theta f_T(x_i)$, $\mathbf{j}_{S,i} = \nabla_\theta f_S(x_i)$.
  Project Jacobians: $\tilde{\mathbf{j}}_{T,i} = \mathbf{j}_{T,i} S_P$, $\tilde{\mathbf{j}}_{S,i} = \mathbf{j}_{S,i} S_P$.
end for
Form sketched Jacobian matrices $\tilde{J}_T, \tilde{J}_S \in \mathbb{R}^{n \times r}$.
Compute sketched kernels: $\tilde{K}_T = \tilde{J}_T \tilde{J}_T^T$, $\tilde{K}_S = \tilde{J}_S \tilde{J}_S^T$.
$d_F = \|\tilde{K}_T - \tilde{K}_S\|_F$.
// CKA Approximation (if d is large)
Get representations $X_T, X_S \in \mathbb{R}^{n \times d}$.
Sample random projection matrix $S_d \in \mathbb{R}^{d \times r}$.
Compute sketched representations: $\tilde{X}_T = X_T S_d$, $\tilde{X}_S = X_S S_d$.
$c = \text{CKA}(\tilde{X}_T, \tilde{X}_S)$.

B Extended Justification of Assumptions

- **Assumption 3.1 (NTK Regime):** This presumption is integral to our established findings. Although contemporary networks are of finite width, there is a body of work demonstrating that many over-parameterized models can train in a manner which is well-approximated by the NTK theory, particularly early in the training process. Our framework provides an expected theoretical baseline and the extent to which models deviate from the NTK regime is an interesting direction for future work.
- **Assumption 3.2 (Boundedness and Lipschitz Continuity):** These are typical regularity assumptions in learning theory. Bounded outputs can be obtained using an activation function such as $\tanh$ or sigmoid in the output layer. Many common architectures satisfy Lipschitz continuity, especially when combined with a technique such as spectral normalization or gradient clipping. These conditions are required to avoid problems associated with pathological cases where the function value or gradient may blow up, which would make generalization impossible.
- **Assumption 3.3 (I.I.D. Data):** In most supervised learning contexts, this is the typical assumption. It assures that the training data is representative of the test data, enabling the generalization of training data to test data. Our outcomes could be generalized to non-i.i.d. contexts (e.g., time series) but would entail different analytical tools.

C Brief Empirical Plan

While the current work is theoretical, we propose the following speculative empirical plan to validate its predictions. We heavily stress testing assumption 3.2.

Objective: To empirically verify the correlation between our proposed metrics (NTK distance, CKA) and the measured sample complexity gap $\Delta(\epsilon, \delta)$.

Methodology:

1. **Setup:** Use a controlled setting, such as distilling a ResNet-18 teacher to a ResNet-9 student on CIFAR-10. Also, use synthetic datasets where teacher properties can be precisely manipulated.
2. **Metric Computation:** Prior to distillation, calculate the CKA by layer between the initialized student and the trained teacher. Use the sketching method to calculate the NTK distance on a subset of the data as described in Algorithm 1. For transformer models, also compute the teacher's average attention entropy ($\mathcal{C}_{\text{attn}}$).
3. **Measuring Sample Complexity Gap:** Train multiple student models via distillation, varying the number of training samples (n). For each n, measure the final student accuracy. The value of n required to reach a target accuracy (e.g., 99% of teacher's accuracy) is n_{distill}. Compare this to n_{scratch} measured by training the student on true labels.
4. **Analysis:** Plot the measured sample complexity gap Δ against the pre-computed metrics (CKA, NTK distance, $\mathcal{C}_{\text{attn}}$). We would expect to see a strong positive correlation between Δ and NTK distance, and a strong negative correlation between Δ and CKA.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction claim to provide a theoretical framework for predicting distillation difficulty.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5, "Discussion of Limitations," explicitly addresses the limitation.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theoretical results are based on the core assumptions listed in Section 3 and appropiate proof blocks.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: This paper is purely theoretical and does not contain experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: This paper is purely theoretical and does not contain data or code.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: This paper is purely theoretical and does not contain experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: This paper is purely theoretical and does not contain experimental results such as statistical significance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: This paper is purely theoretical and does not contain experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper is purely theoretical and does not involve human subjects, data, or experiments that would raise ethical concerns addressed by the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This paper is purely theoretical and while this resaerch has applications, this link is indirect. There are no direct foreseeable negative societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper does not introduce or use any data or models, so there are no associated risks of misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: This paper does not use any external assets such as code, data, or pre-trained models.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: No new assets are introduced in this paper.

Guidelines:

- The answer NA means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The research does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The research does not involve human subjects, so no IRB approval was necessary.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

Justification: Large Language Models were not used in the development of the core theoretical methods or proofs presented in this research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.