Analysis of Knowledge Transfer in Kernel Regime

N.B. When citing this work, cite the original published paper.

(article starts on next page)

Analysis of Knowledge Transfer in Kernel Regime

Ashkan Panahi
Chalmers University of Technology
Gothenburg, Sweden
ashkan.panahi@chalmers.se

Arman Rahbar
Chalmers University of Technology
Gothenburg, Sweden
armanr@chalmers.se

Chiranjib Bhattacharyya
Indian Institute of Science
Bangalore, India
chiru@iisc.ac.in

Devdatt Dubhashi
Chalmers University of Technology
Gothenburg, Sweden
dubhashi@chalmers.se

Morteza Haghir Chehreghani
Chalmers University of Technology
Gothenburg, Sweden
morteza.chehreghani@chalmers.se

ABSTRACT

Knowledge transfer is shown to be a very successful technique for training neural classifiers: together with the ground truth data, it uses the "privileged information" (PI) obtained by a "teacher" network to train a "student" network. It has been observed that classifiers learn much faster and more reliably via knowledge transfer. However, there has been little or no theoretical analysis of this phenomenon. To bridge this gap, we propose to approach the problem of knowledge transfer by regularizing the fit between the teacher and the student with PI provided by the teacher. Using tools from dynamical systems theory, we show that when the student is an extremely wide two layer network, we can analyze it in the kernel regime and show that it is able to interpolate between PI and the given data. This characterization sheds new light on the relation between the training error and capacity of the student relative to the teacher. Another contribution of the paper is a quantitative statement on the convergence of student network. We prove that the teacher reduces the number of required iterations for a student to learn, and consequently improves the generalization power of the student. We give corresponding experimental analysis that validates the theoretical results and yield additional insights.

CCS CONCEPTS

• **Computing methodologies → Neural networks**.

KEYWORDS

Knowledge Transfer, Teacher, Student, Kernel Regime

1 INTRODUCTION

Knowledge transfer considers improving learning processes by leveraging the knowledge learned from other tasks or trained models. Several studies have demonstrated the effectiveness of knowledge transfer in different settings. For instance, in [3], knowledge transfer has been used to improve object detection models. Knowledge transfer has been applied in different levels to neural machine translation in [10]. It has also been employed for Reinforcement learning in [20]. Recommender systems can also benefit from knowledge transfer as shown in [13].

An interesting case of knowledge transfer is *privileged information* [17, 18] where the goal is to supply a student learner with privileged information during training session. Another special case is concerned with knowledge distillation [8] which suggests to train classifiers using the real–valued outputs of another classifier as target values than using actual ground–truth labels. These two paradigms are unified within a consistent framework in [11].

The privileged information paradigm introduced in [18] aims at mimicking some elements of human teaching in order to improve the process of learning with examples. In particular, the teacher provides some additional information for the learning task along with training examples. This privileged information is only available during the training phase. The work in [14] outlines the theoretical conditions required for the additional information from a teacher to a student. If a teacher satisfies these conditions, it will help to accelerate the learning rate. Two different mechanisms of using privileged information are introduced in [16]. In the first mechanism, the concept of similarity in the training examples in the student is controlled. In the second one, the knowledge in the space of privileged information is *transferred* to the space where the student is working.

In 2014, Hinton et al. [8] studied the effectiveness of knowledge distillation. They showed that it is easier to train classifiers using the real–valued outputs of another classifier as target values than using actual ground–truth labels. They introduced the term *knowledge distillation* for this phenomenon. Since then, distillation–based training has been confirmed in several different types of neural networks [3, 21, 22]. It has been observed that optimization is generally more well–behaved than with label-based training, and it needs less regularization or specific optimization tricks.

While the practical benefits of knowledge transfer in neural networks (e.g. via distillation) are beyond doubt, its theoretical justification remains almost completely unclear. Recently, Phuong

and Lampert [15] made an attempt to analyze knowledge distillation in a simple model. In their setting, both the teacher and the student are *linear* classifiers (although the student's weight vector is allowed a over-parametrized representation as a product of matrices). They give conditions under which the student's weight vector converges (approximately) to that of the teacher and derive consequences for generalization error. Crucially, their analysis is limited to linear networks.

Knowledge transfer in neural networks and privileged information are related through a unified framework proposed in [11]. In this work knowledge transfer in neural networks is examined from a theoretical perspective via casting that as a form of learning with privileged information. However, [11] uses a heuristic argument for the effectiveness of knowledge transfer with respect to generalization error rather than a rigorous analysis.

In our knowledge transfer analysis, we assume the so-called *kernel regime*. A series of recent works, e.g., [1, 2, 5, 12] achieved breakthroughs in understanding how (infinitely) wide neural network training behaves in this regime, where the dynamics of training by gradient descent can be approximated by the dynamics of a linear system. We extend the repertoire of the methods that can be applied in such settings.

Contributions: We carry out a theoretical analysis of knowledge transfer which consistently covers aspects of privileged information for non-linear neural networks. We situate ourselves in the recent line of work that analyzes the dynamics of neural networks under the kernel regime. It was shown that the behaviour of training by gradient descent (GD) in the limit of very wide neural networks can be approximated by *linear system dynamics*. This is dubbed the *kernel regime* because it was shown in [9] that a fixed kernel – the *neural tangent kernel* – characterizes the behavior of fully-connected infinite width neural networks in this regime.

Our framework is general enough to encompass Vapnik's notion of *privileged information* and provides a unified analysis of *generalized distillation* in the paradigm of *machines teaching machines* as in [11]. For this analysis, we exploit new tools that go beyond the previous techniques in the literature, as in [1, 2, 5, 12] - we believe these new tools will contribute to further development of the nascent theory. Our main results are:

i) We formulate the knowledge transfer problem as a least squares optimization problem with regularization provided by privileged knowledge. This allows us to characterize precisely what is learnt by the student network in Theorem 3.1 showing that the student converges to an interpolation between the data and the privileged information guided by the strength of the regularizer.

ii) We characterize the speed of convergence in Theorem 3.2 in terms of the overlap between a combination of the label vector and the knowledge vectors and the spectral structure of the data as reflected by the vectors.

iii) We introduce novel techniques from systems theory, in particular, Laplace transforms of time signals to analyze time dynamics of neural networks. The recent line of work e.g. in [1, 2, 5, 12] has highlighted the dynamical systems view in statistical learning by neural networks. We introduce a more coherent framework with a wider ranger of techniques to exploit this view point. This is in fact necessary since the existing approaches are insufficient in our case because of the asymmetry in the associated Gram matrix and its complex eigen-structure. We use the poles of the Laplace transform to analyze the dynamics of the training process in section 4.2.

iv) We discuss the relation of the speed of convergence to the generalization power of the student and show that the teacher may improve the generalization power of the student by speeding up its convergence, hence effectively reducing its capacity.

v) We experimentally demonstrate different aspects of our knowledge transfer framework supported by our theoretical analysis. We exploit the data overlap characterization in Theorem 3.2 using optimal kernel-target alignment [4] to compute kernel embeddings which lead to better knowledge transfer.

2 PROBLEM FORMULATION AND MAIN RESULTS

We study knowledge transfer in an analytically tractable setting: the two layer non–linear model studied in [1, 2, 5, 6]:

$$f(\mathbf{x}) = \sum_{k=1}^{m} \frac{a_k}{\sqrt{m}} \sigma(\mathbf{w}_k^T \mathbf{x}), \qquad (1)$$

Here the weights $\{\mathbf{w}_k\}_{k=1}^{m}$ are the model variables corresponding to m hidden units, $\sigma(.)$ is a real (nonlinear) activation function and the weights $\{a_k\}$ are fixed. While we assume that the student network maintains the form in (1) throughout this paper, the teacher may not assume a definite architecture. However, we often specialize our results for the case that the teacher also takes the form in (1) with a larger number $\bar{m}$ of hidden units.

For training the student, we introduce a general optimization framework that considers knowledge transfer. Given a dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ comprising of n data samples $\{\mathbf{x}_i\}$ and their corresponding labels $\{y_i\}$, our framework is given by

$$\min_{\{\mathbf{w}_k\}} \sum_i (y_i - f(\mathbf{x}_i))^2 + \lambda \sum_i \sum_k \left(\phi^{(k)}(\mathbf{x}_i) - f^{(k)}(\mathbf{x}_i) \right)^2, \quad (2)$$

where $f(.)$ is stated in (1) and $f^{(k)}(\mathbf{x}) = \sigma(\mathbf{w}_k^T \mathbf{x})$ is the corresponding k^{th} hidden feature of the student network. As seen in (2), our framework consists of a least squares optimization problem: $\min \sum (y_i - f(\mathbf{x}_i))^2$ with an additional regularization term incorporating the teacher's knowledge represented by the *privileged knowledge* terms $\phi^{(k)}$. The coefficient $\lambda \geq 0$ is the regularization parameter.

Our analysis considers generic forms of the privileged knowledge functions $\phi^{(k)}$. However, we are particularly interested in a setup where these functions are selected from the hidden neurons of a pre-trained teacher. More precisely, $\phi^{(k)}(\mathbf{x}) = \sigma(\langle \mathbf{w}_k^{\text{teacher}}, \mathbf{x} \rangle)$ where $\mathbf{w}_k^{\text{teacher}}$ is the trained weight of the kth selected unit of the teacher with the architecture in (1). Note that $\left\{\mathbf{w}_k^{\text{teacher}}\right\}$ is a subset of the teacher weights. This case is closely connected to a well-known knowledge distillation (KD) setup empirically studied in e.g. [3].

We study the generic behavior of the gradient descent (GD) algorithm when applied to the optimization in (2). In the spirit of the analysis in [1, 6], shortly explained in Section 4, we carry out an investigation on the dynamics for GD that answers two fundamental questions: *i. What does the (student) network learn? ii. How fast is the convergence by the gradient descent?* The answer to both these questions emerges from the analysis of the dynamics of GD. From the perspective of KD, this approach complements

related recent studies, such as [15] that address similar questions. However, our work is different, as [15] is limited to a single hidden unit ($m = 1$), Sigmoid activation σ and cross-entropy replacing the square-error loss. While convexity plays a major role in [15], our analysis concerns the non-convex setup in (2) with further conditions on initialization. Additionally, our result is applicable to a different regime with a large number m of units and high expression capacity.

3 MAIN RESULTS ON DYNAMICS

3.1 General Linear Systems Theory Framework

The existing analysis of dynamics for neural networks in a series of recent papers [1, 2, 5] is tied centrally to the premise that the behaviour of GD for the optimization can be approximated by a linear dynamics of finite order. To isolate the negligible effect of learning rate μ in GD, it is also conventional to study the case $\mu \to 0$ where GD is alternatively represented by an ordinary differential equation (ODE), known as the gradient flow, with a continuous "time" variable t replacing the iteration number r (being equivalent to the limit of $r\mu$). Let us denote by $\mathbf{f}(t)$ the vector of the output $f(\mathbf{x}_i, t)$ of the network at time t. Then, the theory of linear systems with a finite order suggests the following expression for the evolution of $\mathbf{f}(t)$:

$$\mathbf{f}(t) = \mathbf{f}_\infty + \boldsymbol{\delta}(t), \tag{3}$$

where $\mathbf{f}_\infty$ is a constant and

$$\boldsymbol{\delta}(t) = \mathbf{u}_1 e^{-p_1 t} + \mathbf{u}_2 e^{-p_2 t} + \ldots + \mathbf{u}_d e^{-p_d t}. \tag{4}$$

Here, d is the order of the linear system and complex-valued vectors $\mathbf{u}_1, \ldots, \mathbf{u}_m$ and nonzero complex values $p_1, p_2, \ldots, p_d$ are to be determined by the specifications of the dynamics. The constants $\{p_j \neq 0\}$ are called poles, that also correspond to the singular points of the Laplace transform $\mathbf{F}(s)$ of $\mathbf{f}(t)$ (except for 0, which corresponds to the constant $\mathbf{f}_\infty$ in our formulation). We observe that such a representation may only have a convergence (final) value at $t \to \infty$ if the poles have strictly positive real parts, in which case $\mathbf{f}_\infty$ is the final value. Moreover, the asymptotic rate of convergence is determined by the dominating term in (4), i.e. the smallest value $\Re(p_j)$ with a nonzero vector $\mathbf{u}_j$. We observe that identifying $\mathbf{f}_\infty$ and the dominating term responds to the aforementioned questions of interest. In this paper, we show that these values can be calculated as the number m of hidden units increases.

Definitions: Let us take $\mathbf{w}_k = \mathbf{w}_k(0)$ as the initial values of the weights and define $\mathbf{H}_k = (\sigma'(\mathbf{w}_k^T \mathbf{x}_i)\sigma'(\mathbf{w}_k^T \mathbf{x}_i)\mathbf{x}_i^T \mathbf{x}_j)$ as the kth realization of the "associated gram matrix" where σ' denotes the derivative function of σ (that can be defined in the distribution sense). Further, denote by $\mathbf{f}^k(0)$ the vector of the initial values $f^k(\mathbf{x}_i) = \sigma(\mathbf{w}_k^T \mathbf{x}_i)$ of the kth unit for different data points $\{\mathbf{x}_i\}$ and take $a = \sum_k a_k^2/m$. Finally, take $p_1, p_2, \ldots, p_d$ for $d = n \times m$ as positive values where at $s = -p_i$, the value -1 is the eigenvalue of the matrix $\mathbf{T}(s) = \sum_k \frac{a_k^2}{m} (s\mathbf{I} + \lambda\mathbf{H}_k)^{-1}\mathbf{H}_k$, with $\mathbf{v}^1, \mathbf{v}^2, \ldots, \mathbf{v}^d$ being the corresponding eigenvectors ($\mathbf{T}(s)$ is symmetric).

3.2 What does the student learn?

This result pertains to the first question above, concerning the final value of $\mathbf{f}$. For this, we prove the following result:

THEOREM 3.1. *Suppose that m is large and $\|\boldsymbol{\phi}_k - \mathbf{f}_k(0)\| = O(1/m)$. Under mild conditions (Section 6), it holds that $\lim_{t \to \infty} \mathbf{f}(t) = \mathbf{f}_\infty$ where*

$$\mathbf{f}_\infty = \frac{1}{a + \lambda} \left(a\mathbf{y} + \lambda \sum_k \frac{a_k \boldsymbol{\phi}_k}{\sqrt{m}} \right) + o_m(1). \tag{5}$$

This is an intuitive result: the final output of the student is mixture of the true labels $\mathbf{y}$ and the teacher's provided knowledge vectors $\boldsymbol{\phi}_k$.

Random Privileged Knowledge Setup: Indeed, an interesting case is when the teacher is itself a strong predictor of the labels $\mathbf{b} := \sum_k \frac{a_k \boldsymbol{\phi}_k}{\sqrt{m}} = \mathbf{y} + o_m(1)$, which further yields a perfect prediction by the student. This can be the case when the teacher is a wider network in the form of (1) with $\bar{m}$ hidden neurons and their corresponding weights $\bar{\mathbf{w}}_l^{\text{teacher}}$ and bounded coefficients $\bar{a}_l^{\text{teacher}}$ in the second layer. The results in [6] guarantee that such a network can be perfectly trainable over the samples. Let us consider the case where $\{a_k^{\text{teacher}}, \mathbf{w}_k^{\text{teacher}}\} \subset \{\bar{a}_l^{\text{teacher}}, \bar{\mathbf{w}}_l^{\text{teacher}}\}$ is independently randomly selected. Then, we may invoke Theorem 3.1 by taking $\phi^{(k)}(\mathbf{x}) = \sigma(\langle \mathbf{w}_k^{\text{teacher}}, \mathbf{x} \rangle)$ and $a_k = qa_k^{\text{teacher}}$, where $q = \sqrt{\frac{\bar{m}}{m}}$. Then, we have $\mathbb{E}[\mathbf{b}] = \mathbf{f}^{\text{teacher}}$, $\text{Var}[\mathbf{b}] = O\left(\frac{\bar{m}}{m} - 1\right)$. This shows that for large $\bar{m}$ taking a sufficiently large fraction of the units will introduce negligible harm to the student's solution.

3.3 How fast does the student learn?

Now, we turn our attention to the question of the speed of convergence, for which we have the following result:

THEOREM 3.2. *Define*

$$\mathbf{f}_\infty^{(k)} = \frac{a_k}{\lambda \sqrt{m}} (\mathbf{y} - \mathbf{f}_\infty) + \boldsymbol{\phi}_k.$$

With similar assumptions to Theorem 3.1 (given in Section 6), the dynamics of $\mathbf{f}$ can be written as[1]

$$\mathbf{f}(t) = \mathbf{f}_\infty + \sum_{j=1}^n e^{-p_j t} \alpha_j \mathbf{u}^j + o_m(1),$$

where

$$\alpha_j = \sum_k \frac{a_k^2}{m} \left\langle \mathbf{v}^j, \mathbf{H}_k(p_j \mathbf{I} - \lambda\mathbf{H}_k)^{-1}(\mathbf{f}_\infty^{(k)} - \mathbf{f}^{(k)}(0)) \right\rangle.$$

A straightforward consequence of Theorem 3.2 is that

$$\|\mathbf{f}(t) - \mathbf{f}_\infty\|_2 = O\left(e^{-p_{\min} t}\right),$$

where $p_{\min}$ is the minimum value of p_js. In other words, convergence is linear. In practice, the discrete time process of gradient descent with a step size μ is used. Although our analysis is instead based on the common choice of gradient flow, we remark that with a similar approach, one can show that for a sufficiently small step size, e.g.

---

[1] The rate analysis is in L_1 sense, i.e. $a(t) = b(t) + o_m(1)$ means $\int_0^\infty \|a(t) - b(t)\|_2 dt = o_m(1)$.

$\mu < \frac{1}{2p_{\max}}$ with $p_{\max}$ being the largest of p_js, the convergence rate remains linear:

$$\|\mathbf{f}_n - \mathbf{f}_\infty\|_2 = O\left((1 - \mu p_{\min})^n\right),$$

where with a slight abuse of notation, $\mathbf{f}_n$ denotes the network output in the nth iteration of GD.

From the definition, α_j can be interpreted as an average overlap between a combination of the label vector $\mathbf{y}$ and the knowledge vectors $\boldsymbol{\phi}_k$, and the "spectral" structure of the data as reflected by the vectors $(p_j \mathbf{I} - \lambda \mathbf{H}_k)^{-1} \mathbf{H}_k \mathbf{v}^j$. This is a generalization of the geometric argument in [1] in par with the "data geometry" concept introduced in [15]. We will later use this result in our experiments to improve knowledge transfer by modifying the data geometry of α_j coefficients.

3.4 Further Remarks

The above two results have a number of implications on the knowledge transfer process:

Extreme Cases: First, note that the case $\lambda = 0$ reproduces the results in [6]: The final value simply becomes $\mathbf{y}$ while the poles will become the singular values of the matrix $\mathbf{H} = \sum_k \frac{a_k^2}{m} \mathbf{H}_k$. The other extreme case of $\lambda = \infty$ corresponds to pure transfer from the teacher, where the effect of the first term in (2) becomes negligible and hence the optimization boils down to individually training each hidden unit by ϕ_k. One may then expect the solution of this case to be $f_k = \phi_k$. However, the conditions of the above theorems become difficult to verify, but we shortly present experiments that numerically investigate the corresponding dynamics.

Speed-accuracy trade-off: As previously pointed out, for a finite value of λ, the final value $\mathbf{f}_\infty$ is a weighted average, depending on the quality of $\boldsymbol{\phi}^k$. Defining $e = \|\mathbf{f}_\infty - \mathbf{y}\|$ as the final error, we simply conclude that $e = \frac{\lambda}{a+\lambda} \|\sum_k \frac{a_k \boldsymbol{\phi}_k}{m} - \mathbf{y}\|$, where the term $\|\sum_k \frac{a_k \boldsymbol{\phi}_k}{m} - \mathbf{y}\|$ reflects the quality of teacher in representing the labels. Also for an imperfect teacher the error e monotonically increases with λ. At the same time, we observe that larger λ has a positive effect on the speed of learning.

THEOREM 3.3. *Given, the above definition, it holds that $p_j \geq \lambda \sigma_0$, where σ_0 is the smallest nonzero eigenvalue of all $\mathbf{H}_k$s.*

This sets an intuitive trade-off, where increasing λ, i.e. relying more on the teacher, improves the speed of learning, while magnifying potential teacher's imperfections. Note that σ_0 remains finite and of $O(\sqrt{n})$, even if the number m of $\mathbf{H}_k$s increases.

Effect of data geometry: As we demonstrate in Section 6, the vectors $\mathbf{H}_k(p_j - \lambda \mathbf{H}_k)^{-1} \mathbf{v}_j$ constitute an eigen-basis structure correesoonding to the poles p_j. If the data has a small overlap with the eigen-basis corresponding to small values of p_j, then the values α_j for small poles p_j will drop and the dynamics is mainly identified by the large poles p_j, speeding up the convergence properties. This defines a notion of a suitable geometry for knowledge transfer.

On initializing the student: Finally, the assumption $\|\boldsymbol{\phi}_k - \mathbf{f}^{(k)}(0)\| = O(1/m)$ can be simply satisfied with single hidden layer, where we have $\phi_k(\mathbf{x}) = \sigma(\langle \mathbf{w}_k^{\text{teacher}}, \mathbf{x} \rangle)$ and initializing the weights of the student by that of the teacher $\mathbf{w}_k(0) = \mathbf{w}_k^{\text{teacher}}$ leads to

$\boldsymbol{\phi}_k = \mathbf{f}^{(k)}(0)$. We further numerically investigate the consequences of violating this assumption.

3.5 Consequences on Generalization

In [1], a bound on the generalization of the NN in (1), when trained by GD, is given. Their approach is to show that for GD, the trained weights $\mathbf{w}_k$ and their initial values satisfy

$$\sum_k \|\mathbf{w}_k - \mathbf{w}_k(0)\|^2 \leq \mathbf{y}^T \mathbf{H}^{-1} \mathbf{y} + o(1) \tag{6}$$

They proceed by showing that the family of such neural notworks have a bounded Rademacher complexity and hence the generalization power. The bound in (6) is a natural consequence of the fact that the learning rate of each weight, at each time, is on average proportional to the convergence rate of the network, which is shown to be exponential. In other words, the weights will not have enough time to escape the ball defined by (6). Our analysis of convergence in Theorem 3.2 leads to a similar generalizaton bound for the student network. In fact, as λ increases we get faster convergence by Theorem 3.2 since the weights have less time to update, leading to a tighter bound. This claim is intuitive too: as learning relies more on the teacher, less variation is expected, leading to better generalization. This provides additional insights on the success of knowledge transfer in practice.

4 ANALYSIS AND INSIGHTS

The study in [6] on the dynamics of backpropagation serves as our main source of inspiration, which we review first. The point of departure in this work is to represent the dynamics of BP or gradient descent (GD) for the standard ℓ_2 risk minimization, as in (2) and (1) with $\lambda = 0$. In this case, the associated ODE to GD reads:

$$\frac{d\mathbf{w}_k}{dt}(t) = \frac{a_k}{\sqrt{m}} \mathbf{L}(\mathbf{w}_k(t))(\mathbf{y} - \mathbf{f}(t)), \tag{7}$$

where $\mathbf{y}, \mathbf{f}(t)$ are respectively the vectors of $\{y_k\}$ and $f(\mathbf{x}_k)$, calculated in (1) by replacing $\mathbf{w}_k = \mathbf{w}_k(t)$. Moreover, the matrix $\mathbf{L}(\mathbf{w})$ consists of $\sigma'(\mathbf{w}^T \mathbf{x}_k)\mathbf{x}_k$ as its kth column. While the dynamics in (7) is generally difficult to analyze, we identify two simplifying ingredients in the study of [6]. First, it turns the attention from the dynamics of weights to the dynamics of the function, as reflected by the following relation:

$$\frac{d\mathbf{f}}{dt}(t) = \sum_k \frac{a_k}{\sqrt{m}} \mathbf{L}_k^T \frac{d\mathbf{w}_k}{dt} = \mathbf{H}(t)(\mathbf{y} - \mathbf{f}(t)), \tag{8}$$

where $\mathbf{L}_k = \mathbf{L}_k(t)$ is a short-hand notation for $\mathbf{L}(\mathbf{w}_k(t))$ and $\mathbf{H}(t) = \sum_k a_k^2 \mathbf{L}_k^T \mathbf{L}_k / m$. The second element in the proof can be formulated as follows:

Kernel Hypothesis (KH): In the asymptotic case of $m \to \infty$, the dynamics of $\mathbf{H}(t)$ has a negligible effect, such that it may be replaced by $\mathbf{H}(0)$, resulting to a linear dynamics.

The reason for our terminology of the KH is that under this assumption, the dynamics of BP resembles that of a kernel regularized least squares problem. The investigation in [6] further establishes KH under mild assumptions and further notes that for random initialization of weights $\mathbf{H}(0)$ concentrates on its mean value, denoted by $\mathbf{H}^\infty$.

4.1 Dynamics of Knowledge Transfer

Following the methodology of [6], we proceed by providing the dynamics of the GD algorithm for the optimization problem in (2) with $\lambda > 0$. Direct calculation of the gradient leads us to the following associated ODE for GD:

$$\frac{d\mathbf{w}_k}{dt} = \mathbf{L}_k \left[\frac{a_k}{\sqrt{m}} (\mathbf{y} - \mathbf{f}(t)) + \lambda(\boldsymbol{\phi}^{(k)} - \mathbf{f}^{(k)}) \right], \qquad (9)$$

where $\mathbf{L}_k, \mathbf{y}, \mathbf{f}(t)$ are similar to the previous case in (7). Furthermore, $\mathbf{f}^{(k)}, \boldsymbol{\phi}^{(k)}$ are respectively the vectors of $\{f^{(k)}(\mathbf{x}_i)\}_i$ and $\{\phi^{(k)}(\mathbf{x}_i)\}_i$. We may now apply the methodology of [6] to obtain the dynamics of the features. We also observe that unlike this work, the hidden features $\{\mathbf{f}^{(k)}\}$ explicitly appear in the dynamics:

$$\frac{d\mathbf{f}^{(k)}}{dt}(t) = \mathbf{L}_k^T \frac{d\mathbf{w}_k}{dt} =$$
$$\mathbf{H}_k(t) \left[\frac{a_k}{\sqrt{m}} (\mathbf{y} - \mathbf{f}(t)) + \lambda(\boldsymbol{\phi}^{(k)} - \mathbf{f}^{(k)}(t)) \right], \qquad (10)$$

where $\mathbf{H}_k(t) = \mathbf{L}_k^T \mathbf{L}_k$ and

$$\mathbf{f}(t) = \sum_k \frac{a_k}{\sqrt{m}} \mathbf{f}^{(k)}(t). \qquad (11)$$

This relation will be central in our analysis and we may slightly simplify it by introducing $\boldsymbol{\delta}^{(k)} = \mathbf{f}^{(k)} - \mathbf{f}_\infty^{(k)}$ and $\boldsymbol{\delta} = \mathbf{f} - \mathbf{f}_\infty$. In this case the dynamics in (10) and (11) simplifies to:

$$\frac{d\boldsymbol{\delta}^{(k)}}{dt}(t) = -\mathbf{H}_k(t) \left[\frac{a_k}{\sqrt{m}} \boldsymbol{\delta}(t) + \lambda \boldsymbol{\delta}^{(k)}(t) \right],$$
$$\boldsymbol{\delta}(t) = \sum_k \frac{a_k}{\sqrt{m}} \boldsymbol{\delta}^{(k)}(t). \qquad (12)$$

Finally, we give a more abstract view on the relation in (12) by introducing the block vector $\boldsymbol{\eta}(t)$ where the kth block is given by $\boldsymbol{\delta}^{(k)}(t)$. Then, we may write (12) as $d\boldsymbol{\eta}/dt(t) = -\bar{\mathbf{H}}(t)\boldsymbol{\eta}(t)$, where $\bar{\mathbf{H}}(t)$ is a block matrix with $\mathbf{H}_k(t) \left(\frac{a_k a_l}{m} + \lambda \delta_{k,l} \right)$ as its k, l block ($\delta_{k,l}$ denotes the Kronecker delta function).

4.2 Dynamics Under Kernel Hypothesis: Analysis by Laplace Transform

Now, we follow [6] by simplifying the relation in (10) under the kernel hypothesis, which in this case assumes the matrices $\bar{\mathbf{H}}(t)$ to be fixed to its initial value $\bar{\mathbf{H}} = \bar{\mathbf{H}}(0)$, leading again to a linear dynamics:

$$\boldsymbol{\eta}(t) = e^{-\bar{\mathbf{H}}t} \boldsymbol{\eta}(0). \qquad (13)$$

Despite similarities with the case in [1, 6], the relation in (13) is not simple to analyze due to the asymmetry in $\bar{\mathbf{H}}$ and the complexity of its eigen-structure. For this reason, we proceed by taking the Laplace transform of (12) (assuming $\mathbf{H}_k = \mathbf{H}_k(t) = \mathbf{H}_k(0)$) which after straightforward manipulations gives:

$$\Delta^{(k)}(s) = (s\mathbf{I} + \lambda \mathbf{H}_k)^{-1} \left[\boldsymbol{\delta}^{(k)}(0) - \frac{a_k}{\sqrt{m}} \mathbf{H}_k \Delta(s) \right],$$
$$\Delta(s) = (\mathbf{I} + \mathbf{T}(s))^{-1} \sum_k (s\mathbf{I} + \lambda \mathbf{H}_k)^{-1} \boldsymbol{\delta}^{(k)}(0), \qquad (14)$$

where $\Delta^{(k)}(s)$ and $\Delta(s)$ are respectively the Laplace transforms of $\boldsymbol{\delta}^{(k)}(y)$ and $\boldsymbol{\delta}(t)$.

Hence, $\boldsymbol{\delta}(t)$ is given by taking the inverse Laplace transform of $\Delta(s)$. Note that by construction, $\Delta(s)$ is a rational function, which

shows the finite order of the dynamics. To find the inverse Laplace transform, we only need to find the poles of $\Delta(s)$. These poles can only be either among the eigenvalues of $\mathbf{H}_k$ or the values $-p_k$ where the matrix $\mathbf{I} + \mathbf{T}(s)$ becomes rank deficient. Under Assumption 1 and 2, we may conclude that the poles are only $-p_k$, which gives the result in Theorem 1 and 2. More details of this approach can be found in the Section 6, where the kernel hypothesis for this case is also rigorously proved.

5 EXPERIMENTAL RESULTS

In this section we present validation for the main results in the paper which helps understanding the theorems and reinforces them. We perform our numerical analysis on a commonly-used dataset for validating deep neural models, i.e., CIFAR-10. This dataset is used for the experiments in [1]. As in [1], we only look at the first two classes and set the label $y_i = +1$ if image i belongs to the first class and $y_i = -1$ if it belongs to the second class. The images $\{x_i\}_{i=1}^n$ are normalized such that $||x_i||_2 = 1$ for all $i = 1, \ldots, n$. The weights in our model are initialized as follows:

$$w_i \sim \mathcal{N}(0, k^2 I), \ a_r \sim \text{Unif}(\{-1, 1\}), \forall r \in [m]. \qquad (15)$$

For optimization, we use (full batch) gradient descent with the learning rate η. In our experiments we set $k = 10^{-2}, \eta = 2 \times 10^{-4}$ similar to [1]. In all of our experiments we use 100 hidden neurons for the teacher network and 20 hidden neurons for the student network.
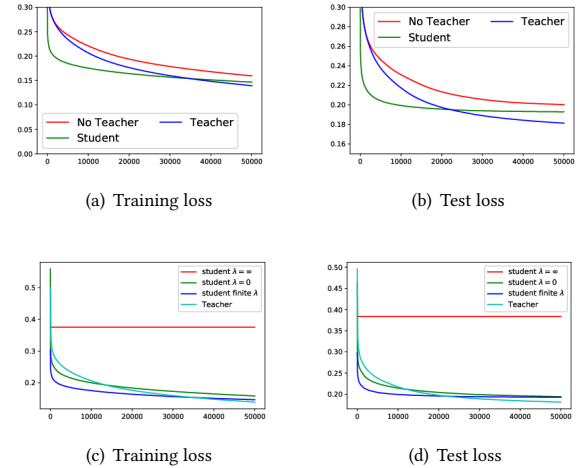

(a) Training loss

(b) Test loss

(c) Training loss

(d) Test loss

Figure 1: Dynamics of knowledge transfer (a,b) and Effect of different regularization λ (c,d).

5.1 Dynamics of knowledge transfer

The experiments in this section show the theoretical justification (in Theorem 3.1 and 3.2) of the experimentally well-studied advantage of a teacher. We study knowledge transfer in different settings. We first consider a finite regularization in Eq. 2 by setting $\lambda = 0.01$. Figures 1(a) and 1(b) show the dynamics of the results in different settings, i) no teacher, i.e., the student is independently trained without access to a teacher, ii) student training, where the student is trained by both the teacher and the true labels according to Eq. 2, and iii) the teacher, trained by only the true labels. For

each setting, we illustrate the training loss and the test loss. The horizontal axes in the plots show the number of iterations of the optimizer. In total, we choose 50000 iterations to be sure of the optimization convergence. This corresponds to the parameter r of the dynamical system proposed in section 3.1. Note that true labels are the same for the teacher and the students. Teacher shows the best performance because of its considerably larger capacity. On the other hand, we observe, i) for the student with access to the teacher its performance is better than the student without access to the teacher. This observation is verified by the result in Theorem 3.1 that states the final performance of the student is a weighted average of the performances of the teacher and of the student with no teacher. This observation is also consistent with the discussion in section 3.4, where the final performance of the student is shown to improve with the teacher. ii) The convergence rate of the optimization is significantly faster for the student with teacher compared to the other alternatives. This confirms the prediction of Theorem 3.2. This experiment implies the importance of a proper knowledge transfer to the student network via the information from the teacher.

In the following we study the effect of the regularization parameter (λ) on the dynamics, when a teacher with a similar structure to the student is utilized. The teacher is wider than the student and randomly selected features of the teacher are used as the knowledge ϕ_k transferred to the student. Specifically, we study two special cases of the generic formulation in Eq. 2 where $\lambda \to 0$ and $\lambda \to \infty$. Figures 1(c) and 1(d) compare these two extreme cases with the student with $\lambda = 0.01$ and the teacher w.r.t. training loss and test loss. We observe that the student with a finite regularization ($\lambda = 0.01$) outperforms the two other students in terms of both convergence rate (optimization speed) and the quality of the results. In particular, when the student is trained with $\lambda \to \infty$ and it is initialized with the weights of the teacher, then the generic loss in Eq. 2 equals 0. This renders the student network to keep its weights unchanged for $\lambda \to \infty$ and the performance remains equal to that of the privileged knowledge $\sum_k \frac{a_k}{\sqrt{m}} \phi_k$ without data labels.

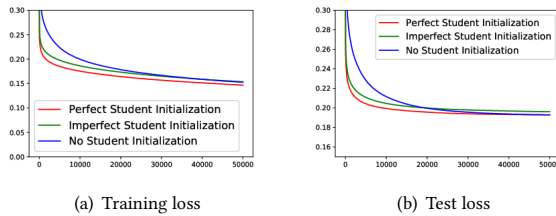(a) Training loss

(b) Test loss

Figure 2: Dynamics of knowledge transfer with perfect and imperfect teacher.

5.2 Dynamics of knowledge transfer with imperfect teacher

In this section, we study the impact of the quality of the teacher on the student network. We consider the student-teacher scenario in three different settings, i) perfect teacher where the student is initialized with the final weights of the teacher and uses the final teacher outputs in Eq. 2, ii) imperfect teacher where the student

is initialized with the intermediate (early) weights of the teacher network and uses the respective intermediate teacher outputs in Eq. 2, and iii) no student initialization where the student is initialized randomly but uses the final teacher outputs. In all the settings, we assume $\lambda = 0.01$.

Figure 2 shows the results for these three settings, respectively w.r.t. training loss and test loss. We observe that initializing and training the student with the perfect (fully trained) teacher yields the best results in terms of both quality (training and test loss) and convergence rate (optimization speed). This observation verifies our theoretical analysis on the importance of initialization of the student with fully trained teacher, as the student should be very close to the teacher.

5.3 Kernel embedding

To provide the teacher and the student with more relevant information and to study the role of the data geometry (Theorem 3.2), we can use properly designed kernel embeddings. Specifically, instead of using the original features for the networks, we could first learn an optimal kernel which is highly aligned with the labels in training data, implicitly improving the combination of $\{\alpha_j\}$ in Theorem 3.2 and then we feed the features induced by that kernel embedding into the networks (both student and teacher).

For this purpose, we employ the method proposed in [4] that develops an algorithm to learn a new kernel from a group of kernels according to a similarity measure between the kernels, namely centered alignment. Then, the problem of learning a kernel with a maximum alignment between the input data and the labels is formulated as a quadratic programming (QP) problem. The respective algorithm is known as *alignf* [4].

Let us denote by K^c the centered variant of a kernel matrix K. To obtain the optimal combination of the kernels (i.e., a weighted combination of some base kernels), [4] suggests the objective function to be centered alignment between the combination of the kernels and yy^T, where y is the true labels vector. By restricting the weights to be non-negative, a QP can be formulated as minimizing $v^T M v - 2v^T a$ w.r.t. $v \in R_+^P$ where P is the number of the base kernels and $M_{kl} = \langle K_k^c, K_l^c \rangle_F$ for $k, l \in [1, P]$, and finally a is a vector wherein $a_i = \langle K_i^c, yy^T \rangle_F$ for $i \in [1, P]$. If v^* is the solution of the QP, then the vector of kernel weights is given by $\mu^* = v^*/\|v^*\|$ [4, 7].

Using this algorithm we learn an optimal kernel based on seven different Gaussian kernels. Then, we need to approximate the kernel embeddings. To do so, we use the Nyström method [19]. Then we feed the approximated embeddings to the neural networks. The results in Figure 3 show that using the kernel embeddings as inputs to the neural networks, helps both teacher and student networks in terms of training loss (Figure 3(a)) and test loss (Figure 3(b)).

5.4 Spectral analysis

Here, we investigate the overlap parameter of different networks, where we compute a simplified but conceptually consistent variant of the overlap parameter α_j in theorem 3.2. For a specific network, we consider the normalized columns of matrix ϕ (as defined in Eq. 2) corresponding to the nonlinear outputs of the hidden neurons, and compute the dot product of each column with the top eigenvectors of H^∞, and take the average. We repeat this for all the columns and
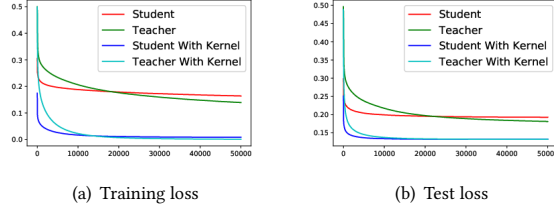
(a) Training loss

(b) Test loss

Figure 3: Dynamics of knowledge transfer with teacher trained using kernel embedding.

depict the histogram. For a small value of λ, the resulting values are approximately equal to $\{\alpha_j\}$ in Theorem 3.2.

Figure 4 shows such histograms for two settings. In Figure 4(a) we compare the overlap parameter for two teachers, one trained partially (imperfect teacher) and the other trained fully (perfect teacher). We observe that the overlap parameter is larger for the teacher trained perfectly, i.e., there is more consistency between its outputs ϕ and the matrix H^∞. This analysis is consistent with the results in Figure 2 which demonstrates the importance of fully trained (perfect) teacher. In Figure 4(b), we show that this improvement is transferred to the student.
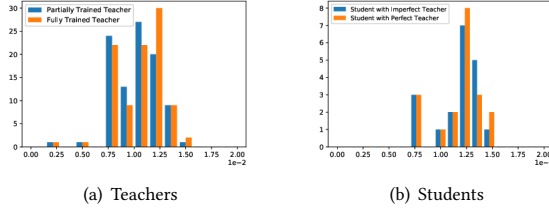

(a) Teachers

(b) Students

Figure 4: Spectral analysis of teachers (4(a)) and students (4(b)) in different settings.

6 PROOFS

6.1 Elaborations on Assumptions

Our analysis will also be built upon a number of assumptions:

ASSUMPTION 1. *Nonzero eigenvalues of the matrices $\{H_k\}$ are all distinct. Note that they are always strictly positive as $\{H_k\}$ are by construction positive semi-definite (psd).*

ASSUMPTION 2. *The values of $p_1, p_2, \ldots, p_d$ are all distinct and different to the eigenvalues of $\{H_k\}$.*

ASSUMPTION 3. *The function σ and its derivative σ' are Lipschitz continuous.*

ASSUMPTION 4. *We assume $m \rightarrow \infty$, such that $\|\phi_k - f_k(0)\| = O(1/\sqrt{m})$ and $\|y - \sum_k \frac{a_k \phi_k}{\sqrt{m}}\| = O(1)$.*

ASSUMPTION 5. *$\|x_i\|$s and a_ks are bounded.*

ASSUMPTION 6. *$\sum_k \frac{\bar{a}_k^2}{m} \|\bar{\phi}_k\|^2$ is bounded.*

Assumption 1-5 are required for Theorem 1. Assumption 1-6 are required for Theorem 3.

In practice, these assumptions are mild, as we elaborate in the following:

1. Assumption 1 and 2 are equivalent to assuming that all eigenvalues of $\bar{H}$, defined in the concluding lines of section 4.1, are distinct. Note that $\bar{H}$ depends on the data and hence inherits its random nature. Accordingly, the event that such a random matrix has an eigenvalue with multiplicity is "zero-measure", i.e almost impossible in reality.
2. Assumption 3 holds for virtually every popular activation functions, including ReLU and Sigmoid.
3. Assumption 4 is equivalent to the scenario in the experiments of section 5.3, referred to as *almost perfect initialization of student by teacher*. Note that it can be rather viewed as a design specification and our experiments verify its merits. The other experiments of section 5.3 identify another consistent (but still weaker) scenario, beyond this assumption.
4. Assumptions 5 can be easily be imposed by normalizing the data. Similar assumptions also exist in [1, 6], but ours are slightly weaker. for example, they merely assume binary values for a_k.
5. Assumption 6 is also very mild. For the student-teacher scenario for example, it requires the features of the teacher to be bounded in the ℓ_2 sense, which is met by standard network architectures.

6.2 Proof of Theorem 1 and 2

We continue the discussion in (12 of paper) and remind that $\sigma_{\max}$ is the largest singular value of matrices and $\|\|$ means the 2-norm of vectors. Note that the values $s = -p_i$ correspond to the points where $\det(I + T(s)) = 0$. We also observe that these values correspond to the negative of eigenvalues of the matrix $\bar{H}(t = 0)$. We conclude that under assumption 2, the eigenvalues of $\bar{H} = \bar{H}(t = 0)$ are distinct and strictly positive, hence this matrix is diagonalizable. Now, we write $\bar{H}(t) = \bar{H} + \Delta\bar{H}(t)$ and state the following lemma:

LEMMA 6.1. *Suppose that $\bar{H}$ is a diagonalizable matrix with strictly positive eigenvalues and denote its smallest eigenvalue by p. Take $\Delta\bar{H}(t)$ as a matrix valued function of the continuous variable t such that for a given fixed value of t*

$$q = q(t) = \frac{\sup_{\tau \in [0\ t]} \sigma_{\max}(\Delta\bar{H}(\tau))}{p} < 1$$

Let $\eta(t)$ denote the solution to $d\eta/dt(t) = -\bar{H}(t)\eta(t)$, with $\bar{H}(t) = \bar{H} + \Delta\bar{H}(t)$. Then,

$$\int_0^t \left\| \eta(\tau) - e^{-\bar{H}\tau}\eta(0) \right\| d\tau \leq \frac{q\|\eta(0)\|}{p(1-q)}$$

PROOF. Consider the iteration $\eta^{r+1} = \mathcal{T}\eta^r$ that generates a sequence of function functions $\eta^r(t)$ for $r = 0, 1, \ldots$ where $\eta^r(0) = e^{-\bar{H}t}\eta(0)$ and $\eta' = \mathcal{T}\eta$ is the solution to

$$\frac{d}{dt}\eta'(t) = -\bar{H}\eta'(t) - \Delta\bar{H}(t)\eta(t) \qquad (16)$$

with $\eta'(0) = \eta(0)$, which can also be written as

$$\eta'(t) = e^{-\bar{H}t}\eta(0) - \int_0^t e^{-\bar{H}(t-\tau)}\Delta\bar{H}(\tau)\eta(\tau)d\tau \qquad (17)$$

We observe that $\mathcal{T}$ on the interval $[0\ t]$ is a contraction map under L_1 norm as we have

$$\Delta\boldsymbol{\eta}'(t) = -\int_0^t e^{-\bar{\mathbf{H}}(t-\tau)}\Delta\bar{\mathbf{H}}(\tau)\Delta\boldsymbol{\eta}(\tau)\mathrm{d}\tau \qquad (18)$$

and hence by the triangle inequality , we get

$$\|\Delta\boldsymbol{\eta}'(t)\|_2 \leq \sup_{\tau\in[0\ t]} \sigma_{\max}(\Delta\bar{\mathbf{H}}(\tau)) \times$$
$$\int_0^t e^{-p(t-\tau)}\|\Delta\boldsymbol{\eta}(\tau)\|_2\mathrm{d}\tau \qquad (19)$$

we conclude that

$$\int_0^t \|\Delta\boldsymbol{\eta}'(\tau)\|_2\mathrm{d}t \leq$$

$$\sup_{\tau\in[0\ t]} \sigma_{\max}(\Delta\bar{\mathbf{H}}(\tau)) \times \int_0^t \frac{1-e^{-p(t-\tau)}}{p}\|\Delta\boldsymbol{\eta}(\tau)\|_2$$

$$\leq q\int_0^t \|\Delta\boldsymbol{\eta}(\tau)\|_2\mathrm{d}t$$

which shows that $\mathcal{T}$ is a contraction. Then, from Banach fixed-point theorem we conclude that $\boldsymbol{\eta}^r$ converges uniformly on the interval $[0\ t]$ to the fixed-point $\boldsymbol{\eta}$ of $\mathcal{T}$, which coincides with the solution of (14 in paper). Moreover,

$$\int_0^t \|\boldsymbol{\eta}-\boldsymbol{\eta}^0(\tau)\|_2\mathrm{d}\tau \leq \frac{\int_0^t \|\boldsymbol{\eta}^1(\tau)-\boldsymbol{\eta}^0(\tau)\|_2\mathrm{d}\tau}{1-q}$$

Now, we observe that

$$\boldsymbol{\eta}^1(t) - \boldsymbol{\eta}^0(t) = -\int_0^t e^{-\bar{\mathbf{H}}(t-\tau)}\Delta\bar{\mathbf{H}}(\tau)e^{-\bar{\mathbf{H}}(\tau)}\boldsymbol{\eta}(0)\mathrm{d}\tau$$

Hence,

$$\|\boldsymbol{\eta}^1(t) - \boldsymbol{\eta}^0(t)\| \leq \int_0^t e^{-p(t-\tau)}\sigma_{\max}(\Delta\bar{\mathbf{H}}(\tau))e^{-p\tau}\|\boldsymbol{\eta}(0)\|\mathrm{d}\tau$$

$$\leq te^{-pt}\sup_{\tau\in[0\ t]}\sigma_{\max}(\Delta\bar{\mathbf{H}}(\tau))\|\boldsymbol{\eta}(0)\|$$

and

$$\int_0^t \|\boldsymbol{\eta}^1(\tau)-\boldsymbol{\eta}^0(\tau)\|\mathrm{d}\tau \leq$$

$$\sup_{\tau\in[0\ t]}\sigma_{\max}(\Delta\bar{\mathbf{H}}(\tau))\|\boldsymbol{\eta}(0)\|\int_0^t \tau e^{-p\tau}\mathrm{d}\tau \leq \frac{q}{p}\|\boldsymbol{\eta}(0)\|$$

which completes the proof. □

Now, we state two results that connect $\sigma_{\max}(\Delta\mathbf{H})$ to the change of $\mathbf{w}_k(t)$:

LEMMA 6.2. *Under Assumption 3, the following relation holds:*

$$\sigma_{\max}(\bar{\mathbf{H}}(t)) \leq \sqrt{2}\sqrt{\lambda^2 + \left(\sum_k \frac{a_k^2}{m}\right)^2}\max_k \sigma_{\max}(\Delta\mathbf{H}_k(t)) \qquad (20)$$

where $\Delta\mathbf{H}_k(t) = \mathbf{H}_k(t) - \mathbf{H}_k(0)$.

PROOF. Take an arbitrary block vector $\boldsymbol{\eta} = [\boldsymbol{\delta}_k]_k$ with $\|\boldsymbol{\eta}\| = 1$ and note that

$$\|\Delta\bar{\mathbf{H}}(t)\boldsymbol{\eta}\|^2 = \sum_k \|\Delta\mathbf{H}_k(t)(\frac{a_k}{\sqrt{m}}\boldsymbol{\delta} + \lambda\boldsymbol{\delta}_k)\|^2$$

$$\leq 2\max_k \sigma_{\max}^2(\Delta\mathbf{H}_k(t))\sum_k \left(\frac{a_k^2}{m}\|\boldsymbol{\delta}\|^2 + \lambda\|\boldsymbol{\delta}_k\|^2\right)$$

$$= 2\max_k \sigma_{\max}^2(\Delta\mathbf{H}_k(t))\left(\|\boldsymbol{\delta}\|^2\sum_k \frac{a_k^2}{m} + \lambda^2\right),$$

where $\boldsymbol{\delta} = \sum_k \frac{a_k}{\sqrt{m}}\boldsymbol{\delta}_k$. We obtain the desired result by observing that

$$\|\boldsymbol{\delta}\|^2 = \left\|\sum_k \frac{a_k}{\sqrt{m}}\boldsymbol{\delta}_k\right\|^2 \leq \left(\sum_k \frac{a_k^2}{m}\right)\sum_k \|\boldsymbol{\delta}_k\|^2 = \sum_k \frac{a_k^2}{m}.$$

□

Next, we show

LEMMA 6.3. *We have*

$$\sigma_{\max}(\Delta\mathbf{H}_k(t)) \leq L^2\sigma_x^2\max_i \|\mathbf{x}_i\|^2 \times$$

$$\|\mathbf{w}_k(t) - \mathbf{w}_k(0)\|(\|\mathbf{w}_k(t) - \mathbf{w}_k(0)\| + 2\|\mathbf{w}_k(0)\|)$$

where σ_x is the maximal eigenvalue of the the data matrix $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2 \ldots \mathbf{x}_n]$ and L is the largest of the Lipschitz constants of σ, σ'.

PROOF. Note that since $\Delta\mathbf{H}_k$ is symmetric, we have (e.g. by eigendecomposition)

$$\sigma_{\max}(\Delta\mathbf{H}_k(t)) = \max_{\boldsymbol{\delta}|\|\boldsymbol{\delta}\|=1} |\boldsymbol{\delta}^T\Delta\mathbf{H}_k(t)\boldsymbol{\delta}|$$

Taking an arbitrary normalized $\boldsymbol{\delta}$, we observe that

$$\boldsymbol{\delta}^T\Delta\mathbf{H}_k(t)\boldsymbol{\delta} = \left|\|\mathbf{L}(\mathbf{w}_k(t))\boldsymbol{\delta}\|^2 - \|\mathbf{L}(\mathbf{w}_k(0))\boldsymbol{\delta}\|^2\right|$$

On other hand,

$$\mathbf{L}(\mathbf{w}_k(t))\boldsymbol{\delta} = \mathbf{L}(\mathbf{w}_k(0))\boldsymbol{\delta} + \sum_i \mathbf{x}_i\sigma_i\delta_i,$$

where $\sigma_i = \sigma'(\mathbf{w}_k(t)^T\mathbf{x}_i) - \sigma'(\mathbf{w}_k(0)^T\mathbf{x}_i)$. Hence,

$$\boldsymbol{\delta}^T\Delta\mathbf{H}_k(t)\boldsymbol{\delta} \leq \left\|\sum_i \mathbf{x}_i\sigma_i\delta_i\right\|^2 + 2\left\|\sum_i \mathbf{x}_i\delta_i\sigma_i\right\|\|\mathbf{L}(\mathbf{w}_k(0))\boldsymbol{\delta}\|$$

We also observe that

$$\left\|\sum_i \mathbf{x}_i\sigma_i\delta_i\right\| \leq \sigma_x\sqrt{\sum_i \delta_i^2\sigma_i^2}$$

and from Lipschitz continuity,

$$\sigma_i^2 \leq L^2\langle\mathbf{x}_i, \mathbf{w}_k(t) - \mathbf{w}_k(0)\rangle^2 \leq L^2\|\mathbf{x}_i\|^2\|\mathbf{w}_k(t) - \mathbf{w}_k(0)\|^2$$

We conclude that

$$\left\|\sum_i \mathbf{x}_i\sigma_i\delta_i\right\| \leq L\sigma_x\|\mathbf{w}_k(t) - \mathbf{w}_k(0)\| \times \max_i \|\mathbf{x}_i\|$$

Similarly, we obtain

$$\|\mathbf{L}(\mathbf{w}_k(0))\boldsymbol{\delta}\| \le L\sigma_x \times \max_i \|\mathbf{x}_i\|$$

which completes the proof. □

We finally connect the magnitude of the change $\mathbf{w}_k(t) - \mathbf{w}_k(0)$ to $\boldsymbol{\delta}_k$:

LEMMA 6.4. *With the same definitions as in Lemma 3, we have*

$$\|\mathbf{w}_k(t) - \mathbf{w}_k(0)\| \le \frac{|a_k|}{\sqrt{m}} L\sigma_x \max_i \|\mathbf{x}_i\| \int_0^t \|\boldsymbol{\delta}_k(\tau)\| d\tau \qquad (21)$$

PROOF. Note that $\mathbf{w}_k(t) - \mathbf{w}_k(0) = -\frac{a_k}{\sqrt{m}} \int_0^t \mathbf{L}(\mathbf{w}_k(\tau))\boldsymbol{\delta}_k(\tau)d\tau$ and hence $\|\mathbf{w}_k(t) - \mathbf{w}_k(0)\| \le \frac{|a_k|}{\sqrt{m}} \int_0^t \|\mathbf{L}(\mathbf{w}_k(\tau))\boldsymbol{\delta}_k(\tau)\|d\tau$ From a similar argument as in Lemma 3, we have

$$\|\mathbf{L}(\mathbf{w}_k(\tau))\boldsymbol{\delta}_k(\tau)\| \le L\sigma_x \max_i \|\mathbf{x}_i\|\|\boldsymbol{\delta}_k(\tau)\|$$

which completes the proof. □

We may now proceed to the proof of Theorem 1 and 2. Define

$$T = \{t \mid \forall \tau \in [0\ t]; \ q(\tau) < \frac{1}{2}\}$$

Note that T is nonempty as $0 \in T$ and open since q is continuous. We show that for sufficiently large m, $T = [0, \infty)$. Otherwise T is an open interval $[0\ t_0)$ where $q(t_0) = \frac{1}{2}$. For any $t \in T$, we have from Lemma 1

$$A = \int_0^t \|\boldsymbol{\eta}(\tau)\|d\tau \le \left(\frac{1 - e^{-pt}}{p} + \frac{q}{p(1-q)}\right) \|\boldsymbol{\eta}(0)\| \le \frac{\|\boldsymbol{\eta}(0)\|}{p(1-q)}$$

Denote $b = \max_k |a_k|$ and $B = \max_i \|\mathbf{x}_i\|$. We further define $C = \frac{\sqrt{2}}{p} L^3 \sigma_B^3 x^3 b \sqrt{\lambda^2 + a^2}$. Then Lemma 2,3 and 4 give us $q \le \frac{CA}{\sqrt{m}}$. We conclude that

$$q(t) \le \frac{C\|\boldsymbol{\eta}(0)\|}{p(1 - q(t))\sqrt{m}}$$

Note that by Assumption 4, we have

$$\boldsymbol{\delta}_k(0) = \frac{a_k}{(\lambda + a)\sqrt{m}} (\mathbf{y} - \sum_k \frac{a_k \boldsymbol{\phi}_k}{\sqrt{m}}) + \boldsymbol{\phi}_k - \mathbf{f}_k(0) = O(1/\sqrt{m})$$

and hence $\boldsymbol{\eta}(0) = O(1)$. This shows that there exists a constant C_0 such that $q(t)(1 - q(t)) \le \frac{C_0}{\sqrt{m}}$ for $t \in T$. But for large values of m this is in contradiction to $q(t_0) = 1/2$. Hence, for such values t_0 does not exist and $q(t) < 1/2$ for all t. We conclude that for sufficiently large values of m we have $q(t) < 3C_0/\sqrt{m}$ for all t. Then, according to Lemma 1 and the monotone convergence theorem, we have

$$\int_0^\infty \left\|\boldsymbol{\eta}(\tau) - e^{-\bar{\mathbf{H}}\tau}\boldsymbol{\eta}(0)\right\| d\tau = O(\frac{1}{\sqrt{m}})$$

This shows that

$$\lim_{t \to \infty} \|\boldsymbol{\eta}(t) - e^{-\bar{\mathbf{H}}t}\boldsymbol{\eta}(0)\| = 0$$

Note that as $\bar{\mathbf{H}}$ is diagonalizable and has strictly positive eigenvalues, we get that

$$\lim_{t \to \infty} e^{-\bar{\mathbf{H}}t}\boldsymbol{\eta}(0) = 0,$$

which further leads to

$$\lim_{t \to \infty} \boldsymbol{\eta}(t) = 0,$$

This proves Theorem 1. For Theorem 2, we see that

$$\boldsymbol{\eta}(t) = e^{-\bar{\mathbf{H}}t}\boldsymbol{\eta}(0) + O(1/\sqrt{m})$$

It suffices to show that the expression in Theorem 2 coincides with $e^{-\bar{\mathbf{H}}t}\boldsymbol{\eta}(0)$. This is simple to see through the following lemma:

LEMMA 6.5. *According to Assumption 2, the right and left eigenvectors of $\bar{\mathbf{H}}$ corresponding to p_j are respectively given by vectors $\boldsymbol{\eta}_j^r = [\mathbf{v}_k^j]_k$ and $\boldsymbol{\eta}_j^l = [\mathbf{u}_k^j]_k$, where $\mathbf{v}_k^j = \frac{a_k}{\sqrt{m}} (p_j \mathbf{I} - \lambda \mathbf{H}_k)^{-1} \mathbf{H}_k \mathbf{v}^j$ and $\mathbf{u}_k^j = \frac{a_k}{\sqrt{m}} (p_j \mathbf{I} - \lambda \mathbf{H}_k)^{-1} \mathbf{u}^j$. Moreover, $\mathbf{v}^j = \sum_k \frac{a_k}{\sqrt{m}} \mathbf{v}_k^j$.*

PROOF. According to the definition of $\bar{\mathbf{H}}$, we have that

$$\mathbf{H}_k(\frac{a_k}{\sqrt{m}}\mathbf{v} + \lambda \mathbf{v}_k^j) = p\mathbf{v}_k^j$$

where $\mathbf{v} = \sum_k \frac{a_k}{\sqrt{m}} \mathbf{v}_k^j$ which gives $\mathbf{v}_k^j = \frac{a_k}{\sqrt{m}} (p_j \mathbf{I} - \lambda \mathbf{H}_k)^{-1} \mathbf{H}_k \mathbf{v}$ Replacing this expression in the definition of $\mathbf{v}$ shows that $\mathbf{v} = \mathbf{v}^j$. The case for $\mathbf{u}_k^j$ is similarly proved. □

Theorem 2 simply follows by replacing the result of Lemma 5 to the eigen-decomposition of $e^{-\bar{\mathbf{H}}t}$:

$$e^{-\bar{\mathbf{H}}t} = \sum_j e^{-p_j t} \mid \boldsymbol{\eta}_j^r \rangle \langle \boldsymbol{\eta}_j^l \mid$$

6.3 Proof of Theorem 3

Note that for $p = p_j$ there exists an eigen vector $\mathbf{v}$ of $\mathbf{T}(-p)$ such that $\mathbf{v}^T \mathbf{T}(-p)\mathbf{v} = -1$. This leads to $-1 = \sum_k \frac{a_k^2}{m} \mathbf{v}^T (-p\mathbf{I} + \lambda \mathbf{H}_k) \mathbf{H}_k \mathbf{v}$. If the result does not hold, we have that $-p\mathbf{I} + \lambda \mathbf{H}_k \succeq \mathbf{0}$. Hence, the right hands side is non-negative, as $\mathbf{H}_k$ is also psd and the two matrices commute, leading to a contradiction. This proves the result.

7 CONCLUSIONS

We give a theoretical analysis of knowledge transfer for non–linear neural networks in the model and regime of [1, 2, 5] which yields insights on both privileged information and knowledge distillation paradigms. We provide results for both what is learnt by the student and on the speed of convergence. We further provide a discussion about the effect of knowledge transfer on generalization. Our numerical studies further confirm our theoretical findings on the role of data geometry and knowledge transfer in the final performance of student.

ACKNOWLEDGMENTS

REFERENCES

[1] Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. 2019. Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. 322–332.

[2] Yuan Cao and Quanquan Gu. 2019. Generalization Bounds of Stochastic Gradient Descent for Wide and Deep Neural Networks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*. 10835–10845.

[3] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. 2017. Learning efficient object detection models with knowledge distillation. In *Advances in Neural Information Processing Systems*. 742–751.

[4] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. 2012. Algorithms for Learning Kernels Based on Centered Alignment. *J. Mach. Learn. Res.* 13 (March 2012), 795–828.

[5] Simon S. Du and Wei Hu. 2019. Width Provably Matters in Optimization for Deep Linear Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. 1655–1664.

[6] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. 2018. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054* (2018).

[7] Mehmet Gönen and Ethem Alpayd. 2011. Multiple Kernel Learning Algorithms. *J. Mach. Learn. Res.* 12 (July 2011), 2211–2268.

[8] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the Knowledge in a Neural Network. *CoRR* abs/1503.02531 (2015).

[9] Arthur Jacot, Franck Gabriel, and Clement Hongler. 2018. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.). Curran Associates, Inc., 8571–8580.

[10] Yoon Kim and Alexander M. Rush. 2016. Sequence-Level Knowledge Distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 1317–1327. https://doi.org/10.18653/v1/D16-1139

[11] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. 2016. Unifying distillation and privileged information. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

[12] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. 2019. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. *arXiv preprint arXiv:1902.06015* (2019).

[13] Yiteng Pan, Fazhi He, and Haiping Yu. 2019. A novel Enhanced Collaborative Autoencoder with knowledge distillation for top-N recommender systems. *Neurocomputing* 332 (2019), 137–148. https://doi.org/10.1016/j.neucom.2018.12.025

[14] Dmitry Pechyony and Vladimir Vapnik. 2010. On the Theory of Learning with Privileged Information. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2* (Vancouver, British Columbia, Canada) *(NIPS'10)*. Curran Associates Inc., Red Hook, NY, USA, 1894–1902.

[15] Mary Phuong and Christoph Lampert. 2019. Towards Understanding Knowledge Distillation. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, Long Beach, California, USA, 5142–5151.

[16] Vladimir Vapnik and Rauf Izmailov. 2015. Learning Using Privileged Information: Similarity Control and Knowledge Transfer. *Journal of Machine Learning Research* 16, 61 (2015), 2023–2049. http://jmlr.org/papers/v16/vapnik15b.html

[17] Vladimir Vapnik and Rauf Izmailov. 2017. Knowledge transfer in SVM and neural networks. *Ann. Math. Artif. Intell.* 81, 1-2 (2017), 3–19.

[18] Vladimir Vapnik and Akshay Vashist. 2009. A new learning paradigm: Learning using privileged information. *Neural Networks* 22, 5-6 (2009), 544–557.

[19] Christopher K. I. Williams and Matthias Seeger. 2001. Using the Nyström Method to Speed Up Kernel Machines. In *Advances in Neural Information Processing Systems 13*, T. K. Leen, T. G. Dietterich, and V. Tresp (Eds.). MIT Press, 682–688.

[20] Zhiyuan Xu, Kun Wu, Zhengping Che, Jian Tang, and Jieping Ye. 2020. Knowledge Transfer in Multi-Task Deep Reinforcement Learning for Continuous Control. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*.

[21] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. 2017. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4133–4141.

[22] Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. 2017. Visual relationship detection with internal and external linguistic knowledge distillation. In *Proceedings of the IEEE international conference on computer vision*. 1974–1982.