Neural Spectrum Alignment: Empirical Study

Dmitry Kopitkov[1][0000−0001−7797−0468] and
Vadim Indelman[2][0000−0002−1863−3442]

[1] Technion Autonomous Systems Program (TASP), Technion - Israel Institute of Technology, Haifa 32000, Israel
[2] Department of Aerospace Engineering, Technion - Israel Institute of Technology, Haifa 32000, Israel
{dimkak,vadim.indelman}@technion.ac.il

Abstract. Expressiveness and generalization of deep models was recently addressed via the connection between neural networks (NNs) and kernel learning, where first-order dynamics of NN during a gradient-descent (GD) optimization were related to *gradient similarity* kernel, also known as Neural Tangent Kernel (NTK) [9]. In the majority of works this kernel is considered to be time-invariant [9,13]. In contrast, we empirically explore these properties along the optimization and show that in practice *top* eigenfunctions of NTK align toward the target function learned by NN which improves the overall optimization performance. Moreover, these *top* eigenfunctions serve as basis functions for NN output - a function represented by NN is spanned almost completely by them for the entire optimization process. Further, we study how learning rate decay affects the *neural spectrum*. We argue that the presented phenomena may lead to a more complete theoretical understanding behind NN learning.

Keywords: Deep Learning · Neural Tangent Kernel · Kernel Learning.

1 Introduction

Understanding expressiveness and generalization of deep models is essential for robust performance of NNs. Recently, the optimization analysis for a general NN architecture was related to *gradient similarity* kernel [9], whose properties govern NN expressivity level, generalization and convergence rate. Under various considered conditions [9,13], this NN kernel converges to its steady state and is invariant along the entire optimization, which significantly facilitates the analyses of Deep Learning (DL) theory [9,13,2,1].

Yet, in a typical realistic setting the *gradient similarity* kernel is far from being constant, as we empirically demonstrate in this paper. Particularly, during training its spectrum aligns towards the target function that is learned by NN, which improves the optimization convergence rate [1,15]. Furthermore, we show that these *gradient similarity* dynamics can also explain the expressive superiority of deep NNs over more shallow models. Hence, we argue that understanding the *gradient similarity* of NNs beyond its time-invariant regime is a must for full comprehension of NN expressiveness power.

To encourage the onward theoretical research of the kernel, herein we report several strong empirical phenomena and trends of its dynamics. To the best of our knowledge, these trends neither were yet reported nor they can be explained by DL theory developed so far. To this end, in this paper we perform an empirical investigation of fully-connected (FC) NN, its *gradient similarity* kernel and the corresponding Gramian at training data points during the entire period of a typical learning process. Our main empirical contributions are:

(a) We show that Gramian serves as a NN memory, with its *top* eigenvectors changing to align with the learned target function. This improves the optimization performance since the convergence rate along kernel *top* eigenvectors is typically higher.
(b) During the entire optimization NN output is located inside a sub-space spanned by these *top* eigenvectors, making the eigenvectors to be a basis functions of NN.
(c) Deeper NNs demonstrate a stronger alignment, which may explain their expressive superiority. In contrast, shallow wide NNs with a similar number of parameters achieve a significantly lower alignment level and a worse optimization performance.
(d) We show additional trends in kernel dynamics as a consequence of learning rate decay, demonstrating that the information of the target function is spread along bigger number of *top* eigenvectors after each decay.
(e) Experiments over various FC architectures, real-world datasets, *supervised* and *unsupervised* learning algorithms and number of popular optimizers were performed. All experiments showed the mentioned above spectrum alignment.

The paper is structured as follows. In Section 2 we define necessary notations. In Section 3 we relate *gradient similarity* with Fisher information matrix (FIM) of NN and in Section 4 we provide more insight about NN dynamics on L2 loss example. In Section 5 the related work is described and in Section 6 we present our main empirical study. Conclusions are discussed in Section 7. Further, additional derivations and experiments are placed in Appendix [12].

2 Notations

Consider a NN $f_\theta(X) : \mathbb{R}^d \to \mathbb{R}$ with a parameter vector θ, a typical sample loss ℓ and an empirical loss L, training samples $D = \left[\boldsymbol{\mathcal{X}} = \{X^i \in \mathbb{R}^d\}, \boldsymbol{\mathcal{Y}} = \{Y^i \in \mathbb{R}\}\right]$, $i \in [1, \dots, N]$ and loss gradient $\nabla_\theta L$:

$$L(\theta, D) = \frac{1}{N}\sum_{i=1}^{N} \ell\left[Y^i, f_\theta(X^i)\right], \quad \nabla_\theta L(\theta, D) = \frac{1}{N}\sum_{i=1}^{N} \ell'\left[Y^i, f_\theta(X^i)\right] \cdot \nabla_\theta f_\theta(X^i), \tag{1}$$

where $\ell'[Y, f_\theta(X)] \triangleq \nabla_{f_\theta} \ell[Y, f_\theta(X)]$. The above formulation can be extended to include *unsupervised* learning methods in [11] by eliminating labels $\boldsymbol{\mathcal{Y}}$ from the equations. Further, techniques with a model $f_\theta(X)$ returning multidimensional outputs are out of scope for this paper, to simplify the formulation.

Consider a GD optimization with learning rate δ, where parameters change at each discrete optimization time t as $d\theta_t \triangleq \theta_{t+1} - \theta_t = -\delta \cdot \nabla_\theta L(\theta_t, D)$. Further, a model output change at any X according to first-order Taylor approximation is:

$$df_{\theta_t}(X) \triangleq f_{\theta_{t+1}}(X) - f_{\theta_t}(X) \approx -\frac{\delta}{N} \sum_{i=1}^{N} g_t(X, X^i) \cdot \ell' \left[Y^i, f_{\theta_t}(X^i)\right], \quad (2)$$

where $g_t(X, X') \triangleq \nabla_\theta f_{\theta_t}(X)^T \cdot \nabla_\theta f_{\theta_t}(X')$ is a *gradient similarity* - the dot-product of gradients at two different input points also known as NTK [9].

In this paper we mainly focus on optimization dynamics of f_θ at training points. To this end, define a vector $\bar{f}_t \in \mathbb{R}^N$ with i-th entry being $f_{\theta_t}(X^i)$. According to Eq. (2) the discrete-time evolution of f_θ at testing and training points follows:

$$df_{\theta_t}(X) \approx -\frac{\delta}{N} \cdot g_t(X, \boldsymbol{\mathcal{X}}) \cdot \bar{m}_t, \quad d\bar{f}_t \triangleq \bar{f}_{t+1} - \bar{f}_t \approx -\frac{\delta}{N} \cdot G_t \cdot \bar{m}_t, \quad (3)$$

where $G_t \triangleq g_t(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}})$ is a $N \times N$ Gramian with entries $G_t(i,j) = g_t(X^i, X^j)$ and $\bar{m}_t \in \mathbb{R}^N$ is a vector with the i-th entry being $\ell' \left[Y^i, f_{\theta_t}(X^i)\right]$.

Likewise, denote eigenvalues of G_t, sorted in decreasing order, by $\{\lambda_i^t\}_{i=1}^N$, with $\lambda_{max}^t \triangleq \lambda_1^t$ and $\lambda_{min}^t \triangleq \lambda_N^t$. Further, notate the associated orthonormal eigenvectors by $\{\bar{v}_i^t\}_{i=1}^N$. Note that $\{\lambda_i^t\}_{i=1}^N$ and $\{\bar{v}_i^t\}_{i=1}^N$ also represent estimations of eigenvalues and eigenfunctions of the kernel $g_t(X, X')$ (see Appendix A for more details). Below we will refer to large and small eigenvalues and their associated eigenvectors by *top* and *bottom* terms respectively.

Eq. (3) describes the first-order dynamics of GD learning, where $\bar{m}_t$ is a functional derivative of any considered loss L, and the global optimization convergence is typically associated with it becoming a zero vector, due to Euler-Lagrange equation of L. Further, G_t translates a movement in θ-space into a movement in a space of functions defined on $\boldsymbol{\mathcal{X}}$.

3 Relation to Fisher Information Matrix

NN Gramian can be written as $G_t = A_t^T A_t$ where A_t is $|\theta| \times N$ Jacobian matrix with i-th column being $\nabla_\theta f_{\theta_t}(X^i)$. Moreover, $F_t = A_t A_t^T$ is known as the empirical FIM of NN[3] [14,10] that approximates the second moment of model gradients $\frac{1}{N} F_t \approx \mathbb{E}_X \left[\nabla_\theta f_{\theta_t}(X) \nabla_\theta f_{\theta_t}(X)^T\right]$. Since F_t is dual of G_t, both matrices share same non-zero eigenvalues $\{\lambda_i^t \neq 0\}$. Furthermore, for each λ_i^t the respectful eigenvector $\bar{\omega}_i^t$ of F_t is associated with appropriate $\bar{v}_i^t$ - they are left and right singular vectors of A_t respectively. Moreover, change of θ_t along the direction $\bar{\omega}_i^t$ causes a change to $\bar{f}_t$ along $\bar{v}_i^t$ (see Appendix C for the proof). Therefore, spectrums of G_t and F_t describe principal directions in function space and θ-space respectively, according to which $\bar{f}_t$ and θ_t are changing during the optimization. Based on the above, in Section 5 we relate some known properties of F_t towards G_t.

[3] In some papers [17] FIM is also referred to as a Hessian of NN, due to the tight relation between F_t and the Hessian of the loss (see Appendix B for details)

4 Analysis of L2 Loss For Constant Gramian

To get more insight into Eq. (3), we will consider L2 loss with $\ell\left[Y^i, f_\theta(X^i)\right] = \frac{1}{2}\left[f_\theta(X^i) - Y^i\right]^2$. In such a case we have $\bar{m}_t = \bar{f}_t - \bar{y}$, with $\bar{y}$ being a vector of labels. Assuming G_t to be fixed along the optimization (see Section 5 for justification), NN dynamics can be written as:

$$\bar{f}_t = \bar{f}_0 - \sum_{i=1}^{N}\left[1 - \left[1 - \frac{\delta}{N}\lambda_i\right]^t\right] < \bar{v}_i, \bar{m}_0 > \bar{v}_i, \tag{4}$$

$$\bar{m}_t = \sum_{i=1}^{N}\left[1 - \frac{\delta}{N}\lambda_i\right]^t < \bar{v}_i, \bar{m}_0 > \bar{v}_i. \tag{5}$$

Full derivation and extension for dynamics at testing points appear in Appendices D-E. Under the stability condition $\delta < \frac{2N}{\lambda_{max}}$ that satisfies $\lim_{t\to\infty}\left[1 - \frac{\delta}{N}\lambda_i\right]^t = 0$, the above equations can be viewed as a transmission of a signal from $\bar{m}_0 = \bar{f}_0 - \bar{y}$ into our model $\bar{f}_t$. At each iteration $\bar{m}_t$ is decreased along each $\{\bar{v}_i : \lambda_i \neq 0\}$ and the same information decreased from $\bar{m}_t$ in Eq. (5) is appended to $\bar{f}_t$ in Eq. (4).

Hence, in case of L2 loss and for a constant Gramian matrix, conceptually GD transmits information packets from the residual $\bar{m}_t$ into our model $\bar{f}_t$ along each axis $\bar{v}_i$. Further, $s_i^t \triangleq 1 - |1 - \frac{\delta}{N}\lambda_i|$ governs a speed of information flow along $\bar{v}_i$. Importantly, note that for a high learning rate (i.e. $\delta \approx \frac{2N}{\lambda_{max}}$) the information flow is slow for directions $\bar{v}_i$ with both very large and very small eigenvalues, since in former the term $1 - \frac{\delta}{N}\lambda_i$ is close to -1 whereas in latter - to 1. Yet, along with the learning rate decay, performed during a typical optimization, s_i^t for very large λ_i is increased. However, the speed along a direction with small λ_i is further decreasing with the decay of δ. As well, in case $\lambda_{min} > 0$, at the convergence $t \to \infty$ we will get from Eqs. (4)-(5) the global minima convergence: $\bar{f}_\infty = \bar{f}_0 - \bar{m}_0 = \bar{y}$ and $\bar{m}_\infty = \bar{0}$.

Under the above setting, there are two important key observations. First, due to the restriction over δ in practice the information flow along small λ_i can be prohibitively slow in case a conditional number $\frac{\lambda_{max}}{\lambda_{min}}$ is very large. This implies that for a faster convergence it is desirable for NN to have many eigenvalues as close as possible to its λ_{max} since this will increase a number of directions in the function space where information flow is fast. Second, if $\bar{m}_0$ (or $\bar{y}$ if $\bar{f}_0 \approx 0$) is contained entirely within *top* eigenvectors, small eigenvalues will not affect the convergence rate at all. Hence, the higher alignment between $\bar{m}_0$ (or $\bar{y}$) and *top* eigenvectors may dramatically improve overall convergence rate. The above conclusions and their extensions towards the testing loss are proved in formal manner in [1,15] for two-layer NNs. Further, the generalization was also shown to be dependent on the above alignment.

In Section 6 we evaluate the above conclusions experimentally, showing them to be true. Moreover, we will demonstrate the exceptional alignment between $\bar{y}$ and *top* eigenvectors of G_t along the optimization process. Such behavior can further explain the expressiveness power of NNs.

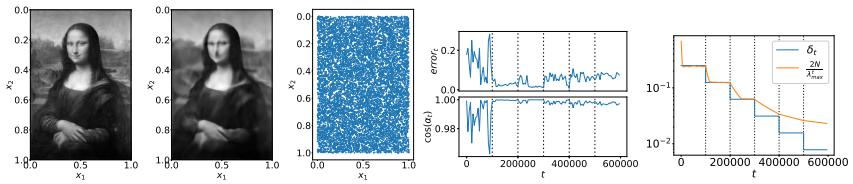
5 Related Work

First-order NN dynamics can be understood by solving the system in Eq. (3). However, its solution is highly challenging due to two main reasons - non-linearity of $\bar{m}_t$ w.r.t. $\bar{f}_t$ (except for the L2 loss) and intricate and yet not fully known time-dependence of Gramian G_t. Although *gradient similarity* $g_t(X, X')$ and corresponding G_t achieved a lot of recent attention in DL community [9,13], their properties are still investigated mostly only for limits under which G_t becomes time-constant. In [9] $g_t(X, X')$ was proven to converge to Neural Tangent Kernel (NTK) in infinite width limit, while in [13] G_0 was shown to accurately explain NN dynamics when θ_t is nearby θ_0. The considered case of constant Gramian facilitates solution of Eq. (3), as demonstrated in Section 4, which otherwise remains intractable.

Yet, in practical-sized NNs the spectrum of G_t is neither constant nor it is similar to its initialization. Recent several studies explored its adaptive dynamics [18,3], with most works focusing on one or two layer NNs. Further, in [4,8] equations for NTK dynamics were developed for a general NN architecture. Likewise, in the Appendix F we derive similar dynamics for the Gramian G_t. Yet, the above derivations produce intricate equations and it is not straightforward to explain the actual behavior of G_t along the optimization, revealed in this paper. In Section 6 we empirically demonstrate that *top* spectrum of G_t drastically changes by aligning itself with the target function. To the best of our knowledge, the presented NN kernel trends were not investigated in such detail before.

Further, many works explore properties of FIM F_t both theoretically and empirically [17,6,10,15]. All works agree that in typical NNs only a small part of FIM eigenvalues are significantly strong, with the rest being negligibly small. According to Section 3 the same is also true about eigenvalues of G_t. Furthermore, in [1,15] authors showed that NN learnability strongly depends on alignment between labels vector $\bar{y}$ and *top* eigenvectors of G_t. Intuitively, it can be explained by fast convergence rate along $\bar{v}_i$ with large λ_i vs impractically slow one along directions with small λ_i, as was shortly described in Section 4. Due to most of the eigenvalues being very small, the alignment between $\bar{y}$ and *top* eigenvectors of G_t defines the optimization performance. Moreover, in [15] authors shortly noted the increased aforementioned alignment comparing ResNet convolutional NN before and after training. In Section 6 we empirically investigate this alignment for FC architecture, in comprehensive manner for various training tasks.

Furthermore, the picture of information flow from Section 4 also explains what target functions are more "easy" to learn. The *top* eigenvectors of G_t typically contain low-frequency signal, which was discussed in [1] and proved in [2] for data uniformly distributed on a hypersphere. In its turn, this explains why low-frequency target functions are learned significantly faster as reported in [19,16,1]. We support findings of [2] also in our experiments below, additionally revealing that for a general case the eigenvectors/eigenfunctions of the *gradient similarity* are not spherical harmonics considered in [2].



(a) **(b)** **(c)** **(d)** **(e)**

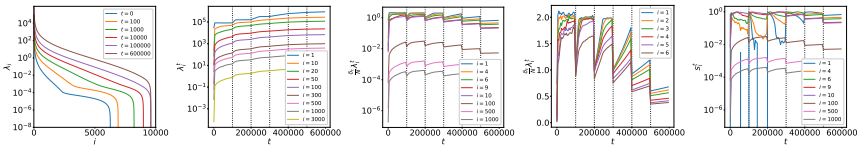
Fig. 1: (a) Mona Lisa target function for a regression task. (b) NN $f_\theta(X)$ at convergence. (c) 10^4 sampled training points. (d) Accuracy of first order dynamics in Eq. (3). Depicted is $error_t = \frac{\left\|d\tilde{f}_t - d\bar{f}_t\right\|}{\left\|d\tilde{f}_t\right\|}$, where $d\bar{f}_t = -\frac{\delta_t}{N} \cdot G_t \cdot \bar{m}_t$ is the first-order approximation of a real differential $d\tilde{f}_t \triangleq \bar{f}_{t+1} - \bar{f}_t$; $\cos(\alpha_t)$ is cosine of an angle between $d\tilde{f}_t$ and $d\bar{f}_t$. As observed, Eq. (3) explains roughly 90% of NN change. (e) Learning rate δ_t and its upper stability boundary $\frac{2N}{\lambda^t_{max}}$ along the optimization. We empirically observe a relation $\lambda^t_{max} \approx \frac{2N}{\delta_t}$.

6 Experiments

In this section we empirically study Gramian dynamics along the optimization process. Our main goal here is to illustrate the alignment nature of the *gradient similarity* kernel and verify various deductions made in Section 4 under a constant-Gramian setting for a real learning case. To do so in detailed and intuitive manner, we focus our experiments on 2D dataset where visualization of kernel eigenfunctions is possible. We perform a simple regression optimization of FC network via GD, where a learning setup is similar to common conventions applied by DL practitioners[4]. **All** empirical conclusions are also validated for high-dimensional real-world data, which we present in Appendix [12].

Setup We consider a regression of the target function $y(X)$ with $X \in [0,1]^2 \subseteq \mathbb{R}^2$ depicted in Figure 1a. This function is approximated via Leaky-Relu FC network and L2 loss, using $N = 10000$ training points sampled uniformly from $[0,1]^2$ (see Figure 1c). Training dataset is normalized to an empirical mean 0 and a standard deviation 1. NN contains 6 layers with 256 neurons each, with $|\theta| = 264193$, that was initialized via Xavier initialization [5]. Such large NN size was chosen to specifically satisfy an over-parametrized regime $|\theta| \gg N$, typically met in DL community. Further, learning rate δ starts at 0.25 and is multiplied by 0.5 each 10^5 iterations, with the total optimization duration being $6 \cdot 10^5$. At convergence $f_\theta(X)$ gets very close to its target, see Figure 1b. Additionally, in Figure 1d we show that first-order dynamics in Eq. (3) describe around 90 percent of the change in NN output along the optimization, leaving another 10 for higher-order Taylor terms. Further, we compute G_t and its spectrum along the optimization, and thoroughly analyze them below.

[4] Related code can be accessed via a repository `https://bit.ly/2kGVHhG`



(a) **(b)** **(c)** **(d)**

Fig. 2: (a) Eigenvalues $\{\lambda_i^t\}_{i=1}^N$ for different t. (b) Individual eigenvalues along t. (c) $\frac{\delta_t}{N}\lambda_i^t$ along time t, for various i. (d) The information flow speed $s_i^t = 1-|1-\frac{\delta}{N}\lambda_i|$ discussed in Section 4, for various i. For first 8 eigenvectors, roughly, this speed is increased at learning rate drop.

Eigenvalues In Figures 2a-2b it is shown that each eigenvalue is monotonically increasing along t. Moreover, at learning rate decay there is an especial boost in its growth. Since $\frac{\delta_t}{N}\lambda_i^t$ also defines a speed of movement in θ-space along one of FIM eigenvectors (see Section 3), such behavior of eigenvalues suggests an existence of mechanism that keeps a roughly constant movement speed of θ within $\mathbb{R}^{|\theta|}$. To do that, when δ_t is reduced, this mechanism is responsible for increase of $\{\lambda_i^t\}_{i=1}^N$ as a compensation. This is also supported by Figure 2c where each $\frac{\delta_t}{N}\lambda_i^t$ is balancing, roughly, around the same value along the entire optimization. Furthermore, in Figure 1e it is clearly observed that an evolution of λ_{max}^t stabilizes[5] only when it reaches value of $\frac{2N}{\delta_t}$, further supporting the above hypothesis.

Neural Spectrum Alignment Notate by $\cos\left[\alpha_t\left(\bar{\phi},k\right)\right] \triangleq \sqrt{\frac{\sum_{i=1}^k <\bar{v}_i^t,\bar{y}>^2}{\left\|\bar{\phi}\right\|_2^2}}$ the cosine of an angle $\alpha_t\left(\bar{\phi},k\right)$ between an arbitrary vector $\bar{\phi}$ and its projection to the sub-space of $\mathbb{R}^N$ spanned by $\{\bar{v}_i^t\}_{i=1}^k$. Further, $E_t(\bar{\phi},k) \triangleq \cos^2\left[\alpha_t\left(\bar{\phi},k\right)\right]$ can be considered as a *relative energy* of $\bar{\phi}$, the percentage of its energy $\left\|\bar{\phi}\right\|_2^2$ located inside $span\left(\{\bar{v}_i^t\}_{i=1}^k\right)$. In our experiments we will use $E_t(\bar{\phi},k)$ as an alignment metric between $\bar{\phi}$ and $\{\bar{v}_i^t\}_{i=1}^k$. Further, we evaluate alignment of G_t with $\bar{y}$ instead of $\bar{m}_0$ since $\bar{f}_0$ is approximately zero in the considered FC networks.

In Figure 3a we depict relative energy of the label vector $\bar{y}$ in *top* k eigenvectors of G_t, $E_t(\bar{y},k)$. As observed, 20 *top* eigenvectors of G_t contain 90 percent of $\bar{y}$ for almost all t. Similarly, 200 *top* eigenvectors of G_t contain roughly 98 percent of $\bar{y}$, with rest of eigenvectors being practically orthogonal w.r.t. $\bar{y}$. That is, G_t aligns its *top* spectrum towards the ground truth target function $\bar{y}$ almost immediately after training starts, which improves the convergence rate since the information flow is fast along *top* eigenvectors, as discussed in Section 4 and proved in [1,15].

[5] Trend $\lambda_{max}^t \to \frac{2N}{\delta_t}$ was consistent in FC NNs for a wide range of initial learning rates, number of layers and neurons, and various datasets (see Appendix [12]), making it an interesting venue for a future theoretical investigation

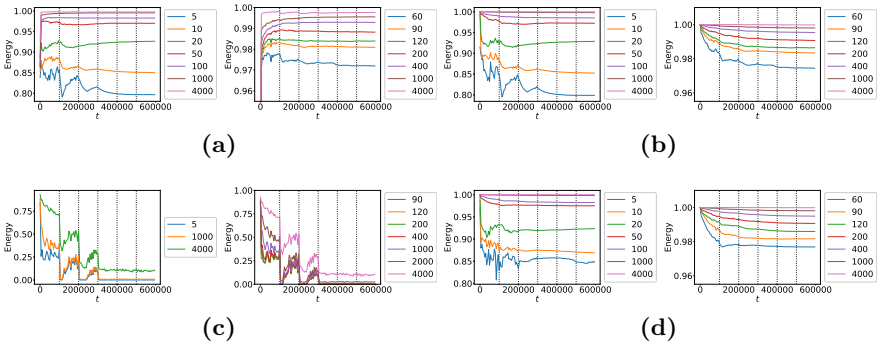
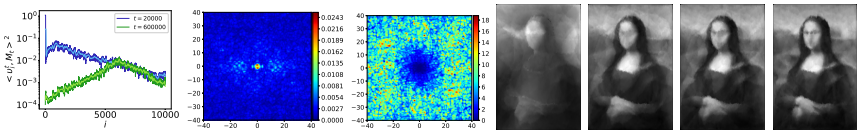


Fig. 3: (a) For different k, relative energy of the label vector $\bar{y}$ in *top* k eigenvectors of G_t, $E_t(\bar{y}, k)$, along the optimization time t. (b) Relative energy of NN output, $E_t(\bar{f}_t, k)$. (c) Relative energy of the residual, $E_t(\bar{m}_t, k)$. (d) Relative energy of NN output, $E_t(\bar{f}_t^{test}, k)$, with both G_t and $\bar{f}_t^{test}$ computed at 10^4 testing points. Dashed vertical lines depict time t at which learning rate δ was decayed (see Figure 1e).

Further, we can see that for $k < 400$ the relative energy $E_t(\bar{y}, k)$ is decreasing after each decay of δ, yet for $k > 400$ it keeps growing along the entire optimization. Hence, the *top* eigenvectors of G_t can be seen as NN memory that is learned/tuned toward representing the target $\bar{y}$, while after each learning rate drop the learned information is spread more evenly among a higher number of different *top* eigenvectors.

Likewise, in Figure 3b we can see that NN outputs vector $\bar{f}_t$ is located entirely in a few hundreds of *top* eigenvectors. In case we consider G_t to be constant, such behavior can be explained by Eq. (3) since each increment of $\bar{f}_t$, $d\bar{f}_t$, is also located within *top* eigenvectors of G_t. Yet, for a general NN with a time-dependent kernel the theoretical justification for the above empirical observation is currently missing. Further, similar relation is observed also at points outside of $\boldsymbol{\mathcal{X}}$ (see Figure 3d), leading to the empirical conclusion that *top* eigenfunctions of *gradient similarity* $g_t(X, X')$ are the basis functions of NN $f_\theta(X)$.

Residual Dynamics Further, a projection of the residual $\bar{m}_t$ onto *top* eigenvectors, shown in Figure 3c, is decreasing along t, supporting Eq. (5). Particularly, we can see that at $t = 600000$ only 10% of $\bar{m}_t$'s energy is located inside *top* 4000 eigenvectors, and thus at the optimization end 90% of its energy is inside *bottom* eigenvectors. Moreover, in Figure 4a we can observe that the projection of $\bar{m}_t$ along *bottom* 5000 eigenvectors almost does not change during the entire optimization. Thus, we empirically observe that the information located in the *bottom* spectrum of G_t was not learned, even for a relatively long optimization process (i.e. 600000 iterations), which can be explained by slow convergence associated with *bottom* eigenvectors. Furthermore, since this spectrum part is



(a) **(b)** **(c)** **(d)**

Fig. 4: (a) Spectral projections of the residual $\bar{m}_t$, $< \bar{v}_i^t, \bar{m}_t >^2$, at $t = 20000$ and $t = 600000$; (b) and (c) Fourier Transform of $\bar{m}_t$ at $t = 20000$ and $t = 600000$ respectively. The high frequency is observed to be dominant in (c). (d) a linear combination $\bar{f}_{t,k} \triangleq \sum_{i=1}^{k} < \bar{v}_i^t, \bar{f}_t > \bar{v}_i^t$ of first $k = \{10, 100, 200, 500\}$ eigenvectors at $t = 600000$. Each vector $\bar{f}_{t,k}$ was interpolated from training points $\{X^i\}_{i=1}^N$ to entire $[0, 1]^2$ via a linear interpolation.

also associated with high-frequency information [2], $\bar{m}_t$ at $t = 600000$ comprises mostly the noise, which is also evident from Figures 4b-4c.

Moreover, we can also observe in Figure 3c a special drop of $E_t(\bar{m}_t, k)$ at times of δ decrease. This can be explained by the fact that a lot of $\bar{m}_t$'s energy is trapped inside first several $\{\bar{v}_i^t\}$ (see $E_t(\bar{m}_t, 5)$ in Figure 3c). When learning rate is decreased, the information flow speed $s_i^t \triangleq 1 - |1 - \frac{\delta_t}{N}\lambda_i^t|$, discussed in Section 4, is actually increasing for a few *top* eigenvectors (see Figure 2d). That is, terms $\frac{\delta_t}{N}\lambda_i^t$, being very close to 2 before δ's decay, are getting close to 1 after, as seen in Figure 2c. In its turn this accelerates the information flow along these first $\{\bar{v}_i^t\}$, as described in Eq. (4)-(5). Further, this leads also to a special descend of $E_t(\bar{m}_t, k)$ and of the training loss (see Figure 7b below).

Eigenvectors We further explore $\{\bar{v}_i^t\}$ in a more illustrative manner, to produce a better intuition about their nature. In Figure 4d a linear combination of several *top* eigenvectors at $t = 600000$ is presented, showing that with only 100 vectors we can accurately approximate the NN output in Figure 1b.

Furthermore, in Figure 5 several eigenvectors are interpolated to entire $[0, 1]^2$. We can see that *top* $\{\bar{v}_i^t\}$ obtained visual similarity with various parts of Mona Lisa image and indeed can be seen as basis functions of $f_\theta(X)$ depicted in Figure 1b. Likewise, we also demonstrate the Fourier Transform of each $\bar{v}_i^t$. As observed, the frequency of the contained information is higher for smaller eigenvalues, supporting conclusions of [2]. More eigenvectors are depicted in Appendices I-N.

Likewise, in Figure 6 same eigenvectors are displayed at $t = 20000$. At this time the visual similarity between each one of first eigenvectors and the target function in Figure 1a is much stronger. This can be explained by the fact that the information about the target function within G_t is spread from first few towards higher number of *top* eigenvectors after each learning rate drop, as was described above. Hence, before the first drop at $t = 100000$ this information is mostly gathered within first few $\{\bar{v}_i^t\}$ (see also $E_t(\bar{y}, 10)$ in Figure 3a).

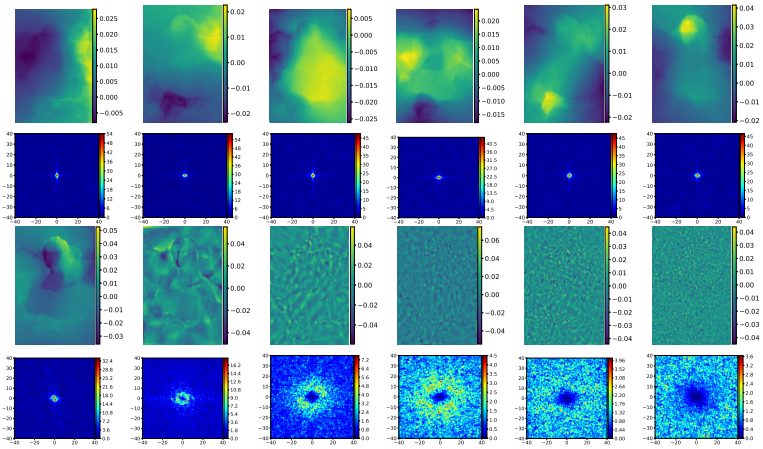


Fig. 5: Eigenvectors of Gramian G_t at $t = 600000$, and their Fourier Transforms (see the Appendix G for technical details). First two rows: from left-to-right, 6 first eigenvectors. Last two rows: 10-th, 100-th, 500-th, 1000-th, 2000-th and 4000-th eigenvectors . As observed, a frequency of signal inside of each eigenvector increases when moving from large to small eigenvalue.

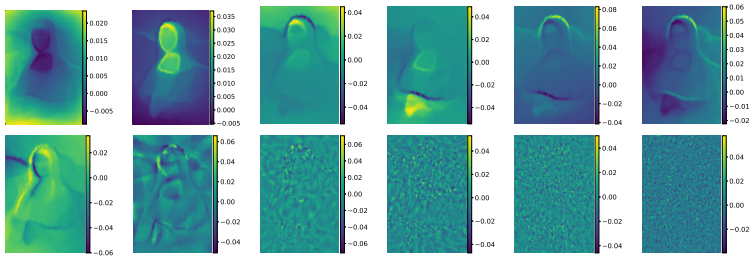
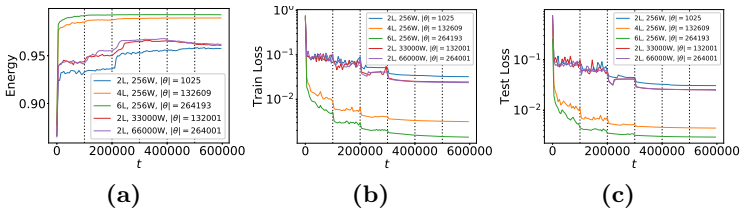


Fig. 6: First line: from left-to-right, 6 first eigenvectors of Gramian G_t at $t = 20000$. Second line: 10-th, 100-th, 500-th, 1000-th, 2000-th and 4000-th eigenvectors.



(a) **(b)** **(c)**

Fig. 7: (a) For NNs with a different number of layers **L** and number of neurons **W**, relative energy of labels $\bar{y}$ in *top* 400 eigenvectors of G_t, $E_t(\bar{y}, 400)$, along the optimization time t; (b) training loss and (c) testing loss of these models.

Alignment and NN Depth / Width Here we further study how the width and the depth of NN affect the alignment between G_t and the ground truth signal $\bar{y}$. To this purpose, we performed the optimization under the identical setup, yet with NNs containing various numbers of layers and neurons. In Figure 7a we can see that in deeper NN *top* eigenvectors of G_t aligned more towards $\bar{y}$ - the relative energy $E_t(\bar{y}, 400)$ is higher for a larger depth. This implies that more layers, and the higher level of non-linearity produced by them, yield a better alignment between G_t and $\bar{y}$. In its turn this allows NN to better approximate a given target function, as shown in Figures 7b-7c, making it more expressive for a given task. Moreover, in evaluated 2-layer NNs, with an increase of neurons and parameters the alignment rises only marginally.

Scope of Analysis The above empirical analysis was repeated under numerous different settings and can be found in Appendix [12]. We evaluated various FC architectures, with and without shortcuts between the layers and including various activation functions. Likewise, optimizers GD, stochastic GD and Adam were tested on problems of regression (L2 loss) and density estimation (noise contrastive estimation [7]). Additionally, various high-dimensional real-world datasets were tested, including MNIST and CIFAR100. **All** experiments exhibit the same alignment nature of kernel towards the learned target function. The results are also consistent with our previous experiments in [11].

7 Discussion and Conclusions

In this paper we empirically revealed that during GD *top* eigenfunctions of *gradient similarity* kernel change to align with the target function $y(X)$ learned by NN $f_\theta(X)$, and hence can be considered as a NN memory tuned during the optimization to better represent $y(X)$. This alignment is significantly higher for deeper NNs, whereas a NN width has only a minor effect on it. Moreover, the same *top* eigenfunctions represent a *neural spectrum* - the $f_\theta(X)$ is a linear combination of these eigenfunctions during the optimization. As well, we showed various trends of the kernel dynamics affected by learning rate decay. Same alignment behavior was observed for various *supervised* and *unsupervised* losses and high-dimensional datasets, optimized via several different optimizers. Likewise, several variants of FC architecture were evaluated. Since the above alignment is critical for a learning [1,2,15], the main question remains to how NN architecture and optimization hyper-parameters affect this spectrum, and what is their optimal configuration for learning a given function $y(X)$. We shall leave it for a future exciting research.

Acknowledgments The authors thank Daniel Soudry and Dar Gilboa for discussions on dynamics of a Neural Tangent Kernel (NTK). This work was supported in part by the Israel Ministry of Science & Technology (MOST) and Intel Corporation. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU, which, among other GPUs, was used for this research.

References

1. Arora, S., Du, S.S., Hu, W., Li, Z., Wang, R.: Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. arXiv preprint arXiv:1901.08584 (2019)
2. Basri, R., Jacobs, D., Kasten, Y., Kritchman, S.: The convergence rate of neural networks for learned functions of different frequencies. arXiv preprint arXiv:1906.00425 (2019)
3. Dou, X., Liang, T.: Training neural networks as learning data-adaptive kernels: Provable representation and approximation benefits. arXiv preprint arXiv:1901.07114 (2019)
4. Dyer, E., Gur-Ari, G.: Asymptotics of wide networks from feynman diagrams. arXiv preprint arXiv:1909.11304 (2019)
5. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 249–256 (2010)
6. Gur-Ari, G., Roberts, D.A., Dyer, E.: Gradient descent happens in a tiny subspace. arXiv preprint arXiv:1812.04754 (2018)
7. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. pp. 297–304 (2010)
8. Huang, J., Yau, H.T.: Dynamics of deep neural networks and neural tangent hierarchy. arXiv preprint arXiv:1909.08156 (2019)
9. Jacot, A., Gabriel, F., Hongler, C.: Neural tangent kernel: Convergence and generalization in neural networks. In: Advances in Neural Information Processing Systems (NIPS). pp. 8571–8580 (2018)
10. Karakida, R., Akaho, S., Amari, S.i.: Universal statistics of fisher information in deep neural networks: mean field approach. arXiv preprint arXiv:1806.01316 (2018)
11. Kopitkov, D., Indelman, V.: General probabilistic surface optimization and log density estimation. arXiv preprint arXiv:1903.10567 (2019)
12. Kopitkov, D., Indelman, V.: Neural spectrum alignment: Empirical study - appendix. `https://bit.ly/3aipgtl` (2019)
13. Lee, J., Xiao, L., Schoenholz, S.S., Bahri, Y., Sohl-Dickstein, J., Pennington, J.: Wide neural networks of any depth evolve as linear models under gradient descent. arXiv preprint arXiv:1902.06720 (2019)
14. Ollivier, Y.: Riemannian metrics for neural networks i: feedforward networks. Information and Inference: A Journal of the IMA **4**(2), 108–153 (2015)
15. Oymak, S., Fabian, Z., Li, M., Soltanolkotabi, M.: Generalization guarantees for neural networks via harnessing the low-rank structure of the jacobian. arXiv preprint arXiv:1906.05392 (2019)
16. Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F.A., Bengio, Y., Courville, A.: On the spectral bias of neural networks. arXiv preprint arXiv:1806.08734 (2018)
17. Sagun, L., Evci, U., Guney, V.U., Dauphin, Y., Bottou, L.: Empirical analysis of the hessian of over-parametrized neural networks. arXiv preprint arXiv:1706.04454 (2017)
18. Woodworth, B., Gunasekar, S., Lee, J., Soudry, D., Srebro, N.: Kernel and deep regimes in overparametrized models. arXiv preprint arXiv:1906.05827 (2019)
19. Zhang, J., Springenberg, J.T., Boedecker, J., Burgard, W.: Deep reinforcement learning with successor features for navigation across similar environments. arXiv preprint arXiv:1612.05533 (2016)