Neural Spectrum and Gradient Similarity

Dmitry Kopitkov and Vadim Indelman
Technion – Israel Institute of Technology

October 2019

Introduction

- Expressiveness and generalization of deep models during a GD optimization was recently addressed via Neural Tangent Kernel (NTK) [1]
- In most works this kernel is considered to be time-invariant [1,2], defined entirely by NN architecture and independent of the learning task.
- In contrast, we show empirically that **top** eigenfunctions of NTK align toward the target function learned by NN, and also serve as basis functions for NN output - a function represented by NN is spanned almost completely by them for the entire optimization process. Further, since the learning along **top** eigenfunctions is typically fast, their alignment with the target function improves the overall optimization performance.

Notations

- Consider a NN $f_\theta(X): \mathbb{R}^d \to \mathbb{R}$, training dataset $D = \{\mathbf{X} = \{X^i \in \mathbb{R}^d\}_{i=1}^N, \mathbf{Y} = \{Y^i \in \mathbb{R}\}_{i=1}^N\}$ and loss:

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

- Define gradient-similarity kernel (NTK) $g_t(X, X') \equiv \nabla_\theta f_{\theta_t}(X)^T \cdot \nabla_\theta f_{\theta_t}(X')$ and its $N \times N$ Gramian $G_t \equiv g_t(\mathbf{X}, \mathbf{X})$, labels vector $\bar{y}$, NN outputs vector $\bar{f}_t$ with entries $\bar{f}_t(i) = f_{\theta_t}(X^i)$ and a functional derivative vector $\bar{m}_t$ with entries $\bar{m}_t(i) = \ell'\left[X^i, Y^i, f_{\theta_t}(X^i)\right]$
- Denote eigenvalues and eigenvectors of G_t by $\{\lambda_i^t\}_{i=1}^N$ and $\{\bar{\upsilon}_i^t\}_{i=1}^N$, with $\lambda_{max}^t \equiv \lambda_1^t$ and $\lambda_{min}^t \equiv \lambda_N^t$.
- GD update: $d\theta_t \equiv \theta_{t+1} - \theta_t = -\delta \cdot \nabla_\theta L(\theta_t, D)$
- First-order Dynamics:

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

$$\boxed{d\bar{f}_t \equiv \bar{f}_{t+1} - \bar{f}_t \approx -\frac{\delta}{N} \cdot G_t \cdot \bar{m}_t}$$

L2 Loss Dynamics and a Constant Gramian

- Functional derivative is the residual: $\bar{m}_t = \bar{f}_t - \bar{y}$
- First-order Dynamics when G_t is constant:

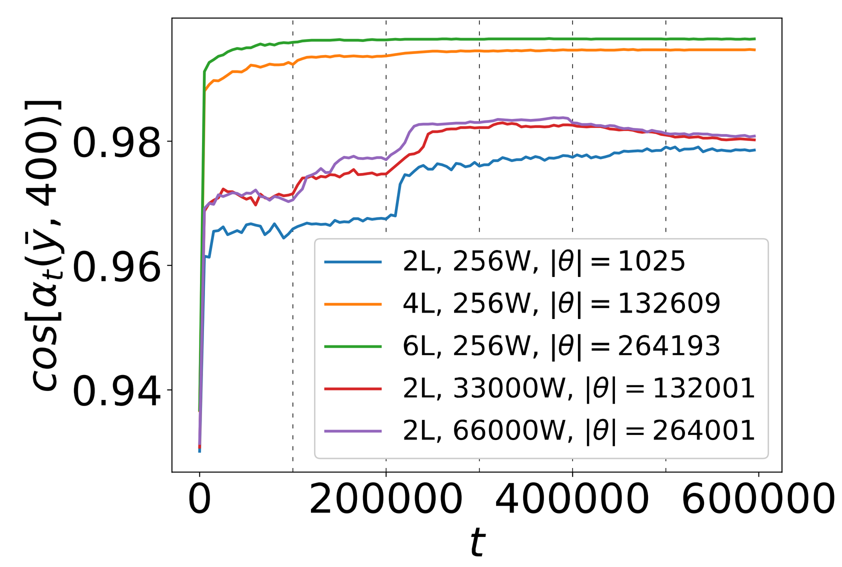
$$\bar{f}_t = \bar{f}_0 - \sum_{i=1}^N \left[1 - \left[1 - \frac{\delta}{N}\lambda_i\right]^t\right] < \bar{\upsilon}_i, \bar{m}_0 > \bar{\upsilon}_i$$

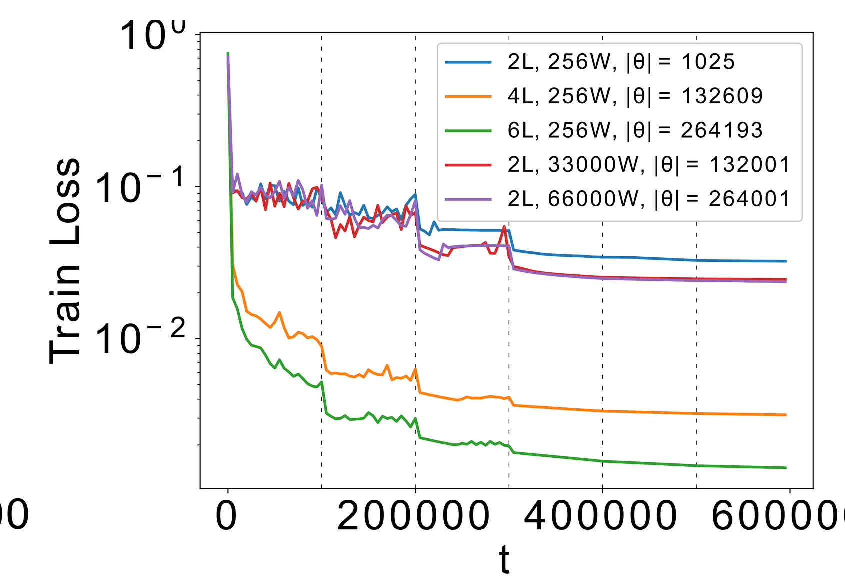
$$\bar{m}_t = \sum_{i=1}^N \left[1 - \frac{\delta}{N}\lambda_i\right]^t < \bar{\upsilon}_i, \bar{m}_0 > \bar{\upsilon}_i$$

- Insights under this setting:
  - $\bar{m}_t$ is reduced and $\bar{f}_t$ is increased along each $\bar{\upsilon}_i$ by the same amount
  - Conceptually, information flows from $\bar{m}_t$ to $\bar{f}_t$ during optimization
  - For $\delta < \frac{2N}{\lambda_{max}}$ and $\lambda_{min} > 0$, global convergence $\bar{f}_\infty = \bar{y}$ at $t \to \infty$
  - $s_i \equiv 1 - |1 - \frac{\delta}{N}\lambda_i|$ governs flow speed along every $\bar{\upsilon}_i$
  - Decay of $\{\lambda_i\}_{i=1}^N$ is typically fast
  - In general, for large λ_i the flow speed is high
  - For small λ_i the flow is slow, sometimes even <u>neglectable</u>
  - For faster convergence <u>we want</u> many eigenvalues close to λ_{max}
  - Alternatively, <u>we want</u> **top** eigenvectors $\{\bar{\upsilon}_i\}_i$ to span $\bar{m}_0 = \bar{f}_0 - \bar{y}$

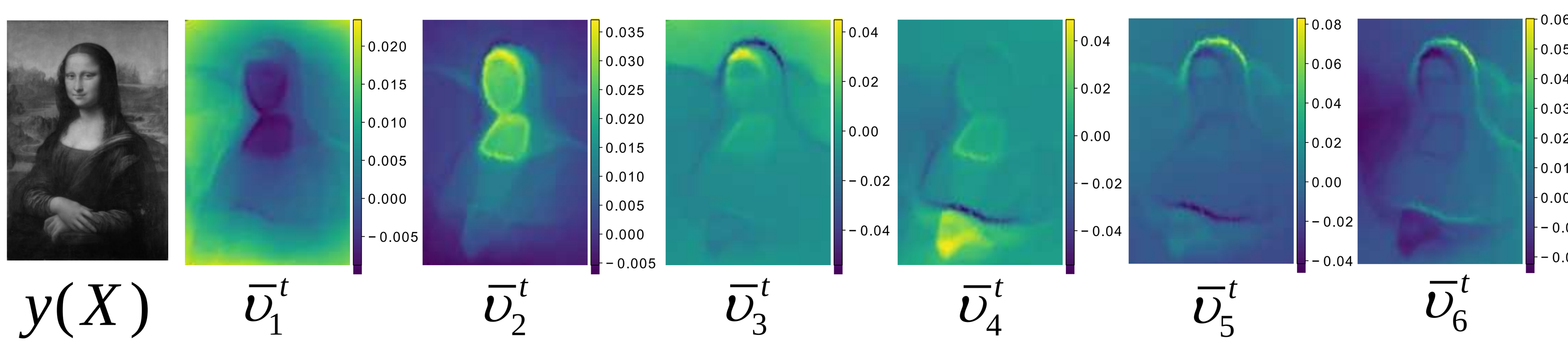
Results

- For an arbitrary vector $\bar{\phi}$ define $\cos\left[\alpha_t(\bar{\phi}, k)\right] \equiv \sqrt{\frac{\sum_{i=1}^k <\bar{\phi}, \bar{\upsilon}_i^t>^2}{\|\bar{\phi}\|_2^2}}$, where $\alpha_t(\bar{\phi}, k)$ is an angle between $\bar{\phi}$ and its projection onto $\text{span}\left(\{\bar{\upsilon}_i^t\}_{i=1}^k\right)$
- <u>Setup</u>: L2 regression, $N = 10000$, X^i sampled uniformly in $[0,1]^2$, $Y^i = y(X^i)$
- Depth increases alignment, alignment improves performance:

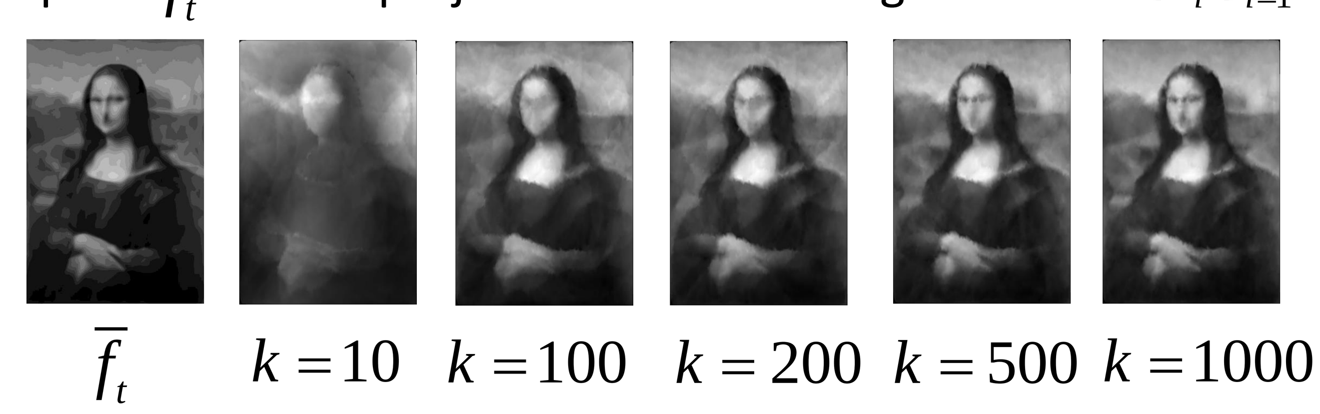




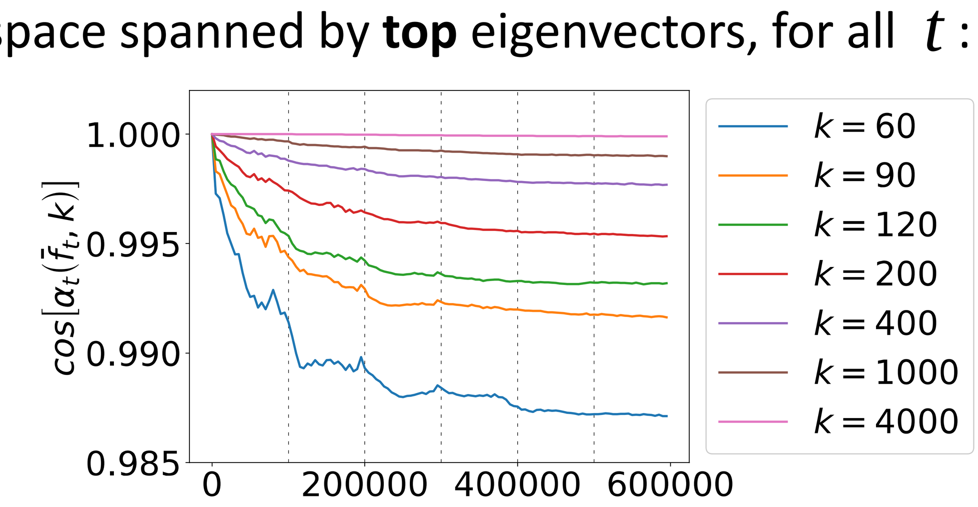
- First **top** eigenvectors for NN with 6 layers at $t = 20000$:



- NN outputs $\bar{f}_t$ and its projection to first k eigenvectors $\{\bar{\upsilon}_i^t\}_{i=1}^k$:



- $\bar{f}_t$ is in a subspace spanned by **top** eigenvectors, for all t:



Conclusions

- Higher <u>alignment</u> between **top** eigenvectors and the target function improves optimization performance
- In actual NNs, **top** spectrum of G_t, and hence also of $g_t(X, X')$, <u>aligns</u> towards target function $\bar{y}$
- Deeper NNs have higher <u>alignment</u>, which also explains their performance superiority
- **Top** eigenvectors/eigenfunctions are basis functions of NN, spanning it almost completely
- Beyond GD and L2 loss, similar behavior was also observed for SGD, Adam and *unsupervised learning* losses in [3]
- More trends of G_t dynamics can be found in:

<u>https://arxiv.org/abs/1910.08720</u>

References

[1] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 8571--8580, 2018.
[2] Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *arXiv preprint arXiv:1902.06720*, 2019.
[3] Dmitry Kopitkov and Vadim Indelman. General Probabilistic Surface Optimization and Log Density Estimation. *arXiv preprint arXiv:1903.10567*, 2019.