Towards Understanding Hierarchical Learning: Benefits of Neural Representations

Minshuo Chen*, Yu Bai◇, Jason D. Lee†, Tuo Zhao*, Huan Wang◇, Caiming Xiong◇, Richard Socher◇

*Georgia Tech, ◇Salesforce Research, †Princeton University

Deep v.s. Shallow

▷ Deep NNs **outperform** shallow ones (He et al. 2016):

+ 8% - 10% top-1 accuracy

ResNet-152

ResNet-18

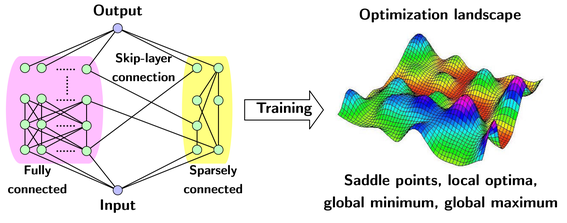
Common belief: deep neural networks are able to perform efficient **hierarchical learning**.

Existing Literature

▷ Classical Approximation Theory

- *Depth separation theory*: deep networks can well approximate certain structured functions, yet shallow networks cannot (Telgarsky 2016; Eldan and Shamir 2016).
- *Universal approximation theory*: deep networks are more efficient in approximating dense function classes, e.g., Sobolev and Hölder functions (Yarotsky 2017).

No **optimization** and **generalization** guarantees.



▷ Neural Tangent Kernels (Jacot et al. 2018):

- *Lazy training*: overparameterized neural networks can be efficiently optimized and generalize as well as shallow learners, e.g., kernels (Allen-Zhu et al. 2018).

Cannot benefit from the **hierarchical structure**.

Three-layer Network with Representation

▷ Prototypical trainable Model:

$$f_{\mathbf{W}}(\mathbf{x}) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \phi(\mathbf{w}_r^\top \mathbf{h}(\mathbf{x})).$$

- $\mathbf{x} \in \mathbb{R}^d$ is the raw input data;
- $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_m]^\top \in \mathbb{R}^{m \times D}$ is the weight matrix;
- ϕ is the activation function;
- $\mathbf{h} : \mathbb{R}^d \mapsto \mathbb{R}^D$ is a **fixed** representation function.

▷ One-layer Neural Representation Function:

$$\mathbf{h}(\mathbf{x}) = \sigma(\mathbf{V}\mathbf{x} + \mathbf{b}).$$

- $\mathbf{V} \in \mathbb{R}^{D \times d}$ and $\mathbf{b} \in \mathbb{R}^D$ are *fixed* weight parameters;
- σ is some nonlinear function.

▷ Similar to a **three-layer neural network**:

$$\widetilde{f}_{\mathbf{W},\mathbf{V},\mathbf{b}}(\mathbf{x}) = \frac{1}{\sqrt{m}} \mathbf{a}^\top \phi(\mathbf{W}\sigma(\mathbf{V}\mathbf{x} + \mathbf{b})),$$

We let $\mathbf{W}$ be trainable and fix $(\mathbf{V}, \mathbf{b})$.

Linearized / Taylorized Models

▷ We consider the linearized / taylorized versions:

$$f^L_{\mathbf{W}}(\mathbf{x}) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \phi'(\mathbf{w}_{0,r}^\top \mathbf{h}(\mathbf{x}))(\mathbf{w}_r^\top \mathbf{h}(\mathbf{x})), \qquad \texttt{(NTK-h)}$$

$$f^Q_{\mathbf{W}}(\mathbf{x}) = \frac{1}{2\sqrt{m}} \sum_{r=1}^{m} a_r \phi''(\mathbf{w}_{0,r}^\top \mathbf{h}(\mathbf{x}))(\mathbf{w}_r^\top \mathbf{h}(\mathbf{x}))^2. \qquad \texttt{(Quad-h)}$$

- $\mathbf{W}_0$ is the initialization;
- $\mathbf{W}$ is overloaded to denote the **increment** of weights.

Note: (NTK-h) and (Quad-h) are the linearized / quadratic expansions of the 3-layer neural network w.r.t. the $\mathbf{W}$ layer.

▷ Supervised Learning — Minimizing Regularized Risk:

$$\min_{\mathbf{W}} \widehat{\mathcal{R}}_\lambda \left(f^Q_{\mathbf{W}}\right) := \frac{1}{n} \sum_{i=1}^{n} \ell \left(f^Q_{\mathbf{W}}(\mathbf{x}_i), y_i\right) + \lambda \|\mathbf{W}\|_{2,4}^4 .$$

- *Data*: $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ are data points with $\mathbf{x}_i \in \mathbb{S}^{d-1}$ and $y \in \mathcal{Y}$;
- *Loss*: $\ell(\cdot, y)$ is a convex loss function, twice differentiable with bounded derivatives, and $|\ell(\mathbf{0}, y)| \le 1$ for any $y \in \mathcal{Y}$, e.g., logistic loss and soft hinge loss;
- *Regularization*: $\|\mathbf{W}\|_{2,4}^4 := \sum_{r=1}^m \|\mathbf{w}_r\|_2^4$.
- *Initialization*: $\mathbf{w}_{0,r} \overset{\text{i.i.d}}{\sim} \mathbb{N}(\mathbf{0}, \mathbf{I}_D)$ and $a_r \overset{\text{i.i.d.}}{\sim} \text{Unif}(\{\pm 1\})$.

Optimization and Generalization

Theorem 1. *Assume*

- *Bounded representation:* $\|\mathbf{h}(\mathbf{x})\|_2 \le B_h$ *almost surely;*
- *Differentiable activation:* $\sup_{t \in \mathbb{R}} |\phi''(t)| \le$ *const.*

(1) (Optimization) Given any $\epsilon > 0$, $\tau = \Theta(1)$, and some radius $B_{w,\star} > 0$, suppose the width $m \ge \widetilde{O}(B_h^4 B_{w,\star}^4 \epsilon^{-1})$ and coefficient $\lambda > 0$ chosen properly.

Then any second-order stationary point (SOSP) $\widehat{\mathbf{W}}$ of the regularized risk $\widehat{\mathcal{R}}_\lambda(f^Q_{\mathbf{W}})$ satisfies $\|\widehat{\mathbf{W}}\|_{2,4} \le O(B_{w,\star})$, and achieves

$$\widehat{\mathcal{R}}_\lambda(f^Q_{\widehat{\mathbf{W}}}) \le (1+\tau) \min_{\|\mathbf{W}\|_{2,4} \le B_{w,\star}} \widehat{\mathcal{R}}(f^Q_{\mathbf{W}}) + \epsilon.$$

(2) (Generalization) For any radius $B_w > 0$, we have with high probability (over $(\mathbf{a}, \mathbf{W}_0)$) that

$$\mathbb{E}_{(\mathbf{x}_i, y_i)} \left[\sup_{\|\mathbf{W}\|_{2,4} \le B_w} \left| \mathcal{R}(f^Q_{\mathbf{W}}) - \widehat{\mathcal{R}}(f^Q_{\mathbf{W}}) \right| \right] \le \widetilde{O}\left(\frac{B_h^2 B_w^2 M_{h,op}}{\sqrt{n}} + \frac{1}{\sqrt{n}} \right),$$

where $M_{h,op}^2 = B_h^{-2} \mathbb{E}_{\mathbf{x}} \left[\left\| \frac{1}{n} \sum_{i=1}^n \mathbf{h}(\mathbf{x}_i)\mathbf{h}(\mathbf{x}_i)^\top \right\|_2 \right]$ and $\mathcal{R}$, $\widehat{\mathcal{R}}$ are unregularized population and empirical risks, respectively.

Implications of Theorem 1:

- **Efficient optimization**: escaping-saddle type algorithms, e.g., noisy SGD, can efficiently optimize $\widehat{\mathcal{R}}(f^Q_{\mathbf{W}})$;
- **Feature isotropicity**: isotropic $\mathbf{h}(\mathbf{x}) \Rightarrow$ small $M_{h,\text{op}} \Rightarrow$ good generalization.

Whitened Representations

▷ Whitened Representation:

$$\mathbf{h}(\mathbf{x}) = \widehat{\mathbf{\Sigma}}^{-1/2} \mathbf{g}(\mathbf{x}), \quad \text{with} \quad \mathbf{g}(\mathbf{x}) = \sigma(\mathbf{V}\mathbf{x} + \mathbf{b}).$$

- Sample convariance matrix $\widehat{\mathbf{\Sigma}} = \frac{1}{n_0} \sum_{i=1}^{n_0} \mathbf{g}(\widetilde{\mathbf{x}}_i)\mathbf{g}(\widetilde{\mathbf{x}}_i)^\top$;
- $\{\widetilde{\mathbf{x}}_i\}_{i=1}^{n_0}$ are additional unlabeled samples;
- $\mathbf{v}_r \overset{\text{i.i.d.}}{\sim} \mathbb{N}(\mathbf{0}, \mathbf{I}_d)$ and $\mathbf{b} \sim \mathbb{N}(\mathbf{0}, \mathbf{I}_D)$; fixed during training.

Assumption 1 (Lower bounded covariance)**.** *For any k and $D \le O(d^k)$, with high probability over $\mathbf{V}, \mathbf{b}$ (as $d \to \infty$), we have the minimum eigenvalue*

$$\lambda_{\min}(\mathbf{\Sigma}) \ge \lambda_k$$

for some constant $\lambda_k > 0$ that only depends on k but not d, where $\mathbf{\Sigma} = \mathbb{E}_{\mathbf{x}}[\sigma(\mathbf{V}\mathbf{x} + \mathbf{b})\, \sigma(\mathbf{V}\mathbf{x} + \mathbf{b})^\top]$.

Main Result: Learning Low-Rank Polynomials

▷ Low-Rank Polynomials:

$$f_\star(\mathbf{x}) = \sum_{s=1}^{r_\star} \alpha_s (\boldsymbol{\beta}_s^\top \mathbf{x})^{p_s}, \qquad \texttt{(Low-rank poly)}$$

where $|\alpha_s| \le 1$, $\left\|(\boldsymbol{\beta}_s^\top \mathbf{x})^{p_s}\right\|_{L_2} \le 1$, $p_s \le p$ for all s.

▷ Sample Complexity Bounds:

- Assumption 1 holds;
- $\exists\ f_\star$ of the form (Low-rank poly) that achieves low risk: $\mathbb{E}[\ell(f_\star(\mathbf{x}), y)] \le \text{OPT}$;
- $\sigma(t) = \mathbb{1}\{t \ge 0\}$ and $\phi(t) = \text{ReLU}^2(t)/2$;
- $D = \Theta\left(\text{poly}(r_\star, p) \sum_s \|\boldsymbol{\beta}_s\|_2^{2\lceil p_s/2 \rceil} \epsilon^{-2} \delta^{-1}\right)$, $m \ge \widetilde{O}\left(\text{poly}(r_\star, D)\epsilon^{-2}\delta^{-1}\right)$, $\quad n_0 = \widetilde{O}(D\delta^{-2})$.

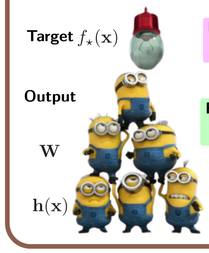
We compare learning with representations and raw inputs:

Quad-h	Quad-Raw
$\widetilde{O}\left(d^{\lceil p/2 \rceil}\right)$	$\widetilde{O}\left(d^{p-1}\right)$
v.s.	
NTK-h	NTK-Raw
$\Omega(d^p)$	$O(d^p)$

Quad-h significantly **outperforms** Quad-Raw (when $p \ge 4$), while NTK-h does not improve on NTK-Raw.

▷ Benefit of Representations — Hierarchical Learning:

Quad-h model can express polynomials hierarchically, using weight matrices with much smaller norms than that of a shallow learner.



Improved sample complexity: hierarchically expressing a degree-p polynomial, NO "heavylifting" on a single layer

Reduced complexity: low complexity functions need to be learned by the top layer based upon the expressivity of $\mathbf{h}$

Rich base: linear combinations of $\mathbf{h}$ yields expressive nonlinear functions, e.g., degree-k polynomials ($k < p$)

Discussions

- *Unwhitened representation*: using a data dependent regularizer can also achieve the improved sample complexity.

$$\min_{\mathbf{W}} \widehat{\mathcal{R}}^{\text{dreg}}_\lambda(f^Q_{\mathbf{W}}) = \frac{1}{n} \sum_{i=1}^{n} \ell(f^Q_{\mathbf{W}}(\mathbf{x}_i), y_i) + \lambda \left\| \mathbf{W}\widehat{\mathbf{\Sigma}}^{1/2} \right\|_{2,4}^4 .$$

- *Implications of Assumption 1*: $\mathbf{h}(\mathbf{x}_i)$'s are not too correlated, which roughly requires that the distribution of $\mathbf{x}$ spans all directions in $\mathbb{R}^d$.