Recovery Guarantees for One-hidden-layer Neural Networks

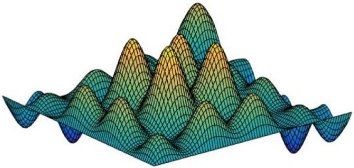
Kai Zhong*

Joint work with
Zhao Song*, Prateek Jain†, Peter L. Bartlett‡, Inderjit S. Dhillon*

*UT-Austin, †MSR India, ‡UC Berkeley

Learning Neural Networks is Hard

- The objective functions of neural networks are highly non-convex.
- Gradient-descent-based methods only achieve local optima.

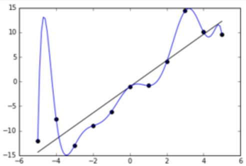


Learning Neural Networks is Hard

- Good News
  - When the size of the network is very large, no need to worry about bad local minima.
  - Every local minimum is a global minimum or close to a global minimum. [Choromanska et al. '15, Nguyen & Hein '17, etc.]

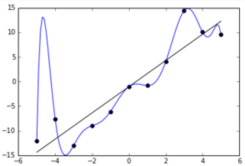
Learning Neural Networks is Hard

- Good News
  - When the size of the network is very large, no need to worry about bad local minima.
  - Every local minimum is a global minimum or close to a global minimum. [Choromanska et al. '15, Nguyen & Hein '17, etc.]
- Bad News
  - Typically over-parameterize
  - May lead to overfitting!!



Learning Neural Networks is Hard

- Good News
  - When the size of the network is very large, no need to worry about bad local minima.
  - Every local minimum is a global minimum or close to a global minimum. [Choromanska et al. '15, Nguyen & Hein '17, etc.]
- Bad News
  - Typically over-parameterize
  - May lead to overfitting!!



- Can we learn a neural net without over-parameterization?

Recover A Neural Network

- Assume the data follows a specified neural network model.
- Try to recover this model.

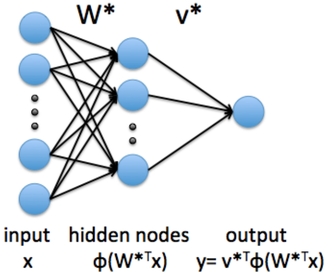
Model: One-hidden-layer Neural Network

Assume n samples $S = \{(\boldsymbol{x}_j, y_j)\}_{j=1,2,\cdots,n} \subset \mathbb{R}^d \times \mathbb{R}$ are sampled i.i.d. from distribution

$$\mathcal{D}: \qquad \boldsymbol{x} \sim \mathcal{N}(0, I), \quad y = \sum_{i=1}^{k} v_i^* \cdot \phi(\boldsymbol{w}_i^{*\top} \boldsymbol{x}),$$

where

- $\phi(z)$ is the activation function,
- k is the number of hidden nodes,
- $\{\boldsymbol{w}_i^*, v_i^*\}_{i=1,2,\cdots,k}$ are underlying ground truth parameters.



General Issues and Our Contribution

- Can we recover the model?
- How many samples are required? (Sample Complexity)
- And how much time? (Computational Complexity)

General Issues and Our Contribution

- Can we recover the model?
  - Yes, by gradient descent following tensor method initialization
- How many samples are required? (Sample Complexity)
- And how much time? (Computational Complexity)

General Issues and Our Contribution

- Can we recover the model?
  - Yes, by gradient descent following tensor method initialization
- How many samples are required? (Sample Complexity)
  - $|S| > d \cdot \log(1/\epsilon) \cdot \mathrm{poly}(k, \lambda)$, where ϵ is the precision and λ is a condition number of W^*.
- And how much time? (Computational Complexity)

General Issues and Our Contribution

- Can we recover the model?
  - Yes, by gradient descent following tensor method initialization
- How many samples are required? (Sample Complexity)
  - $|S| > d \cdot \log(1/\epsilon) \cdot \mathrm{poly}(k, \lambda)$, where ϵ is the precision and λ is a condition number of W^*.
- And how much time? (Computational Complexity)
  - $|S| \cdot d \cdot \mathrm{poly}(k, \lambda)$

General Issues and Our Contribution

- Can we recover the model?
  - Yes, by gradient descent following tensor method initialization
- How many samples are required? (Sample Complexity)
  - $|S| > d \cdot \log(1/\epsilon) \cdot \text{poly}(k, \lambda)$, where ϵ is the precision and λ is a condition number of W^*.
- And how much time? (Computational Complexity)
  - $|S| \cdot d \cdot \text{poly}(k, \lambda)$

The first recovery guarantee with both sample complexity and computational complexity linear in the input dimension and logarithmic in the precision.

Objective Function

- Given v_i^* and a sample set S, consider L2 loss

$$\widehat{f}_S(W) = \frac{1}{2|S|} \sum_{(\boldsymbol{x},y)\in S} \left(\sum_{i=1}^{k} v_i^* \phi(\boldsymbol{w}_i^\top \boldsymbol{x}) - y \right)^2.$$

Objective Function

- Given v_i^* and a sample set S, consider L2 loss

$$\widehat{f}_S(W) = \frac{1}{2|S|} \sum_{(\boldsymbol{x},y)\in S} \left(\sum_{i=1}^{k} v_i^* \phi(\boldsymbol{w}_i^\top \boldsymbol{x}) - y \right)^2.$$

- We show it is locally strongly convex near the ground truth!

Approach

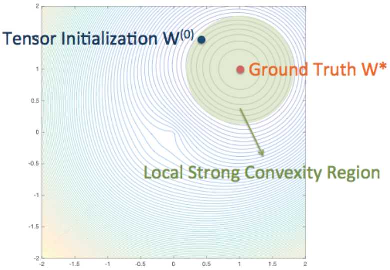
Algorithm:

1. Initialize $v_i = v_i^*$ exactly and W close to W^* by tensor methods → 2. Gradient descent

Corresponding Analysis:

Error bound for tensor decomposition

Local strong convexity & smoothness



Local Strong Convexity (LSC)

- $\nabla^2 f(W)$ is positive definite (p.d.) for $W \in \mathcal{A}$
  $\Rightarrow f(W)$ is LSC in area $\mathcal{A}$

Local Strong Convexity (LSC)

- $\nabla^2 f(W)$ is positive definite (p.d.) for $W \in \mathcal{A}$
  $\Rightarrow f(W)$ is LSC in area $\mathcal{A}$
- Consider the minimal eigenvalue of expected Hessian at ground truth,

$$\lambda_{\min}\big(\nabla^2 f_{\mathcal{D}}(W^*)\big) = \min_{\sum_j \|\boldsymbol{a}_j\|^2=1} \mathbb{E}\left[\left(\sum_j \phi'(\boldsymbol{w}_j^{*\top}\boldsymbol{x})\boldsymbol{x}^\top \boldsymbol{a}_j\right)^2\right]$$

where $f_{\mathcal{D}}$ is the expected risk.

Local Strong Convexity (LSC)

- $\nabla^2 f(W)$ is positive definite (p.d.) for $W \in \mathcal{A}$
  $\Rightarrow f(W)$ is LSC in area $\mathcal{A}$
- Consider the minimal eigenvalue of expected Hessian at ground truth,

$$\lambda_{\min}\big(\nabla^2 f_{\mathcal{D}}(W^*)\big) = \min_{\sum_j \|\boldsymbol{a}_j\|^2=1} \mathbb{E}\left[\left(\sum_j \phi'(\boldsymbol{w}_j^{*\top}\boldsymbol{x})\boldsymbol{x}^\top \boldsymbol{a}_j\right)^2\right]$$

  where $f_{\mathcal{D}}$ is the expected risk.
- $\lambda_{\min}\big(\nabla^2 f_{\mathcal{D}}(W^*)\big) \geq 0$ always holds.

Local Strong Convexity (LSC)

- $\nabla^2 f(W)$ is positive definite (p.d.) for $W \in \mathcal{A}$
  $\Rightarrow f(W)$ is LSC in area $\mathcal{A}$
- Consider the minimal eigenvalue of expected Hessian at ground truth,

$$\lambda_{\min}\big(\nabla^2 f_{\mathcal{D}}(W^*)\big) = \min_{\sum_j \|\boldsymbol{a}_j\|^2=1} \mathbb{E}\left[\left(\sum_j \phi'(\boldsymbol{w}_j^{*\top}\boldsymbol{x})\boldsymbol{x}^\top \boldsymbol{a}_j\right)^2\right]$$

  where $f_{\mathcal{D}}$ is the expected risk.
- $\lambda_{\min}\big(\nabla^2 f_{\mathcal{D}}(W^*)\big) \geq 0$ always holds.
- Does $\lambda_{\min}\big(\nabla^2 f_{\mathcal{D}}(W^*)\big) > 0$ always hold?

Local Strong Convexity (LSC)

- $\nabla^2 f(W)$ is positive definite (p.d.) for $W \in \mathcal{A}$
  $\Rightarrow f(W)$ is LSC in area $\mathcal{A}$
- Consider the minimal eigenvalue of expected Hessian at ground truth,

$$\lambda_{\min}\big(\nabla^2 f_{\mathcal{D}}(W^*)\big) = \min_{\sum_j \|\boldsymbol{a}_j\|^2=1} \mathbb{E}\left[\left(\sum_j \phi'(\boldsymbol{w}_j^{*\top}\boldsymbol{x})\boldsymbol{x}^\top \boldsymbol{a}_j\right)^2\right]$$

  where $f_{\mathcal{D}}$ is the expected risk.
- $\lambda_{\min}\big(\nabla^2 f_{\mathcal{D}}(W^*)\big) \geq 0$ always holds.
- Does $\lambda_{\min}\big(\nabla^2 f_{\mathcal{D}}(W^*)\big) > 0$ always hold? No

Two Examples when LSC doesn't Hold

- Set $v_i^* = 1$ and $W^* = I (k = d)$.

1. When $\phi(z) = z$,

$$\lambda_{\min}\big(\nabla^2 f_{\mathcal{D}}(W^*)\big) = \min_{\sum_j \|\boldsymbol{a}_j\|^2 = 1} \mathbb{E}\left[(\boldsymbol{x}^\top \sum_j \boldsymbol{a}_j)^2 \right] = 0$$

The minimum is achieved when $\sum_j \boldsymbol{a}_j = \mathbf{0}$

Two Examples when LSC doesn't Hold

- Set $v_i^* = 1$ and $W^* = I(k = d)$.

2. When $\phi(z) = z^2$,

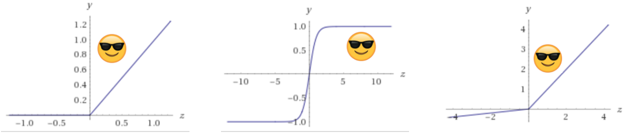
$$\lambda_{\min}\big(\nabla^2 f_{\mathcal{D}}(W^*)\big) = 4 \min_{\sum_j \|\boldsymbol{a}_j\|^2 = 1} \mathbb{E}\big[(\langle \boldsymbol{x}\boldsymbol{x}^\top, A\rangle)^2\big] = 0$$

where $A = [\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_d] \in \mathbb{R}^{d\times d}$.
The minimum is achieved when $A = -A^\top$.

When LSC Holds

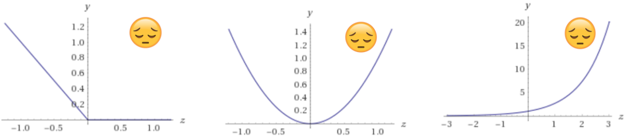
1. $\phi(z)$ satisfies three properties.

   P1 Non-negative and homogeneously bounded derivative
   $0 \leq \phi'(z) \leq L_1|z|^p$ for some constants $L_1 > 0$ and $p \geq 0$.



$\max(z, 0)$ $\tanh(z)$ $\max(z, 0.1z)$

Figure: activations satisfying P1



$\max(-z, 0)$ z^2 e^z

Figure: activations not satisfying P1

When LSC Holds

1. $\phi(z)$ satisfies three properties.

   P2 **"Non-linearity"**[1]

   For any $\sigma > 0$, we have $\rho(\sigma) > 0$, where

   $$\rho(\sigma) := \min\{\alpha_{2,0} - \alpha_{1,0}^2 - \alpha_{1,1}^2, \alpha_{2,2} - \alpha_{1,1}^2 - \alpha_{1,2}^2, \alpha_{1,0}\alpha_{1,2} - \alpha_{1,1}^2\}$$

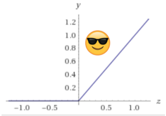
   and $\alpha_{i,j} := \mathbb{E}_{z\sim\mathcal{N}(0,1)}[(\phi'(\sigma z))^i z^j]$.

	ReLU	leaky ReLU	squared ReLU	erf	tanh	linear	quad-ratic
$\rho(0.1)$ $\rho(1)$ $\rho(10)$	0.091	0.089	0.27σ	1.9E-4 5.2E-2 2.5E-5	1.8E-4 4.9E-2 5.1E-5	0	0

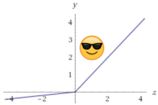
[1] Best name we can find... still need more understanding for $\rho(\sigma)$

When LSC Holds

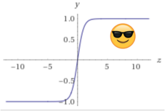
1. $\phi(z)$ satisfies three properties.
   P3 $\phi''(z)$ satisfies one of the following two properties,
   - *(a)* **Smoothness**
     $|\phi''(z)| \leq L_2$ for all z for some constant L_2, or
   - *(b)* **Piece-wise linearity**
     $\phi''(z) = 0$ except for e (e is a finite constant) points.



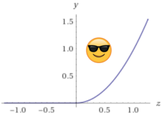
$\max(z, 0)$



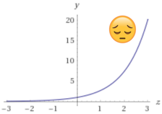
$\max(z, 0.1z)$



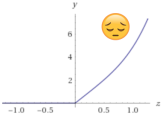
$\tanh(z)$



$\max(z, 0)^2$



e^z



$\phi(z) = 0$ if $z < 0$; $z^4 + 4z$ o.w.

Three Properties in Summary

P1 Non-negative and homogeneously bounded derivative

P2 "Non-linearity"

P3 (a) Smoothness, or (b)Piece-wise linearity

name	$\phi(z)$	P1	P2	P3.a	P3.b	P1,2,3
ReLU	$\max\{z, 0\}$	✓	✓	✗	✓	✓
leaky ReLU	$\max\{z, 0.01z\}$	✓	✓	✗	✓	✓
squared ReLU	$\max\{z, 0\}^2$	✓	✓	✓	✗	✓
sigmoid	$\frac{1}{1+e^{-z}}$	✓	✓	✓	✗	✓
tanh	$\frac{e^z-e^{-z}}{e^z+e^{-z}}$	✓	✓	✓	✗	✓
erf	$\int_0^z e^{-t^2} dt$	✓	✓	✓	✗	✓
linear	z	✓	✗	✓	✓	✗
quadratic	z^2	✗	✗	✓	✗	✗

Local Strong Convexity

Definition

Let $\sigma_i (i = 1, 2, \cdots, k)$ denote the i-th singular value of $W^* \in \mathbb{R}^{d \times k}$. Define $\kappa = \sigma_1 / \sigma_k$ and $\lambda = (\prod_{i=1}^k \sigma_i) / \sigma_k^k$.

Theorem

Let

1. $\phi(z)$ *satisfies Property 1,2,3 with* $\rho(\sigma_k)$
2. $|S| \geq d \cdot \mathrm{poly}(k, \lambda) / \rho^2(\sigma_k)$,
3. $\|W - W^*\| \leq \rho^2(\sigma_k) / \mathrm{poly}(\lambda, k)$.

Then there exist two positives $m_0 = \Theta(\rho(\sigma_k)/(\kappa^2 \lambda))$ *and* $M_0 = \Theta(k \sigma_1^{2p})$ *such that w.h.p.,*

$$m_0 I \preceq \nabla^2 \widehat{f}_S(W) \preceq M_0 I$$

Linear Convergence of Gradient Descent

For smooth activations, gradient descent has linear convergence.

Corollary

Let $\phi(z)$ satisfy Property 1,2,3(a) and $|S|$, W satisfy the conditions in the above theorem. Let

$$W^\dagger = W - \frac{1}{M_0} \nabla \widehat{f}_S(W),$$

then w.h.p.

$$\|W^\dagger - W^*\|_F^2 \leq (1 - \frac{m_0}{M_0}) \|W - W^*\|_F^2.$$

Initialization by Tensor Method

Definition

$\phi(z)$ is called q-homogeneous if $\phi(\sigma \cdot z) = \sigma^q \phi(z)$ for some constant q and any $\sigma > 0$.

Fact

If $(\boldsymbol{x}, y)$ *is sampled from*

$$\mathcal{D}: \qquad \boldsymbol{x} \sim \mathcal{N}(0, I), \quad y = \sum_i v_i^* \cdot \phi(\boldsymbol{w}_i^{*\top} \boldsymbol{x}),$$

and $\phi(z)$ *is* q*-homogeneous, then*

$$\mathbb{E}[y \cdot (\boldsymbol{x} \otimes \boldsymbol{x} \otimes \boldsymbol{x} - \boldsymbol{x} \widetilde{\otimes} I)] = \sum_i c \; v_i^* \|\boldsymbol{w}_i^*\|^{q-3} \boldsymbol{w}_i^* \otimes \boldsymbol{w}_i^* \otimes \boldsymbol{w}_i^*,$$

where $\boldsymbol{v} \widetilde{\otimes} I = \sum_{j=1}^d [\boldsymbol{v} \otimes \boldsymbol{e}_j \otimes \boldsymbol{e}_j + \boldsymbol{e}_j \otimes \boldsymbol{v} \otimes \boldsymbol{e}_j + \boldsymbol{e}_j \otimes \boldsymbol{e}_j \otimes \boldsymbol{v}]$.

Estimate Parameters Using Tensor Decomposition

- W.l.o.g. we can assume $v_i^* \in \{-1, 1\}$ due to the homogeneity.
- Setting $M_3 := \mathbb{E}[y \cdot (\boldsymbol{x} \otimes \boldsymbol{x} \otimes \boldsymbol{x} - \boldsymbol{x} \widetilde{\otimes} I)]$, we can
  1. Compute an empirical M_3, $\widehat{M_3}$, from samples.
  2. Do tensor decomposition on $\widehat{M_3}$.
  3. $v_i^* \in \{-1, 1\}$ can be exactly recovered and $\boldsymbol{w}_i^*$ can be approximated.

Overall Theoretical Guarantees

Theorem

Let the activation function be homogeneous satisfying Property 1, 2, 3(a). Then for any $\epsilon > 0$, if $|S| \geq \widetilde{O}(d \cdot \log(1/\epsilon) \cdot \mathrm{poly}(k, \lambda))$, the tensor method followed by gradient descent takes $\widetilde{O}(|S| \cdot d \cdot \mathrm{poly}(k, \lambda))$ time and outputs $\widehat{W}$ and $\widehat{\boldsymbol{v}}$ satisfying

$$\|\widehat{W} - W^*\|_F \leq O(\epsilon), \text{ and } \widehat{v}_i = v_i^*.$$

The proof mainly follows

- The matrix Bernstein inequality
- Error bound for non-orthogonal tensor decomposition from [Kuleshov-Chaganty-Liang'15]
- Linear convergence of gradient descent

Take-home Message and Future Work

- Take-home message
  1. The squared loss of one-hidden-layer neural nets is locally strongly convex near the ground truth w.r.t. the first-layer parameters.
  2. Tensor method is able to initialize the parameters into the local strong convexity region.
  3. Sample and computational complexities are linear in dim and logarithmic in precision.
- Future work
  1. One-hidden-layer nets have low capacity. –Multiple layers?
  2. Tensor method highly depends on Gaussian assumption. –Random Initialization?