Diverse Neural Network Learns True Target

Le Song

Associate Professor, College of Computing

Associate Director, Center for Machine Learning

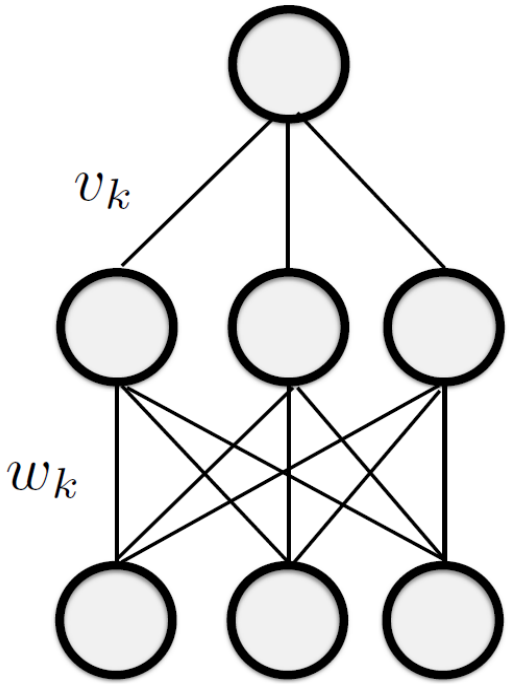
Georgia Institute of Technology

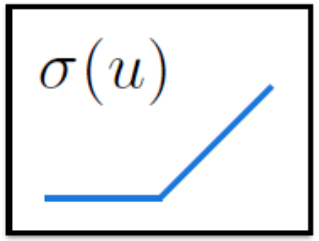
Principal Engineer, Ant Financial, Alibaba

(Joint work with Bo Xie @ Facebook & Yingyu Liang @ UW Madison)

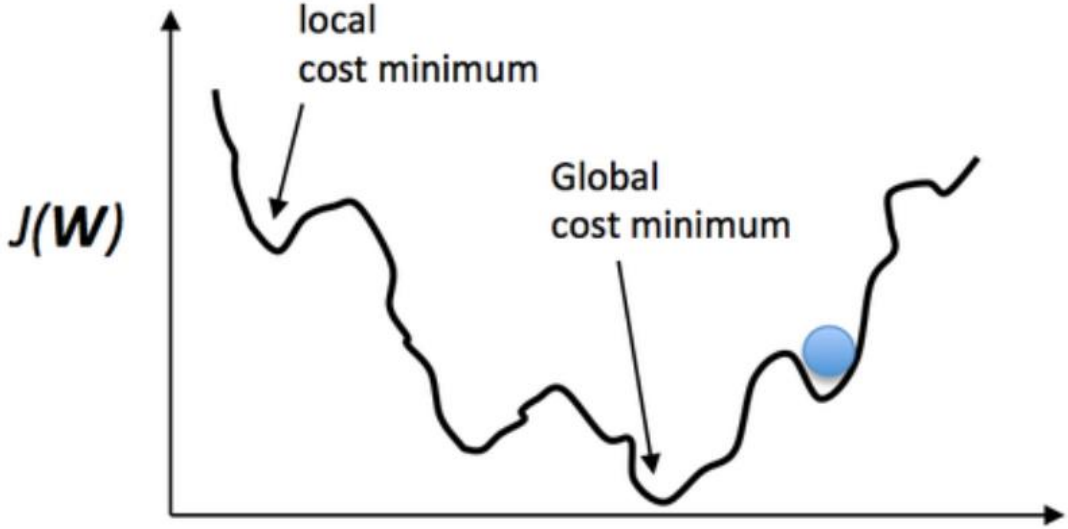
Neural networks learning

- Neural networks extremely successful in learning many nonlinear functions
- Most are trained with simple Gradient Descent (GD) or Stochastic Gradient Descent (SGD)
- Highly non-convex objective function.
  Why GD/SGD work so well?





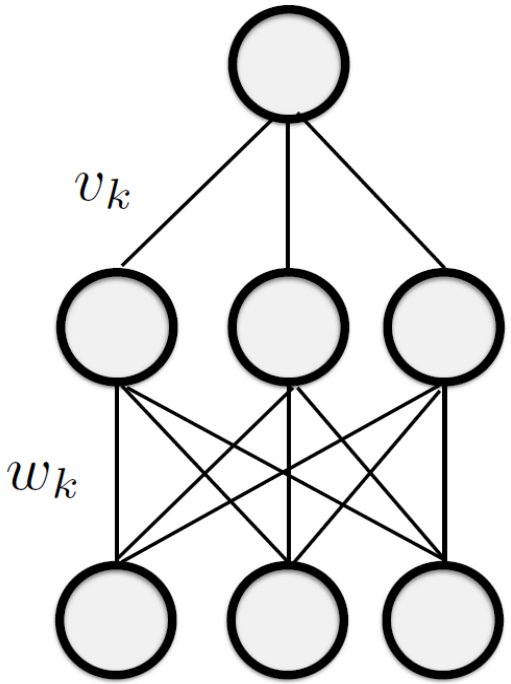
$$\sigma(u) = \max\{0, u\}$$



Problem setting

- One-hidden-layer neural networks with ReLU activation ($v_k \sim \{\pm 1\}$ uniformly, $\|x\|_2 = 1$)

$$f(x) = \sum_{k=1}^{n} v_k \sigma(w_k^\top x)$$

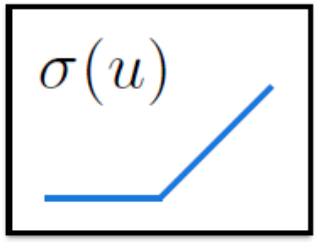


- Least-squares loss

$$L(f) = \frac{1}{2m} \sum_{l=1}^{m} \big(y_l - f(x_l)\big)^2$$

- Main results

For nice neural weights, with high probability, any stationary point is a global optimum



$\sigma(u) = \max\{0, u\}$

Structure of the gradient

- Gradient w.r.t. first layer weights

$$\frac{\partial L}{\partial w_k} = \frac{1}{m}\sum_{l=1}^{m}(f(x_l) - y_l)v_k\sigma'\left(w_k^\top x_l\right)x_l$$

$$\begin{pmatrix} \frac{\partial L}{\partial w_1} \\ \cdots \\ \frac{\partial L}{\partial w_k} \\ \cdots \\ \frac{\partial L}{\partial w_n} \end{pmatrix} = \begin{pmatrix} v_1\sigma'(w_1^\top x_1)x_1 & \cdots & v_1\sigma'(w_1^\top x_m)x_m \\ \cdots & \cdots & \cdots \\ v_k\sigma'(w_k^\top x_1)x_1 & \cdots & v_k\sigma'(w_k^\top x_m)x_m \\ \cdots & \cdots & \cdots \\ v_n\sigma'(w_n^\top x_1)x_1 & \cdots & v_n\sigma'(w_n^\top x_m)x_m \end{pmatrix} \times \frac{1}{m}\begin{pmatrix} f(x_1) - y_1 \\ \cdots \\ f(x_k) - y_k \\ \cdots \\ f(x_m) - y_m \end{pmatrix}$$

$$\frac{\partial L}{\partial W} = D\ r$$

Structure of the gradient

- Gradient w.r.t. first layer weights

$$\frac{\partial L}{\partial w_k} = \frac{1}{m}\sum_{l=1}^{m}(f(x_l) - y_l)v_k\sigma'\left(w_k^\top x_l\right)x_l$$

$$\begin{pmatrix} \frac{\partial L}{\partial w_1} \\ \cdots \\ \frac{\partial L}{\partial w_k} \\ \cdots \\ \frac{\partial L}{\partial w_n} \end{pmatrix} = \begin{pmatrix} v_1\sigma'(w_1^\top x_1)x_1 & \cdots & v_1\sigma'(w_1^\top x_m)x_m \\ \cdots & \cdots & \cdots \\ v_k\sigma'(w_k^\top x_1)x_1 & \cdots & v_k\sigma'(w_k^\top x_m)x_m \\ \cdots & \cdots & \cdots \\ v_n\sigma'(w_n^\top x_1)x_1 & \cdots & v_n\sigma'(w_n^\top x_m)x_m \end{pmatrix} \times \frac{1}{m}\begin{pmatrix} f(x_1) - y_1 \\ \cdots \\ f(x_k) - y_k \\ \cdots \\ f(x_m) - y_m \end{pmatrix}$$

$$\frac{\partial L}{\partial W} = D\,r$$

non-singular?

Intuition

- Key inequality

$$\|r\| \leq \frac{1}{s_m(D)} \left\| \frac{\partial L}{\partial W} \right\|$$

Training error

Minimum singular value

Norm of gradient

Intuition

- Key inequality

$$\|r\| \leq \frac{1}{s_m(D)} \left\| \frac{\partial L}{\partial W} \right\|$$

Training error

Minimum singular value

Norm of gradient

- Need to lower bound minimum singular value

- Directly analyze the singular value

$G_n = D^\top D/n$ a function of the weights; difficult to analyze

Bounding the error

- Key inequality

$$\|r\| \leq \frac{1}{s_m(D)} \left\| \frac{\partial L}{\partial W} \right\|$$

Training error

Minimum singular value

Norm of gradient

- Need to lower bound minimum singular value

- Directly analyze the singular value

$$G_n = D^\top D/n \qquad G = \mathbb{E}_w[G_n]$$

introduce an intermediate variable that does not depend on weights

$w \sim N(0, I)$

Bounding the error

- Key inequality

$$\|r\| \leq \frac{1}{s_m(D)} \left\| \frac{\partial L}{\partial W} \right\|$$

Training error

Minimum singular value

Norm of gradient

- Need to lower bound minimum singular value

- Directly analyze the singular value

$$G_n = D^\top D/n \qquad G = \mathbb{E}_w[G_n]$$

- Decompose into two parts

$$\lambda_m(G_n) \geq \underbrace{\lambda_m(G)}_{\text{I. Ideal spectrum}} - \underbrace{\|G - G_n\|}_{\text{II. discrepancy}}$$

Bounding the first term

- Kernel function associated with ReLU ($w \sim N(0, I)$)

$$G_{ij} = \mathbb{E}_w\left[\sigma'(w^\top x_i)\sigma'\left(w^\top x_j\right)\right]\langle x_i, x_j\rangle$$

$$= \left(\frac{1}{2} - \frac{\arccos\langle x_i, x_j\rangle}{2\pi}\right)\langle x_i, x_j\rangle$$

$$= \sum_{u=1}^{\infty} \gamma_u \phi_u(x_i)\phi_u(x_j)$$

spherical harmonics decomposition

Bounding the first term

- Kernel function associated with ReLU ($w \sim N(0, I)$)

$$G_{ij} = \mathbb{E}_w\left[\sigma'(w^\top x_i)\sigma'\left(w^\top x_j\right)\right]\langle x_i, x_j\rangle$$

$$= \left(\frac{1}{2} - \frac{\arccos\langle x_i, x_j\rangle}{2\pi}\right)\langle x_i, x_j\rangle$$

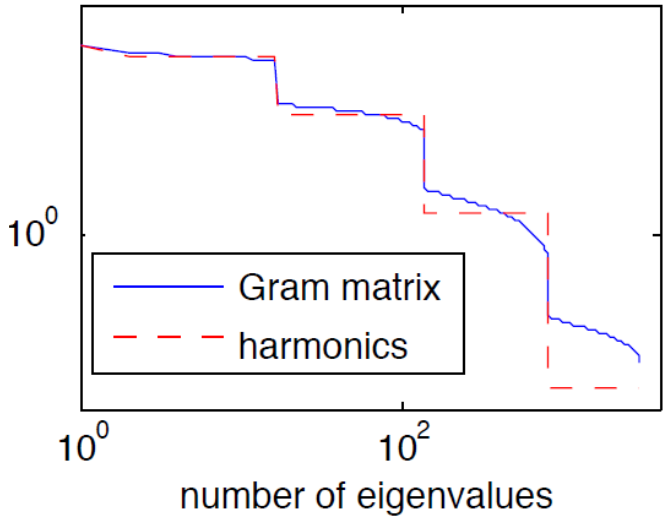
$$= \sum_{u=1}^{\infty} \gamma_u \phi_u(x_i)\phi_u(x_j)$$

With high probability

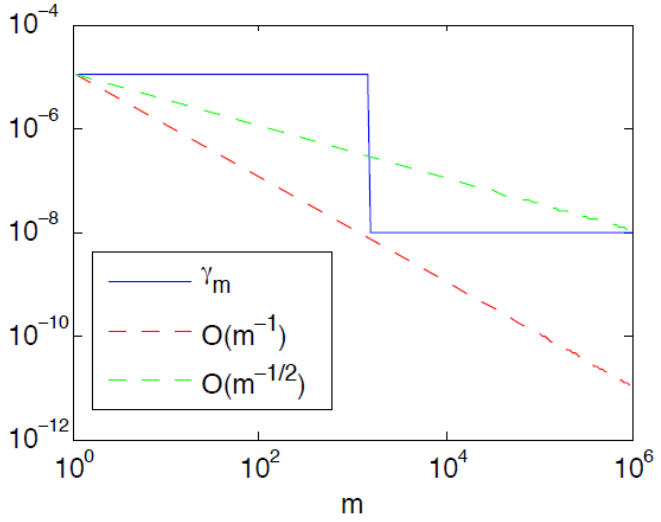
$$\boxed{\lambda_m(G) \geq m\frac{\gamma_m}{2}}$$

The spectrum of ReLU in between $O\left(\frac{1}{m}\right)$ and $O\left(\frac{1}{\sqrt{m}}\right)$

Bounding the first term



Spectrum of G
concentrates around γ_m



The spectrum of ReLU

With high probability

$$\lambda_m(G) \geq m\frac{\gamma_m}{2}$$

The spectrum of ReLU in between $O\left(\frac{1}{m}\right)$ and $O\left(\frac{1}{\sqrt{m}}\right)$

Bounding the second term

- The difference between true weights and the expectation

$$\|G - G_n\| \leq O(\rho\, L_2(W))$$

Weight discrepancy

$$(L_2(W))^2 = \frac{1}{n^2}\sum_{i,j=1}^{n} k\left(w_i, w_j\right)^2 - \mathbb{E}_{u,v}[k(u,v)^2]$$

Where

$$k(x,y) = \frac{1}{2} - \frac{\arccos\langle x, y\rangle}{2\pi}$$

Small discrepancy implies more diverse weights

A bound on the minimum singular value

- With high probability

$$s_m(D)^2 \geq \frac{nm\gamma_m}{2} - cn\rho L_2(W)$$

A simplified result

- With high probability

$$s_m(D)^2 \geq \frac{nm\gamma_m}{2} - cn\rho L_2(W)$$

- Suppose n and d are large enough and weight discrepancy is small

$$n = \widetilde{\Omega}\left(\frac{1}{\gamma_m}\right), \qquad d = \widetilde{\Omega}\left(\frac{1}{\gamma_m}\right), \qquad L_2(W) = \tilde{O}\left(\frac{1}{n^{1/4}d^{1/4}}\right)$$

- Then with high probability

Satisfied by random weights

$$s_m(D)^2 \geq \Omega(m)$$

Final error bound

For n and d large enough

For any W that has small weight discrepancy

With high probability

$$\frac{1}{2m}\sum_{l=1}^{m}(f(x_l)-y_l)^2 \leq O\left(\left\|\frac{\partial L}{\partial W}\right\|^2\right)$$

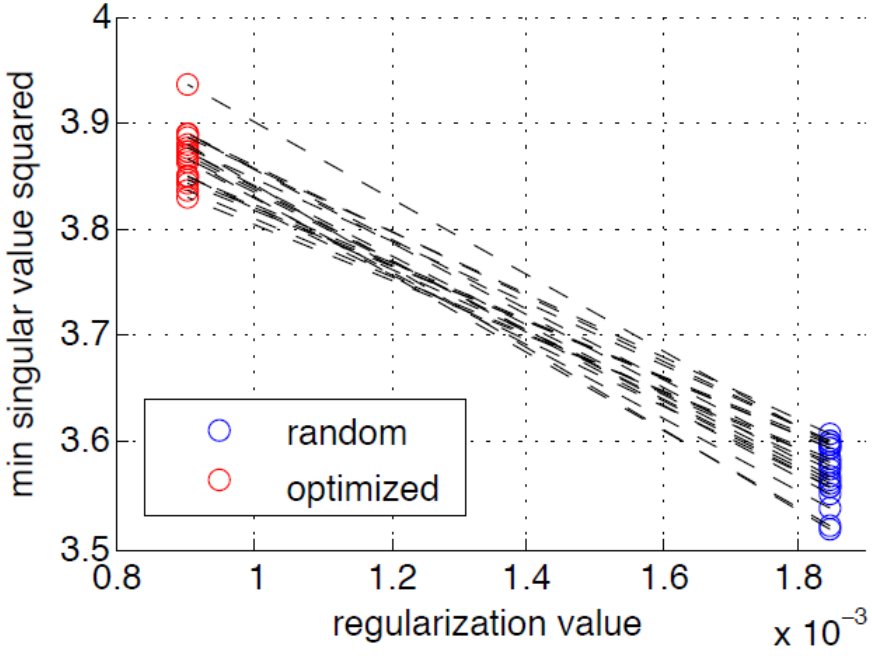
Most W satisfy small enough weight discrepancy

Small gradient means small error!

A regularization term

- To enforce diverse weights, we can use a regularization term

$$R(W) = \frac{1}{n(n-1)} \sum_{i \neq j}^{n} k(w_i, w_j)^2$$



	$n = 100$		$n = 150$	
	train	test	train	test
no-reg	15.42(5.86)	14.80(5.36)	1.79(0.45)	1.86(0.50)
reg	**11.32(1.77)**	**10.63(1.58)**	**1.07(0.84)**	**1.13(0.99)**

	$n = 200$		$n = 300$	
	train	test	train	test
no-reg	**0.38(0.27)**	**0.44(0.35)**	0.39(0.39)	0.44(0.40)
reg	0.50(0.51)	0.58(0.59)	**0.10(0.05)**	**0.12(0.07)**

Recap

- Small weight discrepancy + gradient zero $\approx$ global optimal
  - Random initialization has
    - small weight discrepancy
    - large singular value of the extended feature matrix

- What will gradient descent (GD) do to the random initialization?

- Recent Arxiv draft:
  - S. Du, X. Zha, B. Poczos, and A. Singh. Gradient Descent Provably Optimizes Over-parameterized Neural Networks
    - Over-parameterization and random initialization jointly restrict every weight vector to be close to its initialization for all iterations
  - D. Boob & G. Lan. Theoretical properties of the global optimizer of two layer neural network

Continuous time analysis

- Gradient descent dynamics

$$\frac{dw_i(t)}{dt} = -\frac{dL}{dw_i(t)}$$

- Function prediction dynamics

$$\begin{pmatrix} \frac{df(x_1)}{dt} \\ \\ \frac{df(x_m)}{dt} \end{pmatrix} = D(t)^\top D(t) r(t)$$

If initialize $w(0) \sim N(0, I)$,

$$D(0)^\top D(0) = n\, G_n(0), \qquad G(0) = \mathbb{E}_w[G_n(0)]$$

The spectrum of $D(0)^\top D(0)$ can be lower bounded by λ_0 using spherical harmonics and concentration

Self-consistent argument

- Suppose for a given t, $\|w_i(t) - w_i(0)\|_2 \leq \frac{c\lambda_0}{m^2} = R, \forall i$. Then w.h.p

$$\lambda_{min}\left(nG_n(t)\right) > \frac{\lambda_0}{2}$$

Matrix perturbation

- Suppose for all $0 \leq s \leq t, \lambda_{min}\left(nG_n(s)\right) > \frac{\lambda_0}{2}$. Then

$$\|r(t)\|_2^2 \leq \exp(-\lambda_0 t)\,\|r(0)\|_2^2$$

$$\|w_i(t) - w_i(0)\|_2 \leq \frac{2\sqrt{m}\|r(0)\|_2}{\sqrt{n}\lambda_0} = R', \forall i$$

Dynamical system

- If we set $n = \Omega\left(\frac{m^5\|r(0)\|_2}{\lambda_0^4}\right)$, $R' < R$.

Self consistent

Overparametrization allows gradient decent to maintain weight discprency

Conclusion & future work

- Random initialization + overparametrization $\approx$ global optimal
- Sharper bound?
- What happens for other loss function?
- What happens for multi-layer neural networks?
- How about other optimization algorithms?
- How about stochastic gradients?