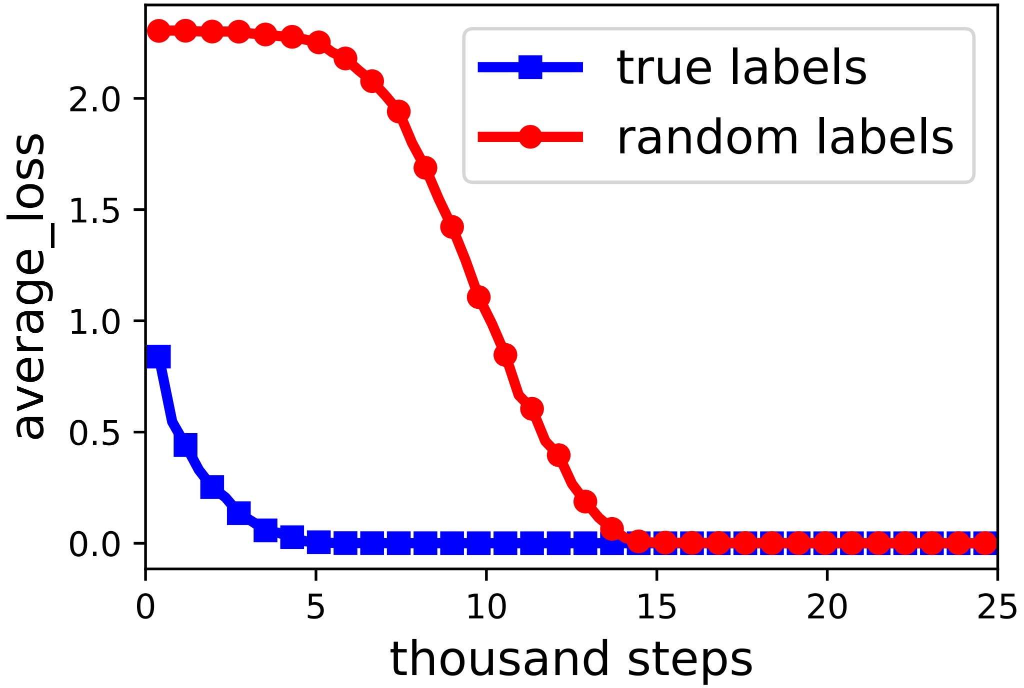
Gradient Descent Finds Global Minima of Deep Neural Networks

Simon S. Du, Jason D. Lee, Haochuan Li, Liwei Wang, Xiyu Zhai

Empirical Observations on Empirical Risk

- Zhang et al, 2017, Understanding Deep Learning Requires Rethinking Generalization.



Randomization Test: replace true labels by random labels.

Observations: Empirical Risk-> 0 for both true labels and random labels.

Conjecture: because neural networks are over-parameterized.

Open Problem: why gradient descent can find a neural network that fits all labels.

Setup

- Training Data: $\{x_i, y_i\}_{i=1}^{n}, x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$
- A Model.
  - Fully connected neural network:

$$f(\theta, x) = W_L \sigma(W_{L-1} \cdots W_2 \sigma(W_1 x) \cdots)$$

- A loss function.
  - Quadratic loss:

$$R(\theta) = \frac{1}{2n} \sum_{i=1}^{n} (f(\theta, x_i) - y_i)^2$$

- An optimization algorithm:
  - Gradient descent:

$$\theta(t+1) \leftarrow \theta(t) - \eta \frac{\partial R(\theta(t))}{\partial \theta(t)}$$

Trajectory-based Analysis

$$\theta(t+1) \leftarrow \theta(t) - \eta \frac{\partial R(\theta(t))}{\partial \theta(t)}$$

- Trajectory of parameters:

$$\theta(0), \theta(1), \theta(2), \cdots$$

- Predictions:

$$u_i(t) \triangleq f(\theta(t), x_i), u(t) \triangleq (u_1(t), \ldots, u_n(t))^\top \in \mathbb{R}^n$$

- Trajectory of predictions:

$$u(0), u(1), u(2), \ldots$$

Proof Sketch

- Simplified form (continuous time):

$$\frac{du(t)}{dt} = -\sum_{\ell=1}^{L} H^{\ell}(t)\left(y - u(t)\right) \quad H_{ij}^{\ell}(t) = \frac{1}{n}\langle \frac{\partial u_i(t)}{\partial W_\ell(t)}, \frac{\partial u_j(t)}{\partial W_\ell(t)} \rangle$$

- Random initialization + concentration + perturbation analysis:

$$\lim_{m\to\infty} \sum_{\ell=1}^{L} H^{\ell}(0) \to H^{\infty} \quad \lim_{m\to\infty} \sum_{\ell=1}^{L} H^{\ell}(t) \to \sum_{\ell=1}^{L} H^{\ell}(0), \forall t \geq 0$$

- Linear ODE theory:

$$\|u(t) - y\|_2^2 \leq \exp\left(-\lambda_0 t\right) \|u(0) - y\|_2^2, \lambda_0 = \lambda_{\min}\left(H^{\infty}\right)$$

Main Results

Theorem 1: For fully-connected neural network with smooth activation, if $m = \mathrm{poly}(n, 2^L, 1/\lambda_0)$ and step size $\eta = O\left(\frac{\lambda_0}{n^2 2^{\Omega(L)}}\right)$, then with high probability over random initialization we have: for $t = 1,2, \ldots$

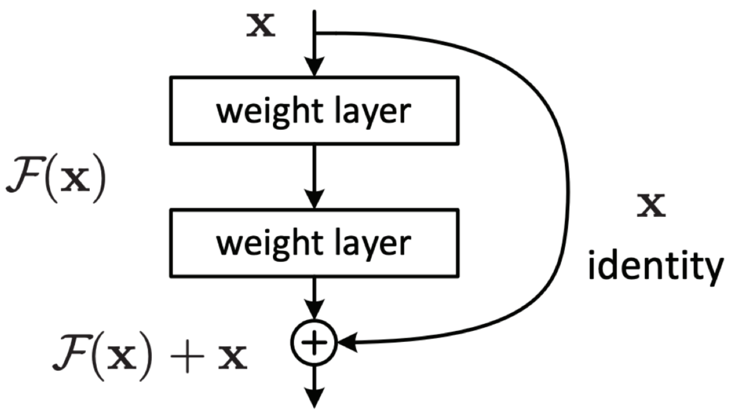
$$R\big(\theta(t)\big) \leq (1 - \eta\lambda_0)^t R(\theta(0)).$$

- First global linear convergence guarantee for deep NN.
- Exponential dependence due to error propagation.

Main Results (Cont'd)

Theorem 2: For ResNet or Convolutional ResNet with smooth activation, if $m = \mathrm{poly}(n, L, 1/\lambda_0)$ and step size $\eta = O\left(\frac{\lambda_0}{n^2}\right)$, then with high probability over random initialization we have: for $t = 1,2,\ldots$

$$R\big(\theta(t)\big) \leq (1 - \eta\lambda_0)^t R(\theta(0)).$$



- ResNet architecture makes the error propagation more stable => exponential improvement over fully-connected neural networks.

Learn more @ Pacific Ball Room #80