neural networks

2020年10月21日 9:13

Parameters $W_1, W_2, \cdots, W_L$

↓

Non-linear $f(x) = W_L \sigma(W_{L-1}\sigma(\cdots \sigma(W_1 x)))$

Three main theory problems

- Representation ✓
- Optimization ✓
- Generalization ?

Overparameterized two layer network

- kernel methods (traditional view)

$$f_{W,a}(x_i) = \sum_{r=1}^{m} a_r \cdot \sigma(W_r^T x_i)$$

m is huge (overpara..)

Init: $W_r(0) \sim \mathcal{N}(0,1)$, then FIX W. (random kernel)

$$\Rightarrow L(a) = \frac{1}{2}\sum_{i=1}^{n}(f_{W,a}(x_i) - y_i)^2$$

$$= \frac{1}{2}\|\underbrace{\sigma(XW^T)}_{\Phi = \sigma(XW^T)} a - y\|_2^2$$

secured by "overparameterized"

$L(a) = \frac{1}{2}\|\Phi a - y\|_2^2$ As long as Φ is full rank, we can optimize it.

↓ Changing Kernels

Fix $a_r = 1$, change w.

Analysis: $\frac{\partial L(W(t))}{\partial W_r} = \sum_{i=1}^{n}(f(x_i) - y_i)\, I_{r,i}\, x_i$

$I_{r,i} = \mathbb{I}[W_r x_i \geq 0]$

Define a "filter" matrix $Z_{r,i} = I_{r,i} x_i$

$\Rightarrow \nabla L(W) = Z(f(x) - y)$.

Analysis

Analysis

- W update

$$W(t+1) = W(t) - \eta Z(t)(f_{W(t)}(X) - y)$$

- $f_W(X)$ update.

$$f_{W(t+1)}(X) = f_{W(t)}(X) - \eta \underbrace{Z^T(t) Z(t)}_{= H(t)} (f_{W(t)}(X) - y)$$

$$\downarrow$$

$$f_{W(t+1)}(X) - y = \underbrace{(I - \eta H(t))}_{\text{almost unchanged since } W \text{ does not change much}} (f_{W(t)}(X) - y).$$

$$\Rightarrow f_{W(t+1)}(X) - y \approx (I - \eta H(0))(f_{W(t)}(X) - y)$$

$$\| f_{W(t)}(X) - y \|_2 = \sqrt{\sum_{i=1}^{n} (1 - \eta \lambda_i)^{2t} (v_i^T y)^2} \pm \varepsilon$$

(where $H(0) v_i = \lambda_i v_i$)

$$f_{W,a}(X) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \sigma(W_r^T x).$$

$W_r(0) \sim \mathcal{N}(0, k^2 I)$. $a_r \sim \text{unif}(\{-1, 1\})$.

$$\downarrow \quad w_r(k+1) = w_r(k) - \eta \frac{a_r}{\sqrt{m}} \sum_{i=1}^{n} (f_{W(k), a}(x_i) - y_i) \mathbb{I}\{w_r(k)^T x_i \geq 0\} x_i$$

Define H^∞.

$$H_{ij}^\infty := \mathbb{E}_{w \sim \mathcal{N}(0, I)} [x_i^T x_j I\{w^T x_i \geq 0, w^T x_j \geq 0\}]$$