Finite-Sample Analysis of Learning High-Dim ReLU Neuron

Jingfeng Wu
with Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, Sham Kakade

High-dim ReLU regression

Linear

Minimize $\quad R(\mathbf{w}) = \mathbb{E}\big(\mathrm{ReLU}(\mathbf{x}^\top \mathbf{w}) - y\big)^2, \quad \mathbf{w} \in \mathbb{R}^d$

With n samples (iid): $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$

Distributional assumption (simplified)

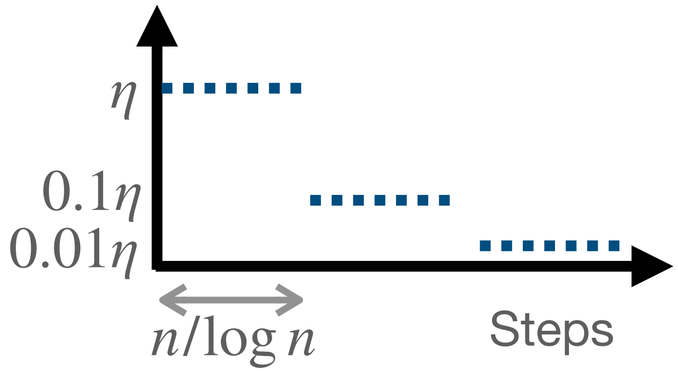
$$y = \mathrm{ReLU}(\mathbf{x}^\top \mathbf{w}_*) + \mathcal{N}(0,1), \quad \|\mathbf{w}_*\|_2 \le 1, \quad \mathbf{x} \sim \mathcal{N}(0, \mathbf{H})$$

$(\lambda_i)_{i \ge 1}$ denote eigenvalues of $\mathbf{H} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$

Learning high-dim linear regression

by (online) SGD

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \eta_t \cdot (\mathbf{x}_t^\top \mathbf{w}_{t-1} - y_t) \cdot \mathbf{x}_t, \quad t = 1, \ldots, n$$



[ZWBGK'21, WZBGK'22]

$$\frac{\texttt{DIM}}{n} \lesssim \mathbb{E}R(\mathbf{w}_n) - \texttt{OPT} \lesssim \frac{\texttt{DIM}}{n}$$

$$\texttt{DIM} := \#\left\{ \lambda_i \geq \frac{1}{n\eta_0} \right\} + n^2\eta_0^2 \cdot \sum_{\lambda_i < \frac{1}{n\eta_0}} \lambda_i^2 \qquad \ll d$$

Learning high-dim ReLU regression

by (online) SGD

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \eta_t \cdot \left(\mathtt{ReLU}(\mathbf{x}_t^\top \mathbf{w}_{t-1}) - y_t\right) \cdot \mathbf{x}_t \cdot 1_{[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]}, \quad t = 1, \ldots, n$$

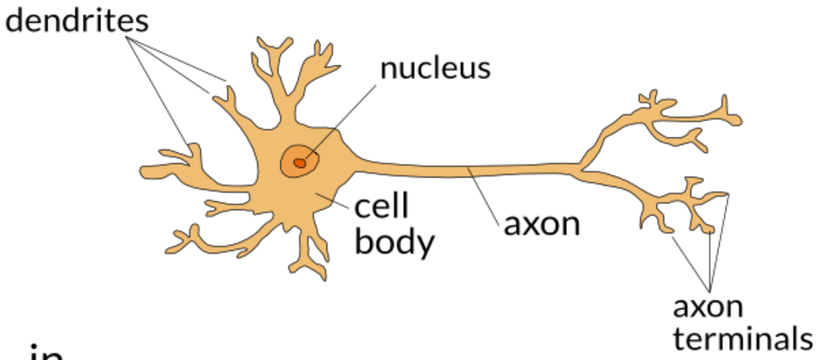
not a good idea

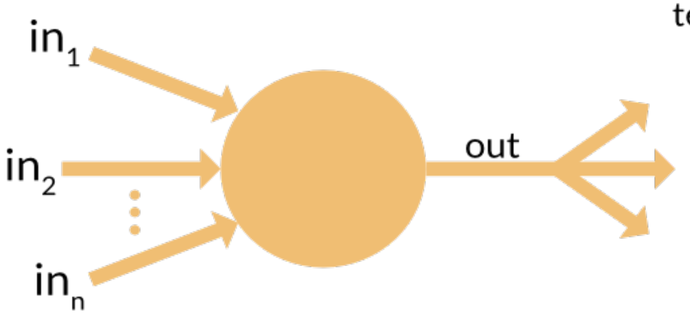
by (online) Perceptron

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \eta_t \cdot \left(\mathtt{ReLU}(\mathbf{x}_t^\top \mathbf{w}_{t-1}) - y_t\right) \cdot \mathbf{x}_t,$$

$$t = 1, \ldots, n$$

our main focus





Result 1: well-specified ReLU regression

$$y = \texttt{ReLU}(\mathbf{x}^\top \mathbf{w}_*) + \mathcal{N}(0,1), \quad \|\mathbf{w}_*\|_2 \le 1, \quad \mathbf{x} \sim \mathcal{N}(0, \mathbf{H})$$

[This work]

$$\frac{\texttt{DIM}}{n} \lesssim \mathbb{E}R(\mathbf{w}_n) - \texttt{OPT} \lesssim \frac{\texttt{DIM}}{n}$$

- Linear reg. bounds port over to ReLU reg.
- Point-wise tight, dim-free, …
- **ReLU isn’t harder!**

misspecified

Result 2: well-specified ReLU regression

$$y = \mathtt{ReLU}(\mathbf{x}^\top \mathbf{w}_*) + \mathcal{N}(0,1), \quad \|\mathbf{w}_*\|_2 \le 1, \quad \mathbf{x} \sim \mathcal{N}(0, \mathbf{H})$$

[This work]

$$\mathbb{E}R(\mathbf{w}_n) \lesssim \mathtt{OPT} + \frac{\mathtt{DIMDIM}}{n} \lesssim \mathbb{E}R(\mathbf{w}_n) - \mathtt{OPT} \lesssim \frac{\mathtt{DIM}}{n}$$

- $\mathbb{E}R(\mathbf{w}) \le \mathtt{OPT} + o(1)$ bound is believed to be "impossible" [GKK'19]
- Previously, best known bound $\mathcal{O}(\mathtt{OPT} + \sqrt{d/n})$ [DGKKS'20]
- A *dim-free,* const. factor approx.

What's more

- In well-specified cases, bounds are ***tight up to const. factor*** as functions of

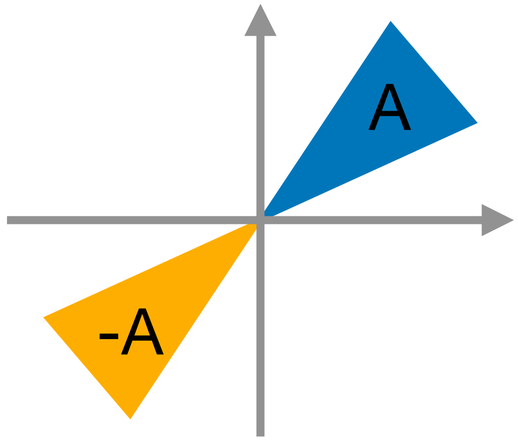
$$\eta,\ n,\ \mathbf{H},\ \mathbf{w}_*,\ \text{and noise scale}$$

- $\mathbf{x} \sim \mathcal{N}(0, \mathbf{H})$ can be relaxed to

  1. hypercontractivity

  2. moments symmetricity

$$\forall \mathbf{v},\ \mathbb{E}(\mathbf{x}^\top \mathbf{v})^4 \le \alpha\left(\mathbb{E}(\mathbf{x}^\top \mathbf{v})^2\right)^2$$



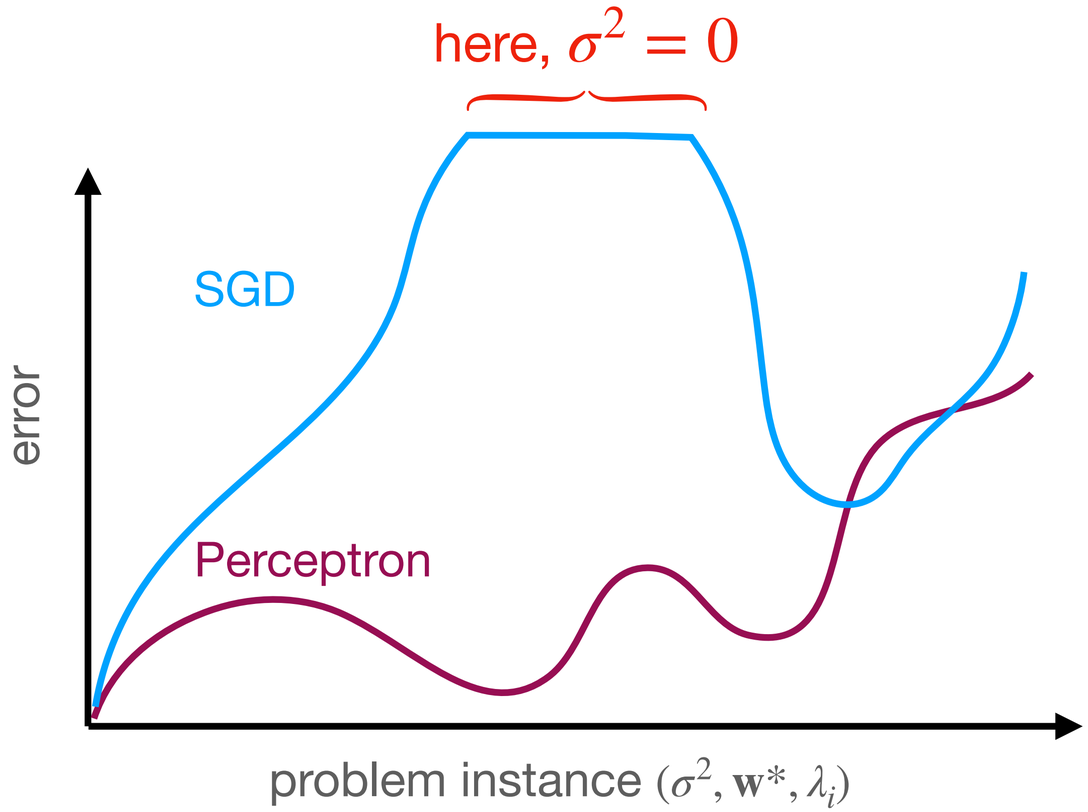
$$\mathbb{E}f(\mathbf{x})\mathbf{1}[\mathbf{x} \in A] = \mathbb{E}f(\mathbf{x})\mathbf{1}[\mathbf{x} \in -A]$$

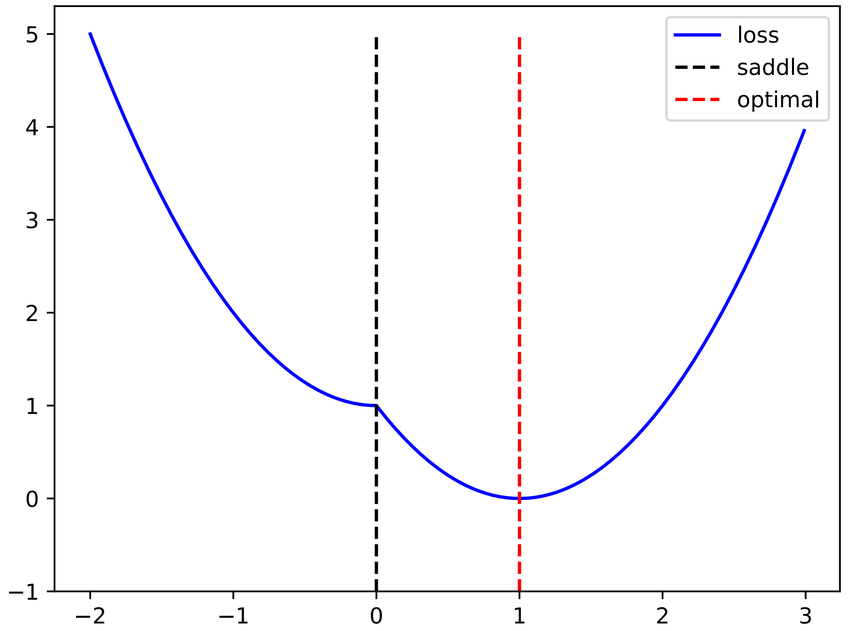
for $f(\mathbf{x}) = \mathbf{x}^{\otimes 2}$ or $\mathbf{x}^{\otimes 4}$

Result 3: SGD is less effective than Perceptron

$$y = \mathtt{ReLU}(\mathbf{x}^\top \mathbf{w}_*) + \mathcal{N}(0, \sigma^2), \qquad \|\mathbf{w}_*\|_2 \le 1,$$

$$\mathbb{P}\{\mathbf{x} = \mathbf{e}_i\} = \mathbb{P}\{\mathbf{x} = -\mathbf{e}_i\} = \frac{\lambda_i}{2}$$





Non-convex landscape

Take Home: ReLU Neuron

1. Well-specified noise
   - ReLU reg. $\approx$ linear reg, $\texttt{OPT} + \mathcal{O}(\texttt{DIM}/n)$ bound
   - **Perceptron**, allowing $d > n$
2. Misspecified noise
   - $\mathcal{O}(\texttt{OPT} + \texttt{DIM}/n)$ bound
   - **Perceptron**, allowing $d > n$
3. Symmetric Bernoulli data
   - SGD no better than Perceptron when noise is well-specified
   - SGD suffers const. error if noiseless

Difan Zou

Zixiang Chen

Vladimir Bravermen

Quanquan Gu

Sham Kakade