Finite-Sample Analysis of Learning High-Dimensional Single ReLU Neuron

Jingfeng Wu*[1], Difan Zou*[2], Zixiang Chen*[3], Vladimir Braverman[4], Quanquan Gu[3], Sham Kakade[5]

[1]Johns Hopkins University, [2]The University of Hong Kong, [3]UCLA, [4]Rice University, [5]Harvard University

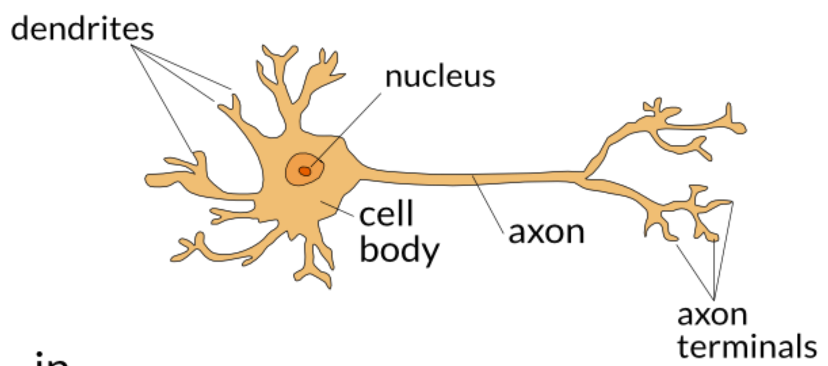
ICML International Conference On Machine Learning

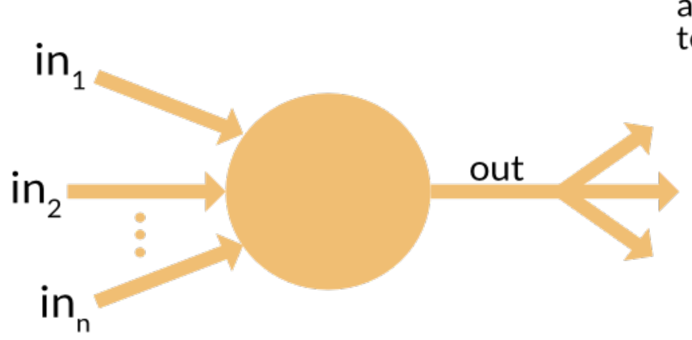
ReLU Regression

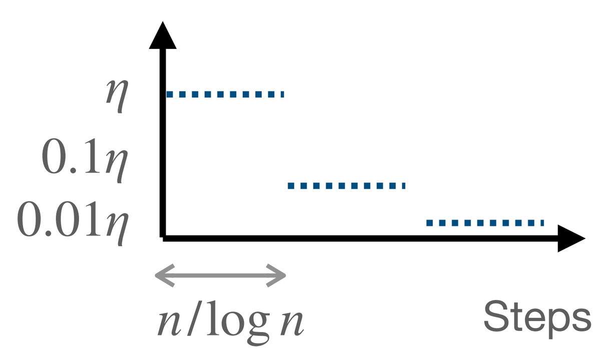
Minimize $R(\mathbf{w}) = \mathbb{E}\big(\underbrace{\mathrm{ReLU}}_{\text{if removed => linear regression}}(\mathbf{x}^\top\mathbf{w}) - y\big)^2, \quad \mathbf{w} \in \mathbb{R}^d$

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

overparameterization: $d > n$







(Online) SGD

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

(Online) GLM-tron [KKSK'11]

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

Notation

$\mathrm{OPT} := \min R(\cdot)$

$\mathbf{w}^* \in \arg\min R(\cdot)$

$\mathbf{H} := \mathbb{E}\mathbf{x}\mathbf{x}^\top$, eigenvalues denoted by $(\lambda_i)_{i\ge 1}$

Prior Works on Linear Regression

Simplified Bounds of SGD [ZWBGK'21, WZBGK'22]

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

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

Bounds $\quad \dfrac{D_{\mathrm{eff}}}{N_{\mathrm{eff}}} \lesssim \mathbb{E}R(\mathbf{w}_n) - \mathrm{OPT} \lesssim \dfrac{D_{\mathrm{eff}}}{N_{\mathrm{eff}}}$

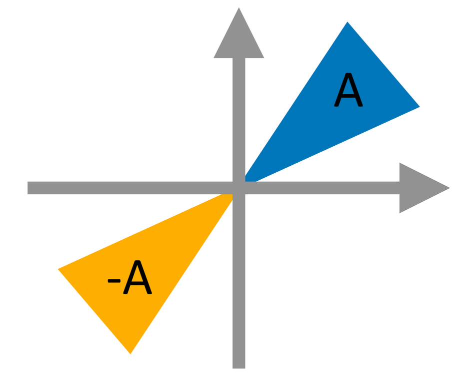
Result 1: Well-Specified Case

Distributional Assumptions

1. Well-specified noise $\quad y = \mathrm{ReLU}(\mathbf{x}^\top \mathbf{w}_*) + \mathcal{N}(0, \sigma^2)$
2. Hypercontractivity $\quad \mathbb{E}\langle \mathbf{v}, \mathbf{x}\rangle^4 \le \alpha\big(\mathbb{E}\langle \mathbf{v}, \mathbf{x}\rangle^2\big)^2$
3. Symmetric moments

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

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



Risk Bound of GLM-tron

Suppose $\eta_0 < 1/(4\alpha \mathrm{tr}(\mathbf{H}))$. Then

$$\mathbb{E}R(\mathbf{w}_n) - \mathrm{OPT} \lesssim \left\| \prod_{t=1}^{n} \left(\mathbf{I} - \frac{\eta_t}{2} \cdot \mathbf{H}\right)(\mathbf{w}_0 - \mathbf{w}^*) \right\|_{\mathbf{H}}^2 + (1 + \mathrm{SNR}) \cdot \sigma^2 \cdot \frac{D_{\mathrm{eff}}}{N_{\mathrm{eff}}}$$

where

$$\mathrm{SNR} := \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}}^2 / \sigma^2$$

$$N_{\mathrm{eff}} := n / \log(n)$$

$$D_{\mathrm{eff}} := \#\left\{\lambda_i \ge \frac{1}{\eta_0 N_{\mathrm{eff}}}\right\} + \eta_0^2 N_{\mathrm{eff}}^2 \cdot \sum_{\lambda_i < \frac{1}{\eta_0 N_{\mathrm{eff}}}} \lambda_i^2$$

Applications

Suppose that $\sigma^2 \lesssim 1,\ \lambda_1 \lesssim 1,\ \|\mathbf{w}_0 - \mathbf{w}_*\|_2 \lesssim 1$.

1. If $\mathrm{tr}(\mathbf{H}) \lesssim 1$, then by choosing $\eta_0 \approx 1/\sqrt{N_{\mathrm{eff}}}$, we have

$$\mathbb{E}R(\mathbf{w}_n) - \mathrm{OPT} \lesssim 1/\sqrt{N_{\mathrm{eff}}}$$

2. If d is finite, then by choosing $\eta_0 \approx 1/\mathrm{tr}(\mathbf{H})$, we have

$$\mathbb{E}R(\mathbf{w}_n) - \mathrm{OPT} \lesssim d/N_{\mathrm{eff}}$$

- *Linear reg. bounds port over to ReLU reg.*
- *Point-wise tight, dim-free, …*
- *Recover existing bound $\tilde{\mathcal{O}}(1/\sqrt{n})$ [KKSK'2011]*
- *ReLU isn't harder!*

Result 2: Misspecified Case

Distributional Assumptions

1. ~~Well-specified noise~~ $\quad \mathbb{E}\big[\big(y - \mathrm{ReLU}(\mathbf{x}^\top \mathbf{w}_*)\big)^2 \mathbf{x}\mathbf{x}^\top\big] \preceq \sigma^2 \cdot \mathbf{H}$
2. Hypercontractivity
3. Symmetric moments

Risk Bound of GLM-tron

Suppose $\eta_0 < 1/(8\alpha \mathrm{tr}(\mathbf{H}))$. Then

$$\mathbb{E}R(\mathbf{w}_n) \lesssim \mathrm{OPT} + \left\| \prod_{t=1}^{n} \left(\mathbf{I} - \frac{\eta_t}{2} \cdot \mathbf{H}\right)(\mathbf{w}_0 - \mathbf{w}^*) \right\|_{\mathbf{H}}^2 + (1 + \mathrm{SNR}) \cdot \sigma^2 \cdot \frac{D_{\mathrm{eff}}}{N_{\mathrm{eff}}}$$

Here, D_{eff} and N_{eff} are as before but

$$\mathrm{SNR} := \big(\mathrm{OPT} + \|\mathbf{w}^*\|_{\mathbf{H}}^2 + \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}}^2\big) / \sigma^2$$

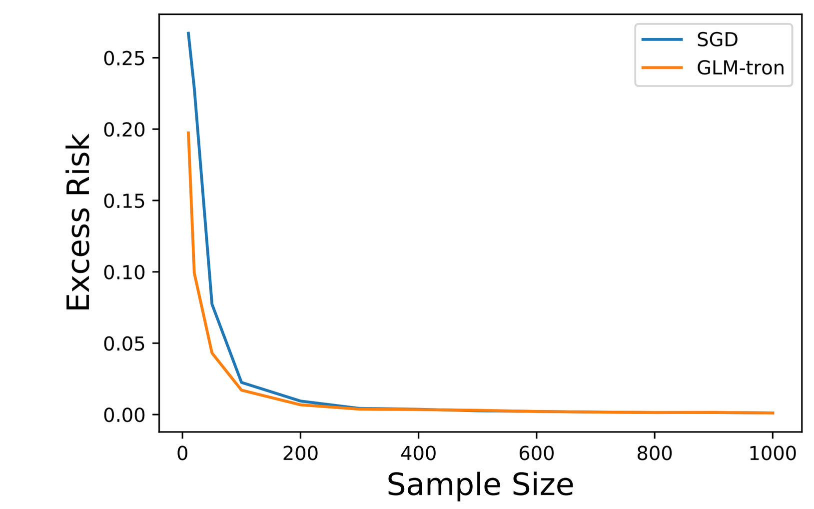
Applications

Suppose that $\sigma^2 \lesssim 1,\ \lambda_1 \lesssim 1,\ \|\mathbf{w}_0 - \mathbf{w}_*\|_2 \lesssim 1,\ \|\mathbf{w}_*\|_2 \lesssim 1$.

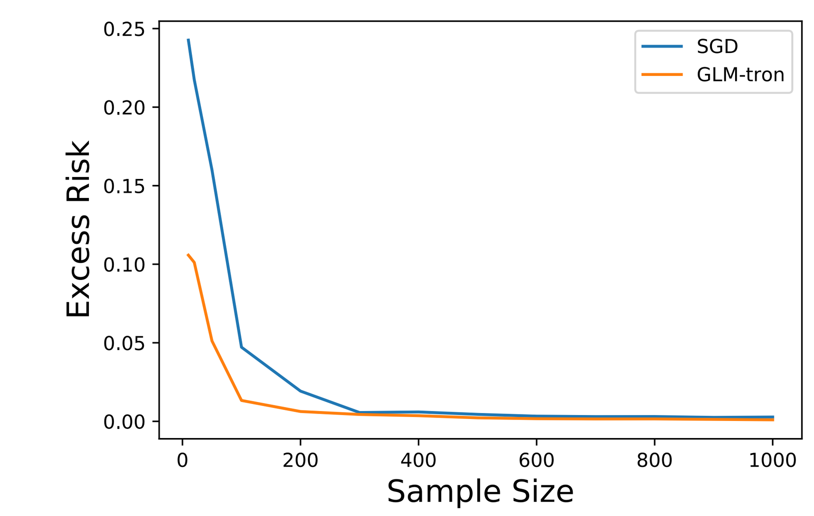
If d is finite, then by choosing $\eta_0 \approx 1/\mathrm{tr}(\mathbf{H})$, we have

$$\mathbb{E}R(\mathbf{w}_n) \lesssim \mathrm{OPT} + d/N_{\mathrm{eff}}$$

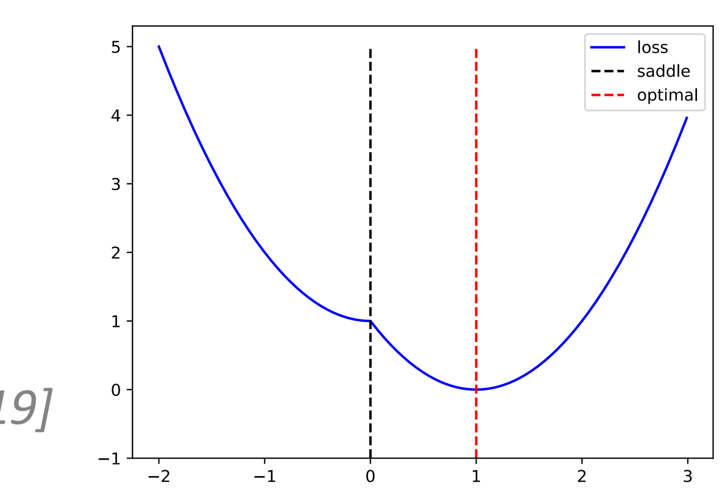
- *$\mathbb{E}R(\mathbf{w}) \le \mathrm{OPT} + o(1)$ bound is believed to be "impossible" [GKK'19]*
- *Previously, best known bound $\mathcal{O}(\mathrm{OPT} + \sqrt{d/n})$ [DGKKS'20]*
- *A dim-free, const. factor approx.*



Well-specified noise + Gaussian data



Well-specified noise + Bernoulli data



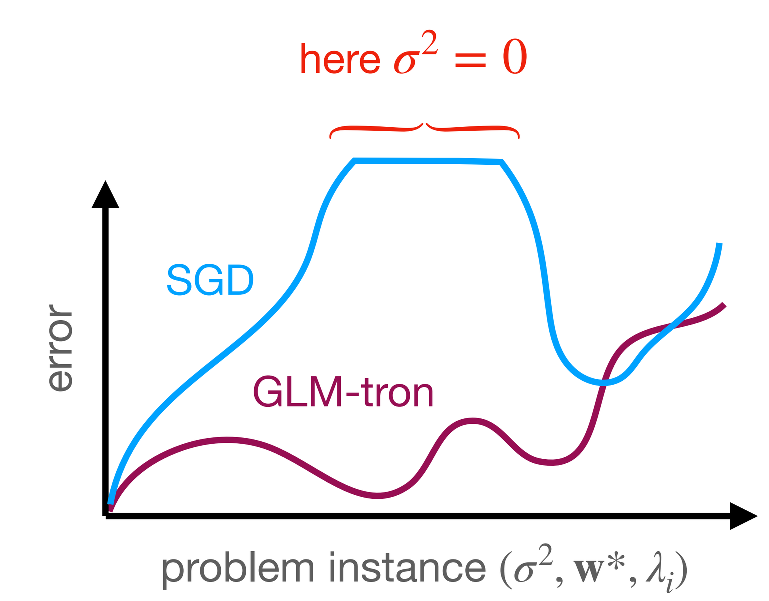
1D non-convex risk landscape
Noiseless + Bernoulli data

Result 3: GLM-tron vs. SGD

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

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

- *SGD no better than GLM-tron when noise is well-specified*
- *SGD suffers constant error if noiseless*



References

- [KKSK'11] Kakade, S., Kanade, V., Shamir, O. and Kalai, A. Efficient learning of generalized linear and single index models with isotonic regression. In NeurIPS 2011.
- [GKK'19] Goel, S., Karmalkar, S. and Klivans, A. Time/accuracy tradeoffs for learning a relu with respect to gaussian marginals. In NeurIPS 2019.
- [DGKKS'20] Diakonikolas, I., Goel, S., Karmalkar, S., Klivans, A. and Soltanolkotabi, M. Approximation schemes for relu regression. In COLT 2020.
- [ZWBGK'21] Zou, D., Wu, J., Braverman, V., Gu, Q. and Kakade, S. Benign overfitting of constant-stepsize sgd for linear regression. In COLT 2021.
- [WZBGK'22] Wu, J., Zou, D., Braverman, V., Gu, Q. and Kakade, S. Last iterate risk bounds of sgd with decaying stepsize for overparameterized linear regression. In ICML 2022.