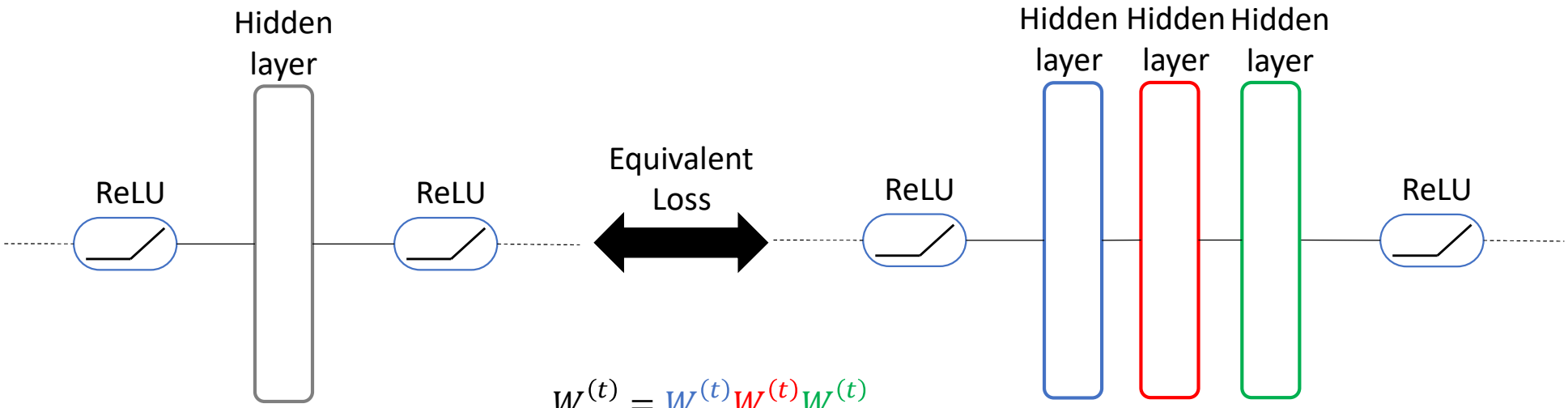
On the Optimization of Deep Networks: Implicit Acceleration by Overparameterization

International Conference on Machine Learning(ICML) 2018

Guillaume LEVY & Clement WANG

Theoretical aspects

Experiment suggests that increasing depth can lead to faster convergence



$$W_e^{(t)} = W_1^{(t)} W_2^{(t)} W_3^{(t)}$$

$$W_e^{(t+1)} \leftarrow (1 - \eta\lambda N) W_e^{(t)} - \eta \| \boldsymbol{W_e^{(t)}} \|_2^{\mathbf{2-\frac{2}{N}}} \cdot \left(\frac{dL^1}{dW}\left(W_e^{(t)}\right) + (N-1) \cdot \boldsymbol{Pr}_{\boldsymbol{W_e^{(t)}}} \left\{ \boldsymbol{\frac{dL^1}{dW}}\left(\boldsymbol{W_e^{(t)}}\right) \right\} \right)$$

Experiments

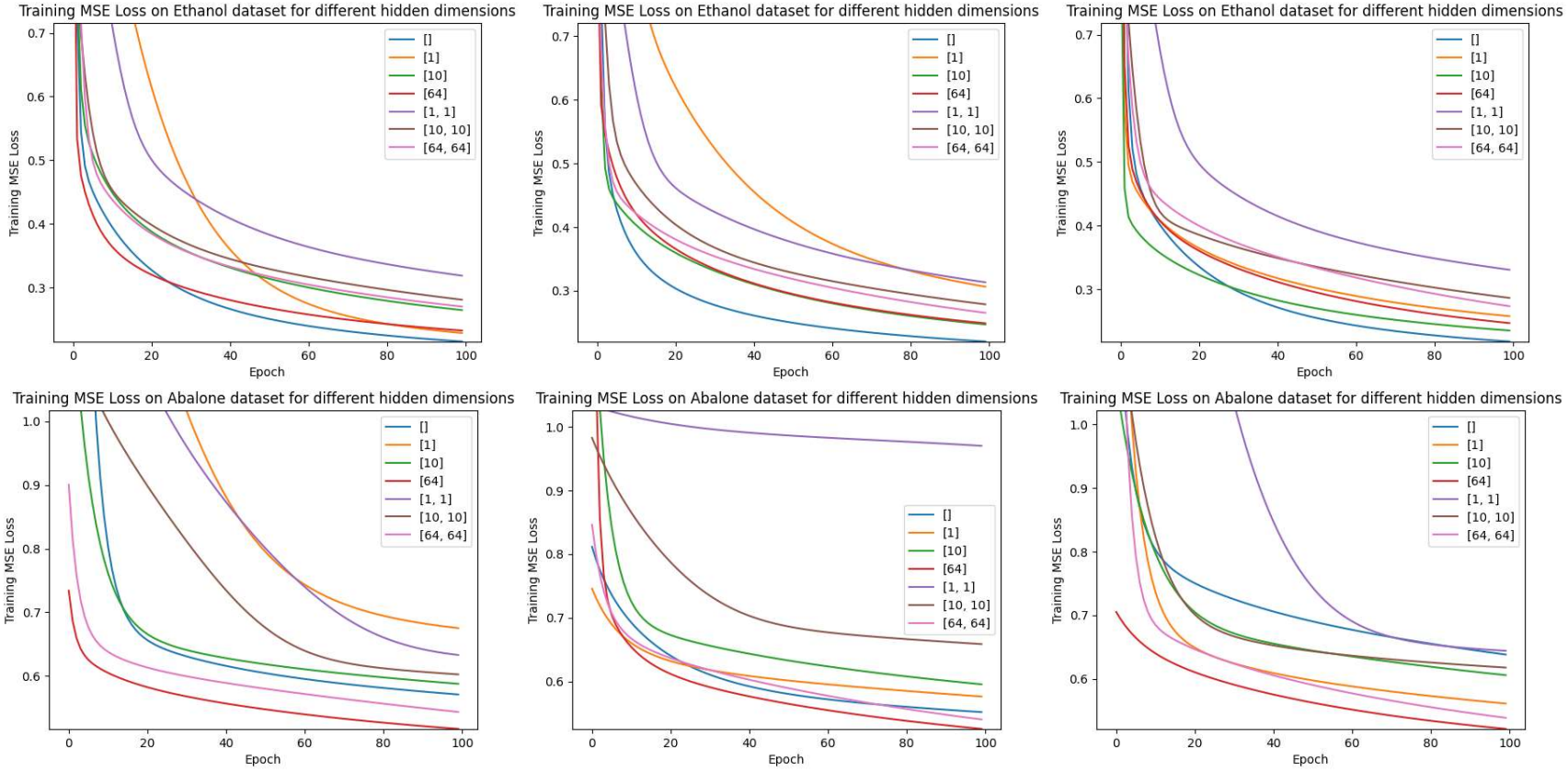
- The paper gives results on one specific dataset
- We tried to reproduce the results on two datasets:
  - Gas Sensor Array Drift Dataset at Different Concentrations [1]:
    - Predict the concentration of the Ethanol
    - 2565 samples, 128 features
  - Abalone length dataset
    - Predict the age of abalones
    - 4177 samples, 8 features
- We were surprised by the schocasticity of the MSE metric with respect to the initialization and to the dataset.
- All experiments were run with a grid search on the learning rate: [0.00001, 0.00005, 0.0001,0.0005, 0.001, 0.005 , 0.01]

[1] Vergara,Alexander. (2013). Gas Sensor Array Drift Dataset at Different Concentrations. UCI Machine Learning Repository. https://doi.org/10.24432/C5MK6M.

[2] Nash,Warwick, Sellers,Tracy, Talbot,Simon, Cawthorn,Andrew, and Ford,Wes. (1995). Abalone. UCI Machine Learning Repository. https://doi.org/10.24432/C55C7W.

Experiments: Linear nets without activation

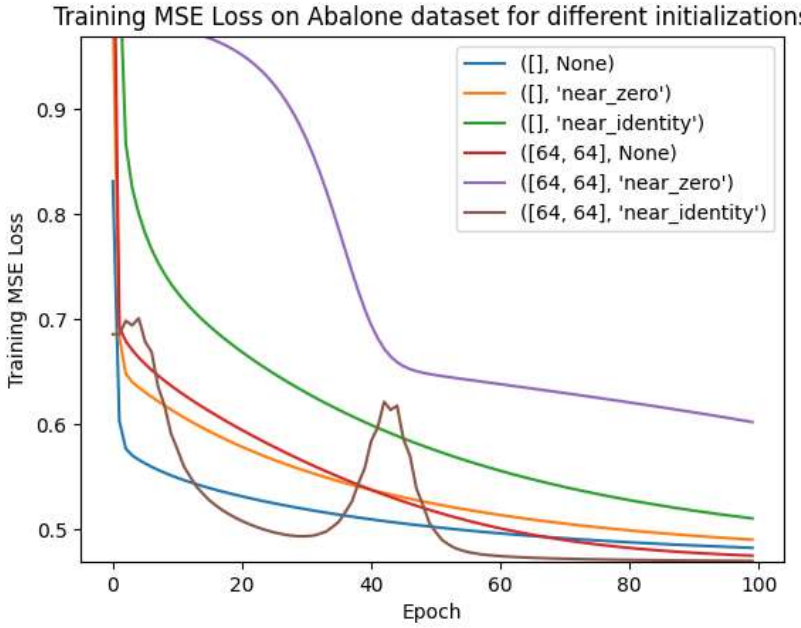
- Same experiment three times with Pytorch default initialization:

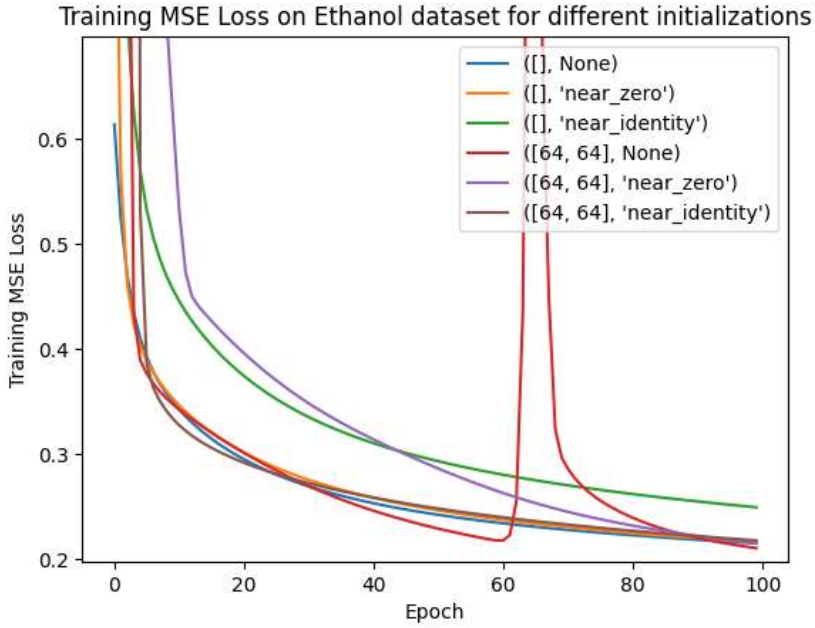


Experiments: Specific initializations

- Default, near zero and near identity initializations

There is much less schosticity when using a particular initialization process.
Sometimes in the over parametrized setting, we get some bursts in the loss



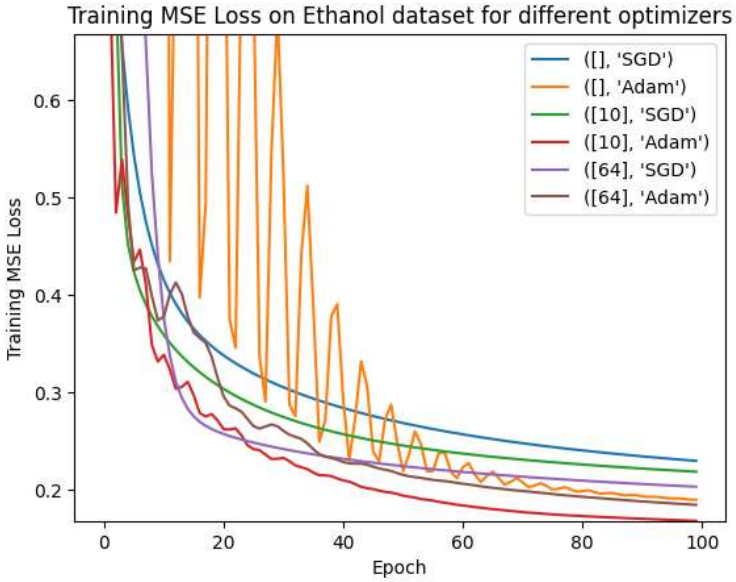


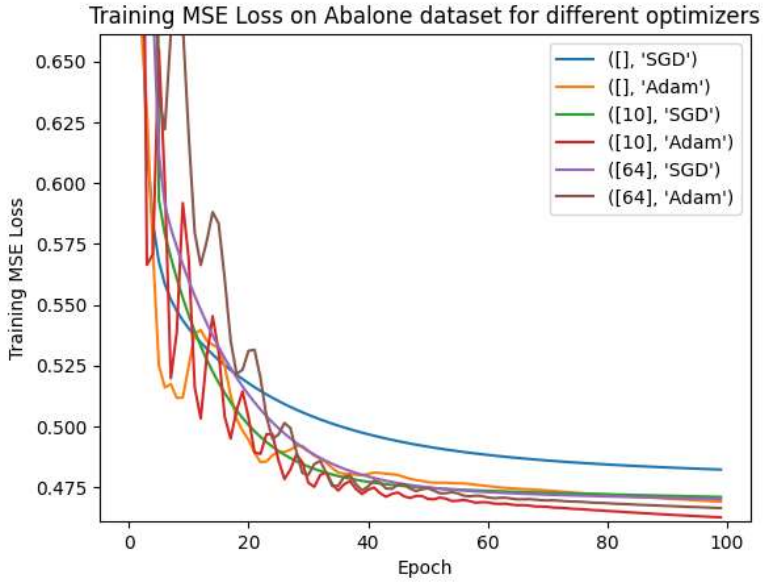
In the overparametrized setting, near identity initialization seems to converge faster

Experiments: Choice of optimizer

- We keep near identity initialization and try Adam VS SGD optimizer

While on the Ethanol dataset, Adam clearly gets better convergence, its benefits on the Abalone dataset are unclear.





The over parametrization seems to make convergence slightly faster in our experiments