"Exponential Convergence Time of Gradient Descent for One-Dimensional Deep Linear Neural Networks" by Ohad Shamir (2018)

Adil Salim

KAUST, Sping 2019

Problem

Supervised Machine Learning with **linear** predictors

$$\begin{array}{rccl} P_{W_1,\dots,W_k}: & \text{features} & \longrightarrow & \text{labels} \\ & x & \longmapsto & \prod_{i=1}^k W_i x \end{array}$$

where W_i matrix.

Example : Deep **Linear** neural networks with depth k. Close to feedforward networks but simpler.

Training :

$$\min_{W_1,\dots,W_k} F(W_1,\dots,W_k) = f\left(\prod_{i=1}^k W_i\right) \tag{1}$$

where f differentiable Lipschitz function. Note that F is not convex.

Example

- Features x_j
- Labels y_j
- $P_{W_1,\dots,W_k}$ should map the features with the labels :

$$\min_{W_1,\dots,W_k} \sum_j \left\| \left(\prod_{i=1}^{k} W_i \right) x_j - y_j \right\|^2$$

- $f(W) = \sum_j \|Wx_j - y_j\|^2$

Result

Problem : Time for training as a function of k *i.e* **Time to solve**

$$\min_{W_1,\ldots,W_k} F(W_1,\ldots,W_k) = f\left(\prod_{i=1}^{k} W_i\right)$$

via Gradient Descent algorithm as a function of k.

Negative result : Training time $= \exp(\Omega(k))$

Gradient Descent algorithm

Assume that $W_1, \ldots, W_k$ are real numbers (no longer matrices).

Algorithm:

- Random initialization $W_1, \ldots, W_k$ close to $1, \ldots, 1$ or zero mean, unit variance.
- $W_j \longleftarrow W_j - \gamma \frac{\partial F}{\partial W_j}(W_1, \ldots, W_k)$

Note that

$$\frac{\partial F}{\partial W_j}(W_1, \ldots, W_k) = \prod_{i \neq j} W_i f' \left(\prod_{i=1}^{k} W_i \right).$$

Statement

Theorem 1

There exists $C, \varepsilon > 0$ *such that if* $\gamma \leq \exp(Ck)$ *then, with probability at least* $1 - \exp(-\Omega(k))$ *over the initialization the following hold :*
the number of iterations n *required such that the* n^{th} *iterate* $W_1, \ldots, W_k$ *of Gradient Descent satisfies* $F(W_1, \ldots, W_k) - \inf F \leq \varepsilon$ *is at least* $\exp(\Omega(k))$ *(i.e* $n \geq \exp(\Omega(k))$*).*

Proof

Consider the random initialization $W_1, \ldots, W_k$. Then $\prod_{i=1}^{k} W_k$ is, with high probability, exponentially small in k. Since

$$\frac{\partial F}{\partial W_j}(W_1, \ldots, W_k) = \prod_{i \neq j} W_i f' \left(\prod_{i=1}^{k} W_i \right),$$

the gradient at the initialization is, with high probability, exponentially small in k as well. One can show that the gradient is also exponentially small at any point from a bounded distance of the initialization. As a result, Gradient Descent only makes exponentially small steps, and hence, exponentially small progress.