The Implicit Bias of Gradient Descent on Separable Data

Soudry, D. et al.(2018 JMLR), cited by 339

InSung Kong

2021 5/13

Seoul National University

Summary

- On linearly **separable** dataset, **logistic regression** with **Gradient Descent**

- **Predictor** converges to the direction of the **max-margin** (hard margin SVM) solution.
  - Normalized vector are convergence in the rate of $O\left(1/\log(t)\right)$
  - It is **slower** than convergence rate of **loss** $(= O(1/t))$

- Can be extended to multi-class problems, and deep network (in a certain restricted setting).

Table of Contents

Table of Contents

Setting

Dataset

- $\{\boldsymbol{x}_n, y_n\}_{n=1}^{N}$, with $\boldsymbol{x}_n \in \mathbb{R}^d$ and binary labels $y_n \in \{-1, 1\}$
- Re-define $y_n\boldsymbol{x}_n$ as $\boldsymbol{x}_n$
- Dataset is linearly separable : $\exists \boldsymbol{w}_*$ s.t $\forall n : \boldsymbol{w}_*^\top \boldsymbol{x}_n > 0$

Model

- We analyze learning by minimizing an empirical loss of the form

$$\mathcal{L}(\boldsymbol{w}) = \sum_{n=1}^{N} \ell\left(\boldsymbol{w}^\top \boldsymbol{x}_n\right)$$

  - $\ell(\cdot)$ is positive, differentiable, monotonically decreasing to zero, β-smooth function, and $-\ell'(\cdot)$ has a tight exponential tail.
- Examples of $\ell(\cdot)$: Exponential loss, Logistic loss

Table of Contents

Main Theorems

Theorem 3

For almost all datasets (i.e., except for a measure zero), any stepsize $0 < \eta < 2\beta^{-1}\sigma_{\max}^{-2}(\mathrm{X})$, any starting point $\boldsymbol{w}(0)$, the gradient descent iterates will be have as :

$$\boldsymbol{w}(t) = \hat{\boldsymbol{w}} \log t + \boldsymbol{\rho}(t)$$

where $\hat{\boldsymbol{w}}$ is the the solution to the hard margin SVM :

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}\in\mathbb{R}^d}{\operatorname{argmin}} \|\boldsymbol{w}\|^2 \text{ s.t. } \boldsymbol{w}^\top \boldsymbol{x}_n \geq 1$$

and $\boldsymbol{\rho}(t)$ is bounded, so

$$\lim_{t\to\infty} \frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|} = \frac{\hat{\boldsymbol{w}}}{\|\hat{\boldsymbol{w}}\|}$$

Theorem 5

With same conditions on previous theorem, predictor converges to the direction of the hard margin SVM solution in terms of

$$\left\| \frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|} - \frac{\hat{\boldsymbol{w}}}{\|\hat{\boldsymbol{w}}\|} \right\| = O\left(\frac{1}{\log t}\right)$$

and in angle

$$1 - \frac{\boldsymbol{w}(t)^\top \hat{\boldsymbol{w}}}{\|\boldsymbol{w}(t)\| \|\hat{\boldsymbol{w}}\|} = O\left(\frac{1}{\log^2 t}\right).$$

Margin converges as

$$\frac{1}{\|\hat{\boldsymbol{w}}\|} - \frac{\min_n \boldsymbol{x}_n^\top \boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|} = O\left(\frac{1}{\log t}\right).$$

On the other hand, the loss itself decrease as

$$\mathcal{L}(\boldsymbol{w}(t)) = O\left(\frac{1}{t}\right)$$

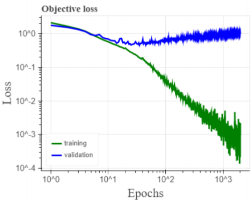
Main Theorems

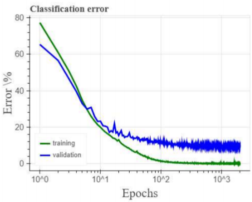
Corollary 6

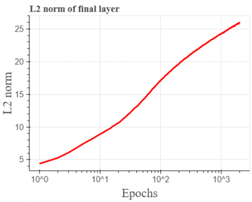
Let ℓ be the logistic loss, and $\mathcal{V}$ be an independent validation set, for which $\exists x \in \mathcal{V}$ such that $\boldsymbol{x}^\top \hat{\boldsymbol{w}} < 0$.
Then the validation loss increases as

$$\mathcal{L}_{\text{val}}\left(\boldsymbol{w}(t)\right) = \sum_{\boldsymbol{x} \in \mathcal{V}} \ell\left(\boldsymbol{w}(t)^\top \boldsymbol{x}\right) = \Omega(\log(t))$$

Example : CNN, CIFAR10







- The traing loss decays as a t^{-1}.
- L_2 norm of last weight layer increases logarithmically.
- After a while, the validation loss starts to increase.
- In contrast, the validation error slowly improves.

Table of Contents

Theorem 7

For almost all multiclass datasets which are linearly separable, any starting point $\boldsymbol{w}(0)$ and any small enough stepsize, the iterates of gradient descent will behave as

$$\boldsymbol{w}(t) = \hat{\boldsymbol{w}} \log t + \rho(t)$$

where $\hat{\boldsymbol{w}}_k$ is the the solution of the K-class SVM :

$$\text{argmin}_{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k} \sum_{k=1}^{K} \|\boldsymbol{w}_k\|^2 \text{ s.t. } \forall n, \forall k \neq y_n : \boldsymbol{w}_{y_n}^\top \boldsymbol{x}_n \geq \boldsymbol{w}_k^\top \boldsymbol{x}_n + 1$$

and $\rho(t)$ is bounded.

Corollary 8

We examine a multilayer neural network with component-wise ReLU functions $f(z) = max(z, 0)$, and weights $\{W_l\}_{l=1}^{L}$. Given input $\boldsymbol{x}_n$ and target $y_n \in \{-1, 1\}$, the DNN produces a scalar output

$$u_n = W_L f\left(W_{L-1} f\left(\cdots W_2 f\left(W_1 \boldsymbol{x}_n\right)\right)\right)$$

If we optimize a single weight layer $\boldsymbol{w}_l = \text{vec}\left(W_l^\top\right)$ using gradient descent, so that $\mathcal{L}(\boldsymbol{w}_l) = \sum_{n=1}^{N} \ell\left(y_n u_n\left(\boldsymbol{w}_l\right)\right)$ converges to zero, and $\exists t_0$ s.t. $\forall t > t_0$ the ReLU inputs do not switch signs, then $\frac{\boldsymbol{w}_l(t)}{||\boldsymbol{w}_l(t)||}$ converges to

$$\hat{\boldsymbol{w}}_l = \underset{\boldsymbol{w}_l}{\text{argmin}} \left\|\boldsymbol{w}_l\right\|^2 \text{ s.t. } y_n u_n\left(\boldsymbol{w}_l\right) \geq 1$$