Qian Qian, Xiaoyuan Qian

October 22, 2019

Our contribution

- We prove that the directions of AdaGrad iterates, with a constant step size sufficiently small, always converge.
- We formulate the asymptotic direction as the solution of a quadratic optimization problem. This achieves a theoretical characterization of the implicit bias of AdaGrad, which also provides insights about why and how the factors involved, such as certain intrinsic properties of the dataset, the initialization and the learning rate, affect the implicit bias.
- We introduce a novel approach to study the bias of AdaGrad. It is mainly based on a geometric estimation on the directions of the updates, which doesn't depend on any calculation on the convergence rate.

Problem Setup

Let $\{(\boldsymbol{x}_n, y_n) : \ n = 1, \cdots, N\}$ be a training dataset with features $\boldsymbol{x}_n \in \mathbb{R}^p$ and labels $y_n \in \{-1, 1\}$.

Consider learning the logistic regression model over the empirical loss:

$$\mathcal{L}(\boldsymbol{w}) = \sum_{n=1}^{N} l\left(\boldsymbol{w}^T \boldsymbol{x}_n\right) , \quad \boldsymbol{w} \in \mathbb{R}^p ,$$

We focus on the following case, same as proposed in **?**:

Assumption 1. There exists a vector $\boldsymbol{w}_*$ such that $\boldsymbol{w}_*^T \boldsymbol{x}_n > 0$ for all n.

Assumption 2. $l(u)$ is continuously differentiable, $\beta-$smooth, and strictly decreasing to zero.

Assumption 3. There exist positive constants a, b, c, and d such that

$$\left| l'(u) + ce^{-au} \right| \leq e^{-(a+b)u} , \quad \text{for } u > d .$$

We are interested in the asymptotic behavior of the AdaGrad iteration scheme. The main problem is: does there exists some vector $\boldsymbol{w}_A$ such that

$$\lim_{t\to\infty} \boldsymbol{w}(t)/\left\|\boldsymbol{w}(t)\right\| = \boldsymbol{w}_A\,?$$

Convergence of the Adaptive Learning Rates

Theorem

The sequence $\{\boldsymbol{h}(t)\}_{t=0}^{\infty}$ *converges as* $t \to \infty$ *to a vector*

$$\boldsymbol{h}_{\infty} = (h_{\infty,1}, \cdots, h_{\infty,p})$$

satisfying $h_{\infty,i} > 0 \quad (i = 1, \cdots, p)$.

Main Results

Theorem

AdaGrad iterates has an asymptotic direction:

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

where

$$\widetilde{\boldsymbol{w}} = \underset{\boldsymbol{w}^T\boldsymbol{x}_n \geq 1, \forall n}{\arg\min} \left\| \frac{1}{\sqrt{\boldsymbol{h}_\infty}} \odot \boldsymbol{w} \right\|^2 . \tag{1}$$