The Implicit Bias of AdaGrad on Separable Data

Qian Qian, Xiaoyuan Qian

November 30, 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

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

It is easy to see that the exponential loss $l(u) = e^{-u}$ and the logistic loss $l(u) = \log\left(1 + e^{-u}\right)$ both meet these assumptions.

Problem Setup

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

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

Difference from the Asymptotic Direction of GD iterates

Example.

Let $\boldsymbol{x}_1 = (\cos\theta, \sin\theta)^T$ and $\mathcal{L}(\boldsymbol{w}) = e^{-\boldsymbol{w}^T \boldsymbol{x}_1}$. Suppose $0 < \theta < \pi/2$. Then $\widehat{\boldsymbol{w}} = \boldsymbol{x}_1$. Selecting $\boldsymbol{w}(0) = (a, b)^T$ and $\epsilon = 0$, then

$$-\boldsymbol{g}(0) = e^{-\boldsymbol{w}(0)^T \boldsymbol{x}_1} \boldsymbol{x}_1 = e^{-a\cos\theta - b\sin\theta} (\cos\theta, \sin\theta)^T,$$

$$\boldsymbol{h}(0) = (h_1(0), h_2(0))^T = e^{a\cos\theta + b\sin\theta} \left(\frac{1}{\cos\theta}, \frac{1}{\sin\theta}\right)^T.$$

In general we can show there is a sequence of positive numbers $p(t)$ such that

$$-\boldsymbol{g}(t) = p(t) (\cos\theta, \sin\theta)^T,$$

and

$$\boldsymbol{h}_\infty = \lim_{t\to\infty} \frac{1}{\sqrt{p(0)^2 + p(1)^2 + \cdots + p(t)^2}} \left(\frac{1}{\cos\theta}, \frac{1}{\sin\theta}\right)^T = \frac{1}{\rho} \left(\frac{1}{\cos\theta}, \frac{1}{\sin\theta}\right)^T.$$

Difference from the Asymptotic Direction of GD iterates

Example (continued). Now

$$\widetilde{\boldsymbol{w}} \quad = \quad \underset{\boldsymbol{w}^T \boldsymbol{x}_1 \geq 1}{\arg\min} \left(w_1^2 \cos\theta + w_2^2 \sin\theta \right) = \left(\frac{1}{\cos\theta + \sin\theta}, \frac{1}{\cos\theta + \sin\theta} \right),$$

and we have $\widetilde{\boldsymbol{w}} / \|\widetilde{\boldsymbol{w}}\| = \left(\sqrt{2}/2, \sqrt{2}/2\right)^T$.

- Note that this direction is invariant when θ ranges between 0 and $\pi/2$, i.e., irrelevant to $\boldsymbol{x}_1$. These two directions coincide only when $\theta = \pi/4$.

Sensitivity to Small Coordinate System Rotations

- If we consider the same setting as in the previous example, but taking $\theta \in (\pi/2, \pi)$. then the asymptotic direction $\widetilde{\boldsymbol{w}}/\|\widetilde{\boldsymbol{w}}\|$ will become $\left(-\sqrt{2}/2, \sqrt{2}/2\right)^T$.
- This implies, if $\boldsymbol{x}_1$ is close to the direction of $y-$axis, then a small rotation of the coordinate system may result in a large change of the asymptotic direction reaching a right angle, i.e., in this case the asymptotic direction is highly unstable even for a small perturbation of its $x-$coordinate.

Cases that the Asymptotic Direction is Stable

Proposition. Let $\boldsymbol{a} = (a_1, \cdots, a_p)^T$ be a vector satisfying $\boldsymbol{a}^T \boldsymbol{x}_n \geq 1 \ \ (n = 1, \cdots, N)$ and $a_1 \cdots a_p \neq 0$. Suppose that $\boldsymbol{w} = (w_1, \cdots, w_p)^T$ satisfies $\boldsymbol{w}^T \boldsymbol{x}_n \geq 1 \ \ (n = 1, \cdots, N)$ and

$$a_i (w_i - a_i) \geq 0 \ \ (i = 1, \cdots, p).$$

Then for any $\boldsymbol{b} = (b_1, \cdots, b_p)^T$ such that $b_1 \cdots b_p \neq 0$,

$$\arg\min_{\boldsymbol{w}^T \boldsymbol{x}_n \geq 1, \, \forall n} \|\boldsymbol{b} \odot \boldsymbol{w}\|^2 = \arg\min_{\boldsymbol{w}^T \boldsymbol{x}_n \geq 1, \, \forall n} \|\boldsymbol{w}\|^2 = \boldsymbol{a},$$

and therefore the asymptotic directions of AdaGrad and GD are equal.

Cases that the Asymptotic Direction is Stable

Proposition. Suppose $N \geq p$ and $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N] \in \mathbb{R}^{p \times N}$ is sampled from any distribution whose density function is nonzero almost everywhere. Then with a positive probability the asymptotic directions of AdaGrad and GD are equal.