A TERMINATION CRITERION FOR STOCHASTIC GRADIENT DESCENT FOR BINARY CLASSIFICATION

Sina Baghal, Courtney Paquette, Stephen Vavasis†

†University of Waterloo

Abstract

In this work [1], we consider the binary classification problem where stochastic gradient descent is used to minimize logistic or hinge loss. We propose a simple and computationally inexpensive termination criterion which exhibits a good degree of predictability on yet unseen data. We provide theoretical and numerical evidence for the effectiveness of our test.

Motivation

Early stopping rules and termination criteria play a central role in machine learning. Models trained via first order methods without early stopping may not predict well on future data since they over-fit the given samples. In the versatile case of deep neural networks, although it is known that even interpolating the training data will result in low generalization errors, early stopping can still be used to correct the adversarially corrupted sample points [2]. Despite all these advantages, our fundamental understanding of explicit stopping rules has lagged far behind.

Model

We suppose that data comes from a mixture model with two means $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$. Without loss of generality, we can assume that $\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1 = 0$.

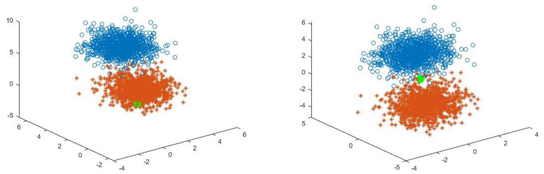


Fig. 1: Re-centring: Estimate $\boldsymbol{v} := \frac{\boldsymbol{\mu}_0+\boldsymbol{\mu}_1}{2}$ using a preliminary sampling phase and then translate the data by $-\boldsymbol{v}$.

We aim to minimize the following expected loss function:

$$\min f(\boldsymbol{\theta}) := \mathbb{E}_{(\boldsymbol{\zeta},y)\sim\mathcal{P}}\ell\left(\boldsymbol{\zeta}^T\boldsymbol{\theta}, y\right),$$

where ℓ is either the logistic or hinge loss. Logistic loss function is as follows.

$$\ell(x,y) = \begin{cases} \log\left(1+\exp(x)\right) & y = 0 \\ \log\left(1+\exp(-x)\right) & y = 1 \end{cases}$$

And hinge loss function is as follows.

$$\ell(x,y) = \begin{cases} \max\{1+x, 0\} & y = 0 \\ \max\{1-x, 0\} & y = 1 \end{cases}$$

For brevity, we present our results only for the logistic loss. Results for hinge loss are similar. Via a folding argument, it suffices to minimize the following loss function.

$$\min \hat{f}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\xi}\sim\hat{\mathcal{P}}}\hat{\ell}\left(\boldsymbol{\xi}^T\boldsymbol{\theta}\right) \tag{1}$$

Assumption 1. *In our analysis, we assume that the data is isotropically Gaussian distributed i.e.*

$$\hat{\mathcal{P}} \equiv N(\boldsymbol{\mu}, \sigma^2 I_d).$$

Results

Lemma 1. *Under the Gaussian data assumption, any optimal classifier will be aligned with $\boldsymbol{\theta}^*$ where $\boldsymbol{\theta}^*$ denotes the unique solution to* (1)*. In other words,*

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathbb{P}_{\boldsymbol{\xi}\sim\hat{\mathcal{P}}}\left(\boldsymbol{\xi}^T\boldsymbol{\theta} > 0\right) = \mathbb{R}_{++}\cdot\boldsymbol{\theta}^*.$$

Furthermore, the following is true.

$$\boldsymbol{\theta}^* = \frac{\boldsymbol{\mu}}{2\sigma^2}.$$

Any classifier aligned with $\boldsymbol{\mu}$ is therefore an optimal classifier. We present our termination criterion below.

Algorithm 1: Stochastic gradient descent with termination test

1: **initialize:** $\boldsymbol{\theta}_0 = \mathbf{0}$, $\alpha > 0$, $k = 0$
2: **while True**
3: Pick data point $\boldsymbol{\xi}_{k+1} \sim \hat{\mathcal{P}}$
4: If $\boldsymbol{\xi}_{k+1}^T\boldsymbol{\theta}_k \geq 1$
break
5: Update $\boldsymbol{\theta}$ by setting

$$\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k + \frac{\alpha\boldsymbol{\xi}_{k+1}}{1+\exp\left(\boldsymbol{\xi}_{k+1}^T\boldsymbol{\theta}_k\right)}$$

6: $k \leftarrow k+1$
7: **end**

We therefore formally define our stopping time as follows.

$$T_0 := \inf\left\{k : \boldsymbol{\xi}_{k+1}^T\boldsymbol{\theta}_k \geq 1\right\}. \tag{2}$$

In order to analyse the stopping time T_0, we borrow techniques from probability theory. To do so, however, we crucially need to assume that two σ-algebras $\mathcal{F}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots)$ and $\mathcal{F}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots)$ are independent. This assumption obviously fails to hold and as such we carry our analysis for the following substitute stopping time.

$$T := \inf\left\{k : \hat{\boldsymbol{\xi}}_{k+1}^T\boldsymbol{\theta}_k \geq 1\right\},$$

where $\hat{\boldsymbol{\xi}}_1, \hat{\boldsymbol{\xi}}_2, \cdots$ are sampled i.i.d. from $\hat{\mathcal{P}}$. Notably, our numerical results did not show any significant difference between the two tests T_0 and T. We present our main results next. The first theorem provides a bound for the expected value of the stopping time T.

Theorem 1. *(Bounds for expected value of T)*

- *Suppose that $\sigma \leq 0.33\|\boldsymbol{\mu}\|$ and $\alpha > 0$ is arbitrary. The following is then true.*

$$\mathbb{E}[T] \leq 2 + \frac{2(c_1 + c_2\alpha\|\boldsymbol{\mu}\|^2)^2}{\alpha\|\boldsymbol{\mu}\|^2}\cdot\left(\Phi^c\left(\frac{\|\boldsymbol{\mu}\|}{\sigma}\right) + \frac{\alpha\sigma^3}{\|\boldsymbol{\mu}\|}\cdot\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\|\boldsymbol{\mu}\|^2}{2\sigma^2}\right) + 1\right).$$

- *Suppose that $\sigma > 0.33\|\boldsymbol{\mu}\|$ and the step-size α satisfies*

$$\alpha \leq c_3\cdot\frac{\|\boldsymbol{\mu}\|^2}{\sigma^2(\|\boldsymbol{\mu}\|^2 + d\sigma^2)}.$$

It then holds that $\mathbb{E}[T] < +\infty$.

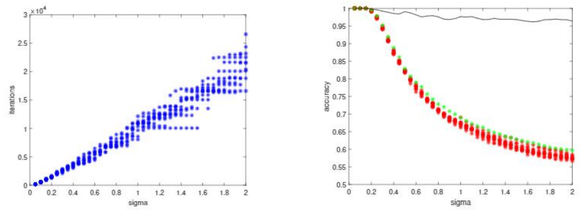
The next theorem provides a bound on the expected value of the misalignment with respect to the direction of the optimal classifier.

Theorem 2. *(Angle bound) Let $\boldsymbol{v} \in \mathbb{S}^{d-1}$ such that $\boldsymbol{v} \perp \boldsymbol{\theta}^*$. The following bound holds.*

$$\mathbb{E}\left[|\boldsymbol{v}^T\boldsymbol{\theta}_T|\right] \leq \sigma\alpha\sqrt{\frac{2}{\pi}}\cdot\mathbb{E}[T].$$

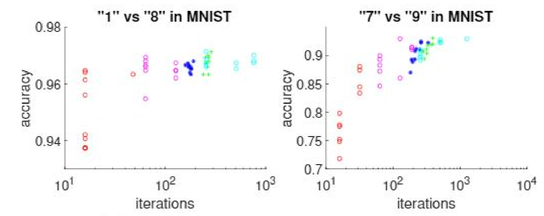
Numerical Experiments

We investigate the performance of our termination test on two popular data sets, MNIST and CIFAR-10, as well as synthetic data generated from Gaussian distributions. All tests were performed using our zero overhead stopping criteria outlined in (2); experiments using our theoretical test which required an extra sample are not presented since the behaviors of the two criteria were indistinguishable on all data sets.

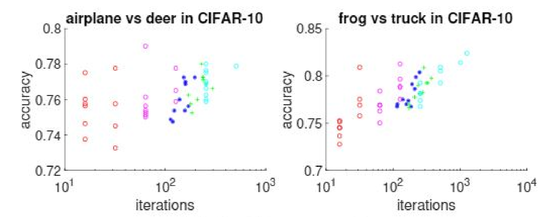


(a) (b)

Fig. 2: Performance of stopping criterion (2) on a mixture of Gaussians as σ is varied. The black curve on the right is the ratio of the average accuracy (over 10 trials) of the classifier when (2) holds to the accuracy of the optimal classifier.



(a) MNIST handwritten digits results



(b) CIFAR-10 results

Fig. 3: Results for MNIST and CIFAR-10 image classification data. Each plot shows 10 random runs of SGD applied to the indicated data set, and for each of the ten runs, five termination tests corresponding to five colors were applied. SVS (small validation set) was tried with $p = 8, 32, 128$, depicted as red, magenta and cyan circles respectively. Test (2) is indicated with a blue asterisk. A green '+' corresponds to termination after $1.5k$ iterations, where k is the iteration index that (2) first holds.

References

[1] Sina Baghal, Courtney Paquette, and Stephen A. Vavasis. *A termination criterion for stochastic gradient descent for binary classification.* 2020. arXiv: 2003.10312 [math.OC].

[2] Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. "Gradient Descent with Early Stopping is Provably Robust to Label Noise for Overparameterized Neural Networks". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics.* Vol. 108. PMLR, 2020.