

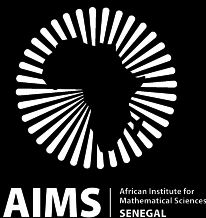
REPORT RESTITUTION



AIMS | African Institute for
Mathematical Sciences
SENEGAL

$$\tilde{C}_{SGD}(\omega) = C(\omega) + \frac{\epsilon}{4m} \sum_{k=0}^{m-1} \|\nabla \hat{C}_k(\omega)\|^2.$$

Mikhaël KIBINDA



AIMS | African Institute for
Mathematical Sciences
SENEGAL



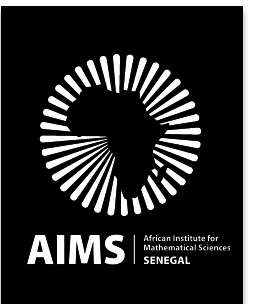
AIMS | African Institute for
Mathematical Sciences
SENEGAL

TITLE:

**ON THE ORIGIN OF IMPLICIT REGULARIZATION IN
STOCHASTIC GRADIENT DESCENT**

1. State of the art
 - Regularization, algorithms of optimisation, learning rate, etc.
2. Introduction
 - How deep learning works ?
3. Core idea
4. Implementation
 - Set Up
5. Experimental Results
6. Conclusion

Mikhaël KIBINDA



State of the art

When we train a machine learning model.

- To find the model that fits or generalizes well on the unseen data.

When the model gives high error on the test and small error on the train, we observe a phenomenon called Overfitting.

- Regularization applies a “penalty” to the input parameters with the larger coefficients, which subsequently limits the amount of variance in the model.

State of the art

Several algorithms are used to find a local solution of a given optimization problem.

- GD, MBGD, SGD, etc.
- Regarding to the pros and cons of each algorithm, SGD seems to be the good one.
- It's computationally fast as only one sample is processed at a time.
- The learning rate is a hyperparameter that controls how much to change the model in response to the estimated error each time the model weights are updated.
- This is one of the challenge (choosing the learning rate) fixed in this work, by proposing a new approach called: "implicit regularization".

INTRODUCTION

How deep learning works ?

- It's not sufficient to focus our mind on the loss function or the model class only.

For the specific training data, there are several minima, some of them generalize well (i.e result in low test error) others can be arbitrary badly overfit.

- Our interest is not to find which algorithm converges quickly to a local minimum but is to find in which of the available minima it prefers to reach first.

INTRODUCTION

- Optimisation algorithms have certain preference in their convergence to a minimum among the possible available minima and this preference is often described as an "implicit regularization".

PURPOSE OF THE WORK OR CORE IDEA

- The learning rate plays in some case an important role.
- Managing it can significantly achieve the performance both in test and train accuracies.
- Large learning rate can give the high test accuracy and this effect can minimize the training accuracy.
- It's often difficult to generalize this phenomenon.

PURPOSE OF THE WORK OR CORE IDEA

Purpose:

- **Modify the loss function in order to see how finite learning rate and small batch size can aid generalization.**

The use of finite learning rates and small batch sizes introduces **implicit regularization**, which can enhance test accuracy of deep networks.

PURPOSE OF THE WORK OR CORE IDEA

- The modified loss function can be written as:

$$C_{modSGD}(w) = C_{Org}(w) + \frac{\epsilon}{4m} \sum_{j=1}^m \|\nabla C_j(w)\|^2$$

The scale of this implicit regularization term is proportional to the learning rate ϵ .

APPROACH USED IN TERM OF IMPLEMENTATION

Set Up:

- The Fashion-MNIST Dataset which comprises 1,000 classes.
- 60,000 training examples and 10,000 examples in the test set.
- A simple fully connected MLP which comprises 3 nonlinear layers and each with width 4096 ReLU activations, a final linear softmax layer and then flatten the input to a 784 dimensional vector.
- epoch = 10; batch_size = 16, 32, 64; learning rate = $2e-2$, $1e-2$, $1e-3$, etc.

EXPERIMENTAL RESULTS

For the batch_size = 64

	Train_acc	Test_acc	Loss
lr = 1e-2	0.9907	0.9468	0.9950
lr = 1e-3	0.9975	0.9491	1.0165
lr = 2e-2	0.112	0.1135	11.2999

For the batch_size = 32 (TO DO)

	Train_acc	Test_acc	Loss
lr = 1e-2
lr = 1e-3
lr = 2e-2

EXPERIMENTAL RESULTS

For the `batch_size = 16` (TO DO)

In progress...

After applying implicit regularization, we will find that the smaller we make the batch size and learning rate, the better and more significant we expect the performance to be on the test set. Doing the opposite would lead to poorer performance.

CONCLUSION

In fact, this paper is a very interesting analysis of stochastic gradient descent. The authors proposed a very interesting new technique for analysing optimization algorithms with finite stepsize.

Implicit regularization is a very interesting approach, it gives us an insight into how SGD works and what kind of minima it prefers to tend to first.



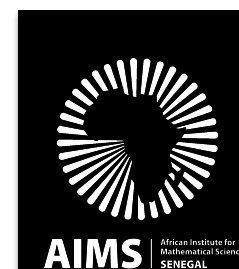
AIMS

African Institute for
Mathematical Sciences
SENEGAL

Thank you
for
listening!



Handwritten signature in red ink.



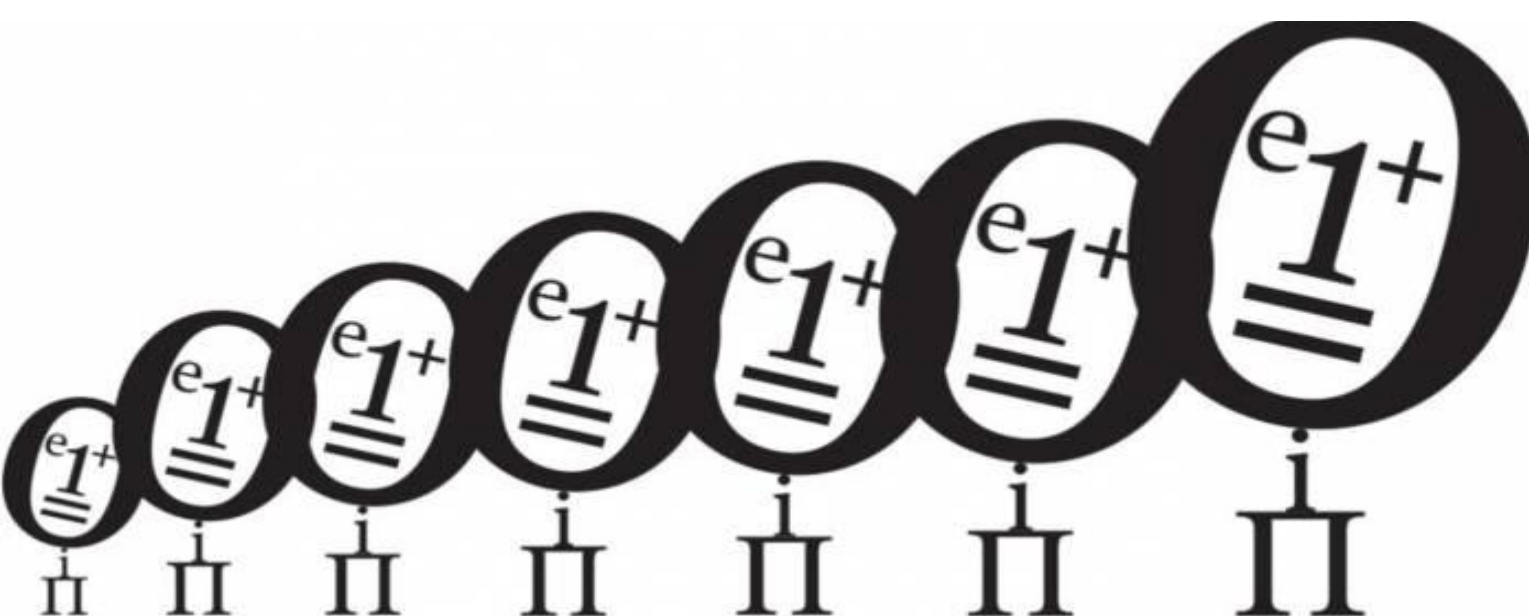
AIMS
African Institute for
Mathematical Sciences
SENEGAL



AIMS

African Institute for
Mathematical Sciences
SENEGAL

Questions

$$H(t) |\psi(t)\rangle = i\hbar \frac{d}{dt} |\psi(t)\rangle$$




AIMS African Institute for
Mathematical Sciences
SENEGAL