# REPORT RESTITUTION

Mikhaël KIBINDA

AIMS | African Institute for Mathematical Sciences SENEGAL

Mikhaël KIBINDA

# TITLE:

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

First part:

- Literature (State of the art)
- Implementation

Second part:

- Mathematical aspect
- Implementation (complete)

## Mathematical aspect

- **Where does this modified loss come from ?**

$$\tilde{C}_{SGD}(w) = C(w) + \frac{\epsilon}{4m}\sum_{j=1}^{m} \|\nabla \hat{C}_j(w)\|^2$$

- **Backward Error Analysis**

## Mathematical aspect

- The gradient flow follows the ODE: $\dot{\omega} = -\nabla C(\omega)$
- This system cannot be solved analytically, we can only approximate the solution by using discrete update like Euler step:

$$\omega(t+\epsilon) \approx \omega(t) + \epsilon f(\omega(t))$$

- The modified system will be $\dot{\omega} = \tilde{f}(\omega)$ where $\tilde{f}(\omega) = f(\omega) + \epsilon f_1(\omega) + \epsilon^2 f_2(\omega) + \ldots$
- The standard derivation of BEA begins by taking a Taylor expansion in $\varepsilon$ of the solution to the modified flow.

$$w_{t+n} = w_t + \alpha\tilde{f}(w_t) + \alpha\tilde{f}(w_{t+1}) + \alpha\tilde{f}(w_{t+2}) + \ldots$$

$$w_{t+n} = w_t + \alpha\tilde{f}(w_t) + \alpha\tilde{f}(w_t + \alpha\tilde{f}(w_t)) + \alpha\tilde{f}(w_t + \alpha\tilde{f}(w_t) + \alpha\tilde{f}(w_t + \alpha\tilde{f}(w_t))) + \ldots$$

$$w_{t+n} = w_t + n\alpha\tilde{f}(w_t) + \frac{n(n-1)}{2}\alpha^2\nabla\tilde{f}(w_t)\tilde{f}(w_t) + O(n^3\alpha^3) \tag{1}$$

When the number of step $n \to \infty$ $\alpha = \epsilon/n$.

## Mathematical aspect

$$w_{t+\epsilon} = w_t + \epsilon \tilde{f}(w_t) + (\epsilon^2/2)\nabla \tilde{f}(w_t)\tilde{f}(w_t) + O(\epsilon^3)$$

$$w_{t+\epsilon} = w_t + \epsilon f(w_t) + \epsilon^2(f_1(w_t) + (1/2)\nabla f(w_t) f(w_t)) + O(\epsilon^3) \qquad (2)$$

- We now derive the influence of **m** SGD updates. Following the similar approach in **(1)**:

$$w_m = w_0 - \epsilon\nabla\hat{C}_0(w_0) - \epsilon\nabla\hat{C}_1(w_1) - \epsilon\nabla\hat{C}_2(w_2) - \ldots$$

$$w_m = w_0 - \epsilon\sum_{j=0}^{m-1}\nabla\hat{C}_j(w_0) + \epsilon^2\sum_{j=0}^{m-1}\sum_{k<j}\nabla\nabla\hat{C}_j(w_0)\nabla\hat{C}_k(w_0) + O(m^3\epsilon^3)$$

$$w_m = w_0 - m\epsilon\nabla C(w_0) + \epsilon^2\xi(w_0) + O(m^3\epsilon^3)$$

The Euler step for SGD is: $w_{i+1} = w_i - \epsilon\nabla\hat{C}_{i\%m}(w_i)$

## Mathematical aspect

- The second order correction $\xi(w) = \sum_{j=0}^{m-1} \sum_{k<j} \nabla\nabla\hat{C}_j(w)\nabla\hat{C}_k(w)$ is a random variable which depends on the order of the mini-batches.

$$\begin{aligned}
\mathbb{E}(\xi(w)) &= \frac{1}{2}\left(\sum_{j=0}^{m-1} \sum_{k\neq j} \nabla\nabla\hat{C}_j(w)\nabla\hat{C}_k(w)\right) \\
&= \frac{1}{2}\nabla \sum_{j=0}^{m-1} \nabla\hat{C}_j(w) \sum_{k=0}^{m-1} \nabla\hat{C}_k(w) - \frac{1}{2}\nabla \sum_{j=0}^{m-1} \nabla\hat{C}_j(w)\nabla\hat{C}_j(w) \\
&= \frac{m^2}{4}\nabla\left(\|\nabla C(w)\|^2 - \frac{1}{m^2}\sum_{j=0}^{m-1} \|\nabla\hat{C}_j(w)\|^2\right)
\end{aligned}$$

- Let’s compute the expected value of the SGD iterate after one epoch

$$\mathbb{E}(w_m) = w_0 - m\epsilon\nabla C(w_0) + \frac{m^2\epsilon^2}{4}\nabla\left(\|\nabla C(w_0)\|^2 - \frac{1}{m^2}\sum_{j=0}^{m-1}\|\nabla\hat{C}_j(w_0)\|^2\right) + O(m^3\epsilon^3) \qquad (3)$$

## Mathematical aspect

- Use (2) to obtain:

$$w(m\epsilon) = w_0 - m\epsilon\nabla C(w_0) + m^2\epsilon^2\left(f_1(w_0) + \frac{1}{2}\nabla\nabla C(w_0)\nabla C(w_0)\right) + O(m^3\epsilon^3)$$

$$= w_0 - m\epsilon\nabla C(w_0) + m^2\epsilon^2\left(f_1(w_0) + \frac{1}{4}\nabla\|\nabla C(w_0)\|^2\right) + O(m^3\epsilon^3) \qquad (4)$$

- $\mathbb{E}(w_m) = w(m\epsilon) + O(m^3\epsilon^3)$ and let set $\dot{w} = -\nabla C(w) + m\epsilon f_1(w)$ where $f_1(w) = -\frac{1}{4m^2}\nabla\sum_{j=0}^{m-1}\|\nabla\hat{C}_j(w_0)\|^2$

- We conclude:

$$\dot{w} = -\nabla\tilde{C}_{SGD}(w)$$

$$-\nabla\tilde{C}_{SGD}(w) = -\nabla C(w) - \frac{\epsilon}{4m}\nabla\sum_{j=0}^{m-1}\|\nabla\hat{C}_j(w_0)\|^2$$

$$\tilde{C}_{SGD}(w) = C(w) + \frac{\epsilon}{4m}\sum_{j=0}^{m-1}\|\nabla\hat{C}_j(w_0)\|^2 \qquad \blacksquare$$

# IMPLEMENTATION/EXPERIMENTAL RESULTS

- For the batch_size = 64

| | Train_acc | Test_acc |
|---|---|---|
| lr = 1e-2 | 0.9912 | 0.9256 |
| lr = 1e-3 | 0.9981 | 0.9312 |
| lr = 2e-2 | 0.112 | 0.1135 |

- For the batch_size = 32

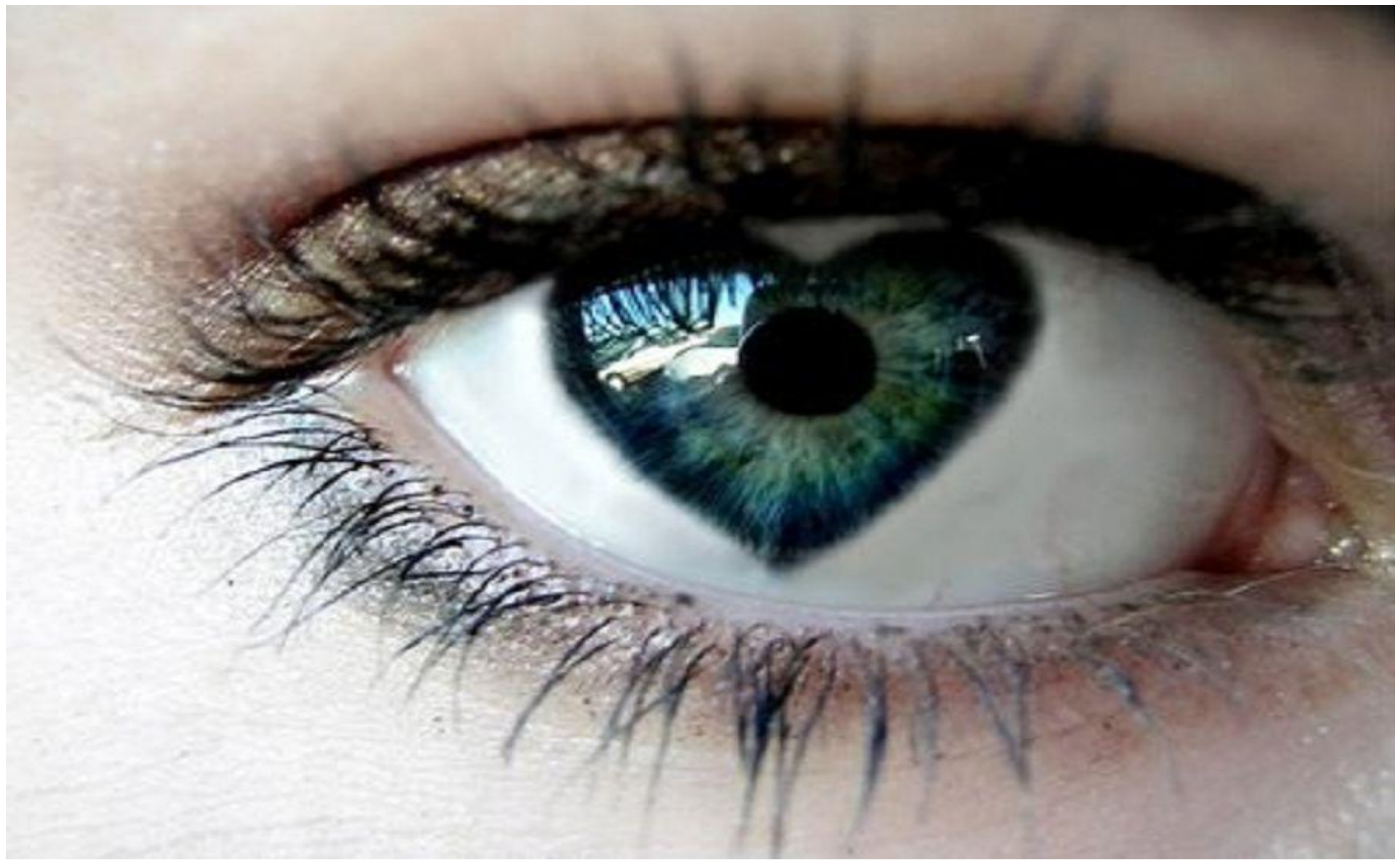
| | Train_acc | Test_acc |
|---|---|---|
| lr = 1e-2 | 0.993 | 0.928 |
| lr = 1e-3 | 0.964 | 0.9655 |
| lr = 2e-2 | 0.9893 | 0.984 |

## IMPLEMENTATION/EXPERIMENTAL RESULTS

- For the batch_size = 16

| | Train_acc | Test_acc |
|---|---|---|
| lr = 1e-2 | 0.9629 | 0.9637 |
| lr = 1e-3 | 0.9932 | 0.9892 |
| lr = 2e-2 | 0.9898 | 0.9461 |

After applying implicit regularization, we see the smaller we make the batch size and learning rate, the better and more significant we expect the test accuracy.



Attention is all we need, thank you for your attention !

# Questions