Adagrad Optimization – Sparse

Adaptive Gradient Algorithm

$w_n = w_o - \eta \frac{\partial L}{\partial w_o}$

→ learning rate η

→ Sparse

$\frac{1}{0} = \infty$

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{\sum_{i=1}^{n} g_i^2 + \epsilon}} * \frac{\partial L}{\partial w_t}$$

gradient current iteration

very very small No. = 10^{-8}

1	2	3	4
5	6	0	7
8	9	3	4
0	1	3	5

⇒ more no. of non-zero = Dense

Text Data ← OHE

BOW / TF-IDF / N-GRAMS / Word2Vec / Glove

0	0	0	1	0
0	0	1	0	0
1	0	0	0	1
0	0	0	2	0
0	0	3	0	0

⇒ More No. of zeros = Sparse

You are screen sharing

Stop share

Ravinder raised hand

very very small

BackProp

$$\boxed{w_t = w_{t-1} - \eta' \frac{\partial L}{\partial w_0}}$$

2

0.001

$$\eta' = \frac{\eta}{\sqrt{\sum_{i=1}^{n} \alpha_i + \epsilon}}$$

$\frac{1}{10}$ $\frac{1}{100}$ $\frac{1}{1000}$ $\frac{1}{10000}$ $\frac{1}{\infty}$ 22

smaller

$\alpha = 1, 1.5, 0.5, 0.6, 2, 2.5 \ldots$

658

$\sum_{i=1}^{n} (1 + 1.5 + 0.5 + 0.6 + 2 + 2.5 \ldots)$

$0.000000001 * 2$

$\eta' = 0.00000000002$

$= 0.000$

Learnvista Priva...

Learnvista Private Limited

Webinar Chat

Read twice but difficult to digest..Anyway read it again

PRAMOD K. to Hosts and panelists

PK yes

abhishek to Hosts and panelists

A what is alpha?

yes sir

no

1/10

Urmil Shah to Hosts and panelists

US The idea of very small eta value is to address exploding gradient problem but if it gets very small it never reach local minima? is my understanding correct.

PRAMOD K. to Hosts and panelists

PK Bigger denominator means very small value will receive

Who can see your messages? Recording on

To: Hosts and panelists

Type message here...

2 / 2

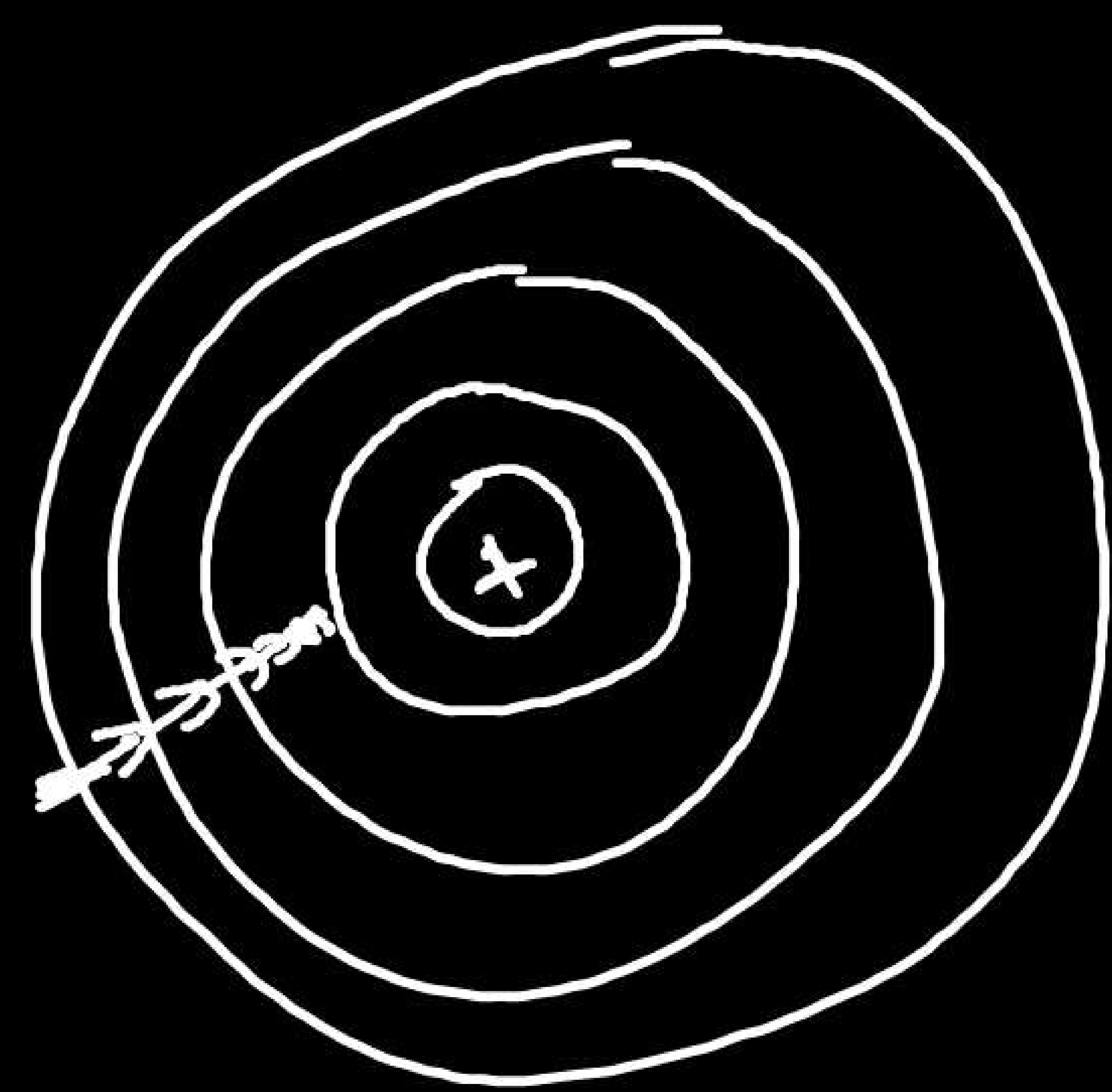
You are screen sharing

Stop share

Ravinder raised hand

Learnvista Priva...

Learnvista Private Limited



$$W_n = W_0 - \eta' \frac{\partial L}{\partial w_0}$$

$$\boxed{\eta' = \frac{\eta}{\sqrt{\sum_{i=1}^{2} \alpha_i + \epsilon}}}$$

$w = 50$

$w_1 = 48$

$w_2 = 53$

$W_t = - - -$

0.000001

50.000001

49.999999

Webinar Chat

abhishek to Hosts and panelists

what is alpha?

yes sir

no

1/10

Urmil Shah to Hosts and panelists

The idea of very small eta value is to address exploding gradient problem but if it gets very small it never reach local minima? is my understanding correct.

PRAMOD K. to Hosts and panelists

Bigger denominator means very small value will receive

abhishek to Hosts and panelists

but how does it help to solve sparse data?

Ravinder to Hosts and panelists

How is the formula of Alpha ?

Who can see your messages? Recording on

To: Hosts and panelists

Type message here...

3 / 3

RMS Prop → Root mean Squared Propagation

→ extension of gradient descent and the Adagrad version of GD that uses a decaying average of partial gradient in the adaptation of the step size for each parameter.

sparse + Dense

RMSprop

$$v_t = \beta v_{t-1} + (1-\beta)(\nabla w_t)^2$$

$\beta = 0.95$

exponential decaying Avg

You are screen sharing | Stop share

Ravinder

Ravinder

$$w_1 = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$$

$$\sum w_i = 55$$

$$(w_i)_{Avg} = \frac{55}{10} = 5.5$$

$$\frac{1}{5.5}$$

$$\frac{\partial L}{\partial w_0}$$

$$\nabla w_t^2$$

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{V_t + \epsilon}} \nabla (w_t)^2$$

Avg current gradient exp.

Webinar Chat

Urmil Shah to Hosts and panelists
Thanks

Usha Kumari to Hosts and panelists
but where is squaring happening in RMS prob?

abhishek to Hosts and panelists
yes sir

Rohan to Hosts and panelists
Yes Clear

Usha Kumari to Hosts and panelists
yes sir clear

Urmil Shah to Hosts and panelists
Understood Sir

Rohan to Hosts and panelists
Adagrad - Can handle Sparse Matrix Only. RMS Prop - both Sparse and Dense

Who can see your messages? Recording on

To: Hosts and panelists

Type message here...

You are screen sharing Stop share

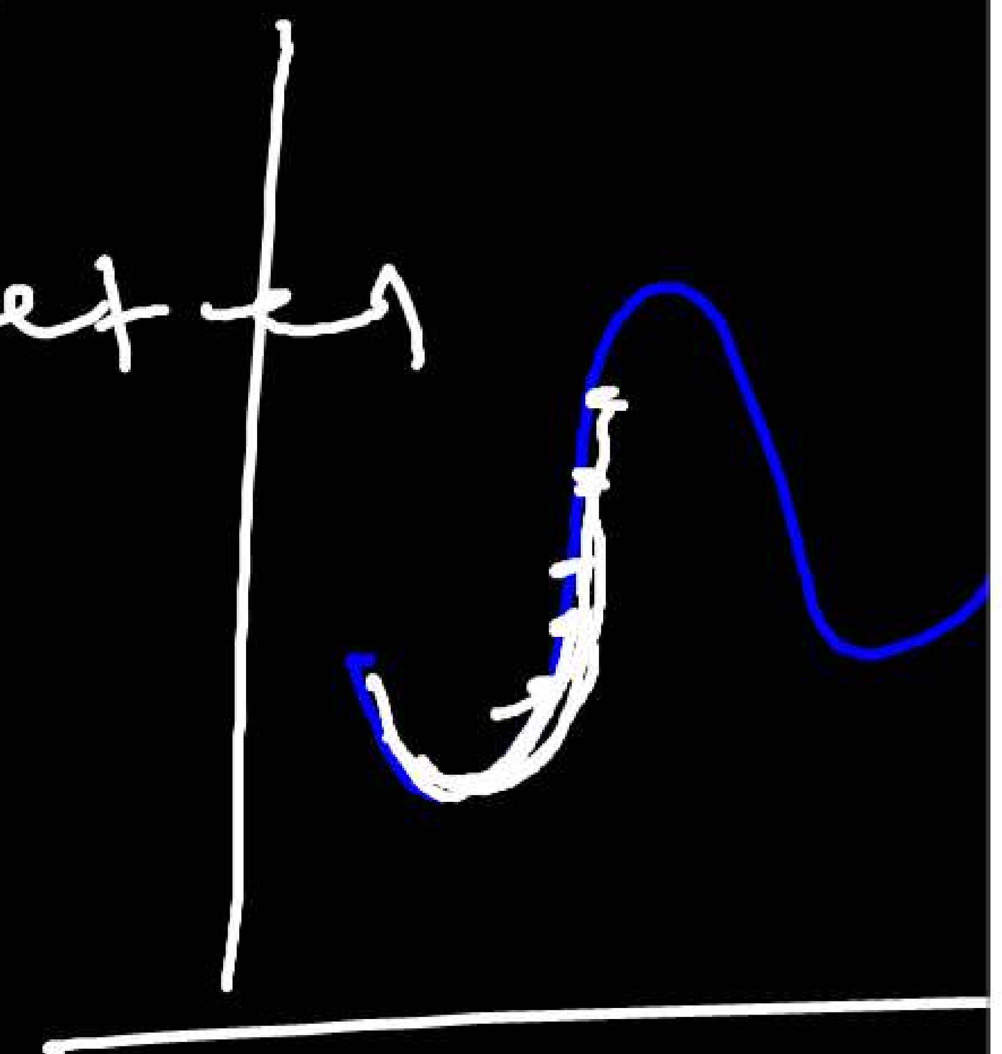
Ravinder

Ravinder

RMSProp - Dis-adv → NO

2014 - Adam - most significant

2024 - Almost 10 years completed



Adaptive Momentum Estimation,

→ it requires less memory

→ large data set

→ combination of Momentum + RMSProp $\beta = 0.9$

$$\frac{1}{1-\beta}$$

Webinar Chat

Urmil Shah to Hosts and panelists
Thanks

Usha Kumari to Hosts and panelists
but where is squaring happening in RMS prob?

abhishek to Hosts and panelists
yes sir

Rohan to Hosts and panelists
Yes Clear

Usha Kumari to Hosts and panelists
yes sir clear

Urmil Shah to Hosts and panelists
Understood Sir

Rohan to Hosts and panelists
Adagrad - Can handle Sparse Matrix Only. RMS Prop - both Sparse and Dense

Who can see your messages? Recording on

To: Hosts and panelists

Type message here...

6 / 6

You are screen sharing
Stop share

Adam :-

SGD / MBGD / BGD
Momentum
NAG
Adagrad
RMSprop / Adadelta → $w_n = w_o - \eta \frac{\partial L}{\partial w_o}$

→ Adam < Dense, Sparse
→ RMSprop
→ Momentum
] combine merge

Dis-

Most of the algorithm

ANN / MLP / CNN / RNN / LSTM / GRU etc.

↳ Adam

Ravinder

Webinar Chat

Rohan to Hosts and panelists
Yes Clear

Usha Kumari to Hosts and panelists
yes sir clear

Urmil Shah to Hosts and panelists
Understood Sir

Rohan to Hosts and panelists
Adagrad - Can handle Sparse Matrix Only. RMS Prop - both Sparse and Dense

abhishek to Hosts and panelists
can u explain more on adam

Usha Kumari to Hosts and panelists
no sir

abhishek to Hosts and panelists
adam formulae

Who can see your messages? Recording on

To: Hosts and panelists

Type message here...

7 / 8

You are screen sharing Stop share

Urmil Shah

Urmil Shah

$$\boxed{W_{t+1} = W_t - \frac{\eta}{\sqrt{V_t + \epsilon}} * m_t}$$ ← Formulas

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}$$

$$\hat{V}_t = \frac{V_t}{1-\beta_2^t}$$

$\beta_1 = 0.9$, $\beta_2 = 0.99$ — by default

-keras

where,

$$m_t = \beta_1 m_{t-1} + (1-\beta)\nabla W_t$$ — momentum

$$V_t = \beta_2 V_{t-1} + (1-\beta_2)(\nabla W_t)^2$$ — RMSprop

Webinar Chat

Urmil Shah to Hosts and panelists

Understood Sir

Rohan to Hosts and panelists

Adagrad - Can handle Sparse Matrix Only. RMS Prop - both Sparse and Dense

abhishek to Hosts and panelists

can u explain more on adam

Usha Kumari to Hosts and panelists

no sir

abhishek to Hosts and panelists

adam formulae

Usha Kumari to Hosts and panelists

so finally we need to concentrate more on Adam, Sigmoid and ReLU as of now

Who can see your messages? Recording on

To: Hosts and panelists

Type message here...

8 / 8

How to improve a neural Network

→ fine tuning Neural Network hyperparameters.

- epochs
- # hidden layer
- # Neurons
- Learning Rate ↓ Optimization
- Activation ↓

→ By Solving Problem
→ vanishing Gradient Prob
→ Less data → Slow training
→ overfitting

You are screen sharing
Stop share

Ravinder

Ravinder

- → No. of hidden layer
- → No. of Neurons per layer
- → learning rate
- → best optimization — Adam/RMSprop
- → Batch Normalization / early stopping / Regularization / Dropout
- → batch size — BGD, SGD, MBGD
- → Activation
- → epochs