NPTEL ONLINE CERTIFICATION COURSES

Course Name: Deep Learning

Faculty Name: Prof. P. K. Biswas

Department : E & ECE, IIT Kharagpur

Topic

Lecture 45: Optimizing Gradient Descent III

CONCEPTS COVERED

Concepts Covered:

- CNN
  - Gradient Descent Challenges
  - Momentum Optimizer
  - Nesterov Accelerated Gradient
  - Adagrad
  - RMSProp
  - etc.

Adagrad

$$g_t = \frac{1}{n} \sum_{\forall X \in Minibatch} \nabla_W L(W_t, X) \qquad r_t = \sum_{\tau=1}^{t} g_\tau \circ g_\tau$$

$$W_{t+1} = W_t - \frac{\eta}{\sqrt{\in I + r_t}} \circ g_t$$

$\circ \rightarrow$ element - wise product

Adagrad

Positive Side:

- Adagrad adaptively scales the learning rate for different dimensions by normalizing with respect to the gradient magnitude in the corresponding dimension.
- Adagrad eliminates the need to manually tune the learning rate.
- Reduces learning rate faster for parameters showing large slope and slower for parameters giving smaller slope.
- Adagrad converges rapidly when applied to convex functions.

Adagrad

Negative side:

- If the function is non-convex:- trajectory may pass through many complex terrains eventually arriving at a locally region.
- By then learning rate may become too small due to the accumulation of gradients from the beginning of training.
- So at some point the model may stop learning.

RMSProp

RMSProp

- RMSProp uses exponentially decaying average of squared gradient and discards history from the extreme past.
- Converges rapidly once it finds a locally convex bowl.
- Treats this as an instance of Adagrad algorithm initialized within that bowl.

RMSProp

$$g_t = \frac{1}{n} \sum_{\forall X \in Minibatch} \nabla_W L(W_t, X)$$

$$r_t = \beta r_{t-1} + (1-\beta) g_t \circ g_t$$ → Exponentially decaying average

$$W_{t+1} = W_t - \frac{\eta}{\sqrt{\in I + r_t}} \circ g_t$$

RMSProp with Nesterov Momentum

$$\tilde{W} = W_t + \alpha v \qquad g_t = \frac{1}{n} \sum_{\forall X \in Minibatch} \nabla_W L(\tilde{W}, X)$$

$$r_t = \beta r_{t-1} + (1 - \beta) g_t \circ g_t$$

$$v_{t+1} = \alpha v_t - \frac{\eta}{\sqrt{\in I + r_t}} \circ g_t \qquad W_{t+1} = W_t + v_t$$

Adaptive Moments (Adam)

Adam

- Variant of the combination of RMSProp and Momentum.
- Incorporates first order moment (with exponential weighting) of the gradient (Momentum term).
- Momentum is incorporated in RMSProp by adding momentum to the rescaled gradients.
- Both first and second moments are corrected for bias to account for heir initialization to zero.

Adam

$$g_t = \frac{1}{n} \sum_{\forall X \in Minibatch} \nabla_W L(W, X)$$

Biased first and second moments

$$s_t = \beta_1 s_{t-1} + (1 - \beta_1) g_t$$

$$r_t = \beta_2 r_{t-1} + (1 - \beta_2) g_t \circ g_t$$

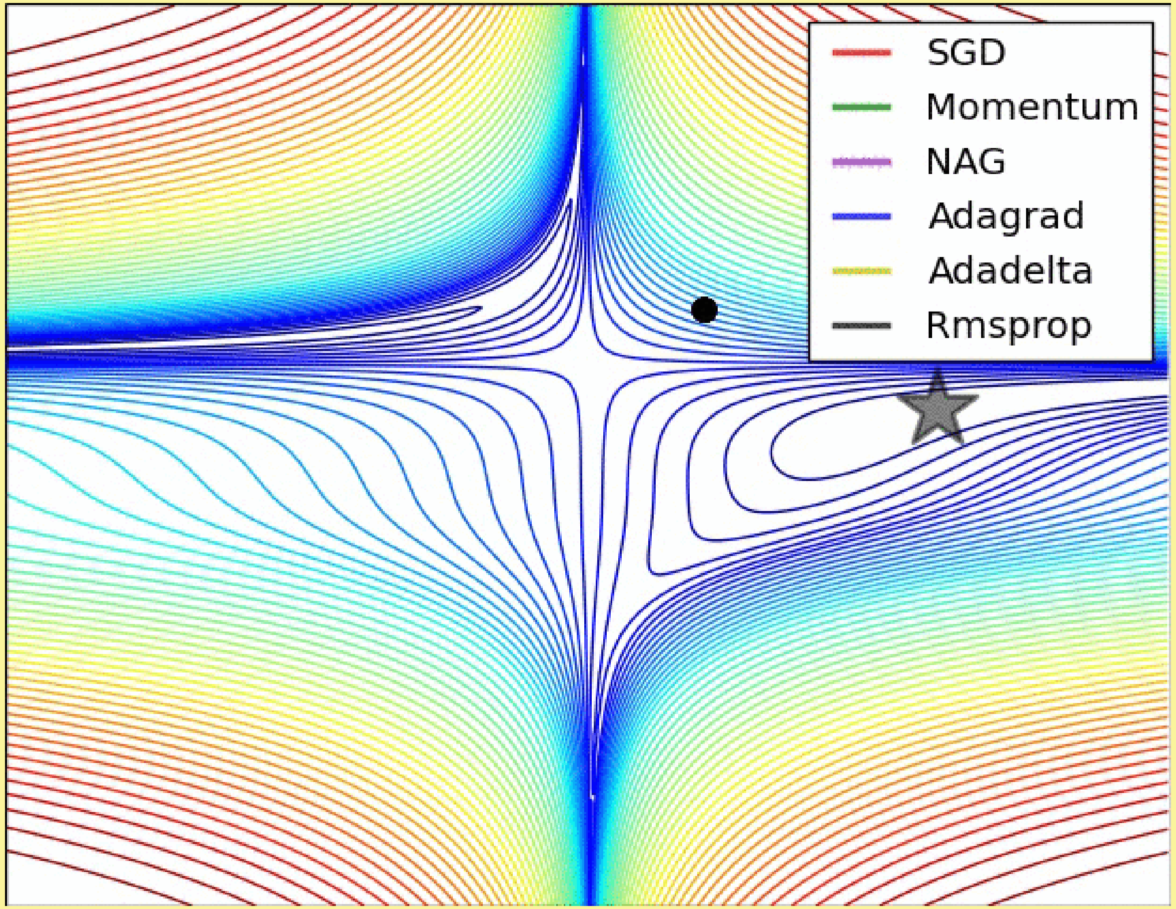
Adam

Bias corrected first and second moments

$$\hat{s}_t = \frac{s_t}{1-\beta_1} \qquad \hat{r}_t = \frac{r_t}{1-\beta_2}$$

$$W_{t+1} = W_t - \eta \frac{\hat{s}_t}{\sqrt{\in I + \hat{r}_t}}$$

Momentum Optimizer



Animation Source:-
https://imgur.com/a/Hqolp

NPTEL ONLINE CERTIFICATION COURSES

Thank you