Week 6 Notes

**Optimization with Momentum**

If we go repeatedly in the same direction, then we should increase the step size to reach the target faster.

Text

Description automatically generated

Alpha is between 0 and 1.

Velocity is exponentially decaying moving average of negative gradients.



The recent velocities have higher importance compared to older ones.

Drawback. Because of high velocity, the gradient may overshoot its target.

Nesterov momentum looks ahead for the next step in the gradient descent and figures out if the target is overshot and updates the velocity accordingly.

Text

Description automatically generated

Diagram

Description automatically generated

**Stochastic Gradient Descent**

Graphical user interface, text

Description automatically generatedText

Description automatically generated

**AdaGrad, RMSProp and Adam**

**Text, letter

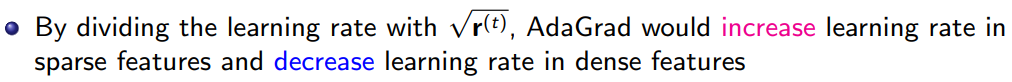
Description automatically generated**

SGD/GD applies the same step size to all the features.

In adaptive GD we need to apply different step size to different features.

AdaGrad

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated

RMSProp – Root Mean Square Propagation

Graphical user interface, text, application

Description automatically generated

Adam – Adaptive Momentum Estimation

Graphical user interface, text

Description automatically generatedGraphical user interface, text, application, chat or text message

Description automatically generated

Text, letter

Description automatically generated