Neural Networks

Seminar - Week 4

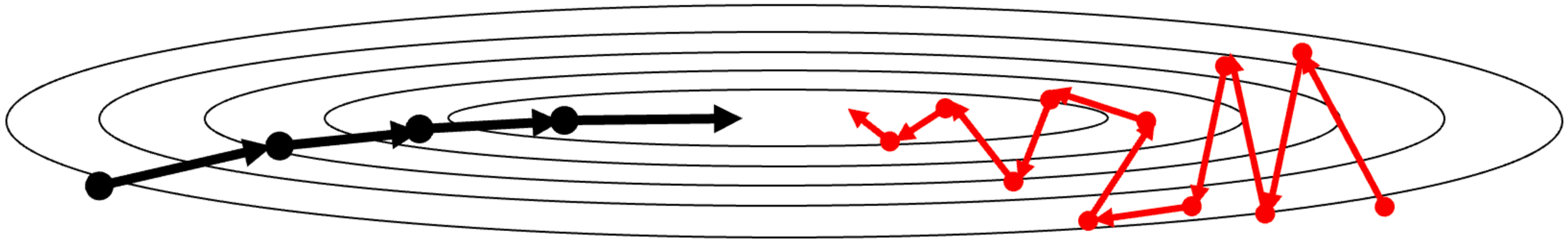
Igor Janos

Seminar - Week 4

Assignment 3 Goals

- Implement optimizers
    - SGD with Momentum
    - RMSProp
    - Adam

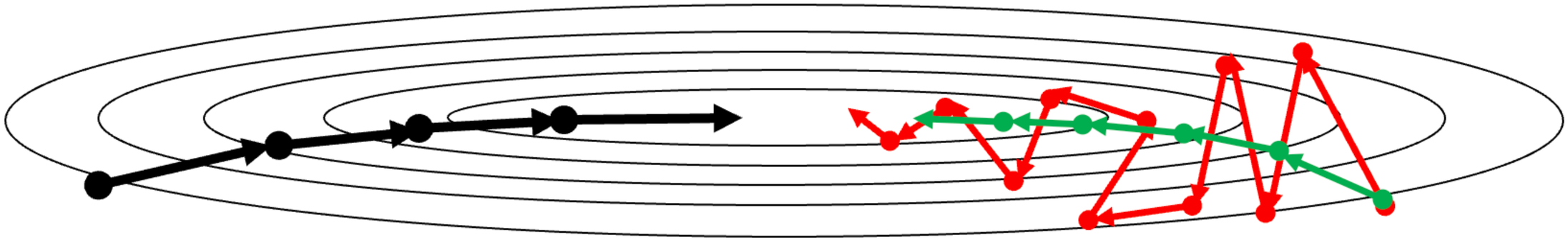
Gradient Descent



Gradient Descent

$$W := W - \alpha dW$$
$$b := b - \alpha db$$

Gradient Descent - Momentum



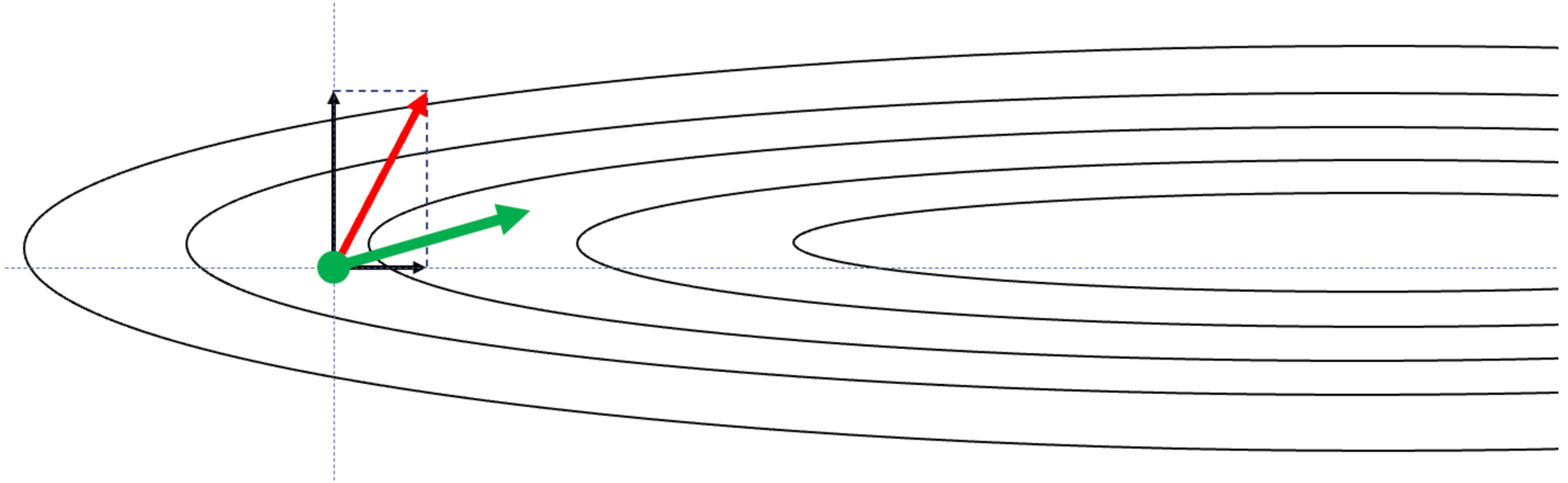
Gradient Descent

$$W := W - \alpha dW$$
$$b := b - \alpha db$$

Gradient Descent with Momentum

$$V_{dW} := (1 - \beta)dW + \beta V_{dW}$$
$$V_{db} := (1 - \beta)db + \beta V_{db}$$
$$W := W - \alpha V_{dW}$$
$$b := b - \alpha V_{db}$$

RMSProp



$$S_{dW} := (1-\beta)dW^2 + \beta S_{dW}$$

$$S_{db} := (1-\beta)db^2 + \beta S_{db}$$

$$W := W - \alpha \frac{dW}{\sqrt{S_{dW}}}$$

$$b := b - \alpha \frac{db}{\sqrt{S_{db}}}$$

Unpublished, Proposed by Geoffrey Hinton

Adam

- Combines ideas from momentum and RMSProp

$$V_{dW} := (1-\beta_1)dW + \beta_1 V_{dW}$$
$$S_{dW} := (1-\beta_2)dW^2 + \beta_2 S_{dW}$$

$$\widehat{V}_{dW} := \frac{V_{dW}}{1-\beta_1^t}$$
$$\widehat{S}_{dW} := \frac{S_{dW}}{1-\beta_2^t}$$

$$W := W - \alpha \frac{\widehat{V}_{dW}}{\sqrt{\widehat{S}_{dW}} + \epsilon}$$

Small value
for numerical stability

Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).