Machine and Deep Learning

Holger Schmidt

May 8, 2019

Inhaltsverzeichnis

Überblick Themen

(00) Mathematik/Python
(01) Linear Regression
(02) Binary Classification
(03) Multiclass Classification
(04) Neural Networks I
(05) Dimensionality Reduction/ Principal Component Analysis
(06) Decision Trees
(07) Neural Networks II (Tensorflow/Keras)
(08) Convolutional Networks I
(09) Convolutional Networks II
(10) Recurrent Networks

Gradient Descent

(1) Calculate Gradient $\mathrm{d}\mathbf{W}$ (using **all** m training examples)

(2) Perform parameter update ("1 epoch")

$$W = W - \alpha \cdot \mathrm{d}\mathbf{W}$$

Gradient Descent

(1) Calculate Gradient $\mathrm{d}\mathbf{W}$ (using **all** m training examples)

(2) Perform parameter update ("1 epoch")

$$W = W - \alpha \cdot \mathrm{d}\mathbf{W}$$

Stochastic Gradient Descent

For ech epoch:
Parameter update is performed m times for every single training example

Gradient Descent

(1) Calculate Gradient $\mathrm{d}\mathbf{W}$ (using **all** m training examples)

(2) Perform parameter update ("1 epoch")

$$W = W - \alpha \cdot \mathrm{d}\mathbf{W}$$

Stochastic Gradient Descent

For ech epoch:
Parameter update is performed m times for every single training example

Mini Batch Gradient Descent

For ech epoch:
Parameter update is performed $\frac{m}{m_{\mathrm{mb}}}$ times for each mini batch of size m_{mb}

Gradient Descent

(1) Calculate Gradient $\mathrm{d}\mathbf{W}$ (using **all** m training examples)

(2) Perform parameter update ("1 epoch")

$$W = W - \alpha \cdot \mathrm{d}\mathbf{W}$$

Stochastic Gradient Descent

For ech epoch:
Parameter update is performed m times for every single training example

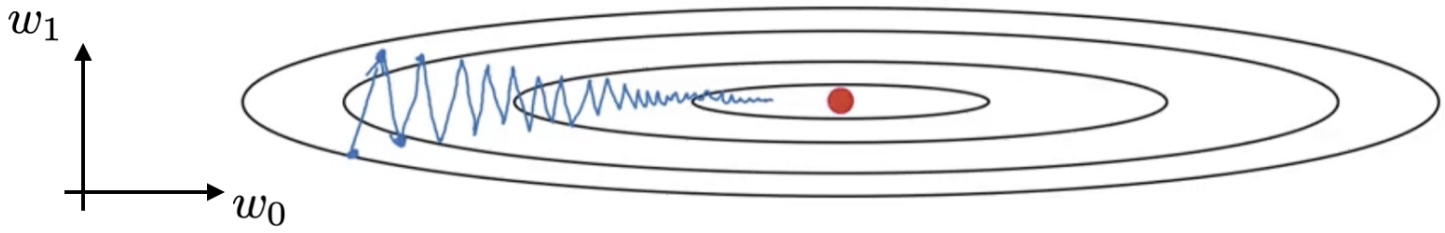
Mini Batch Gradient Descent

For ech epoch:
Parameter update is performed $\frac{m}{m_{\mathrm{mb}}}$ times for each mini batch of size m_{mb}

Batch Gradient Descent

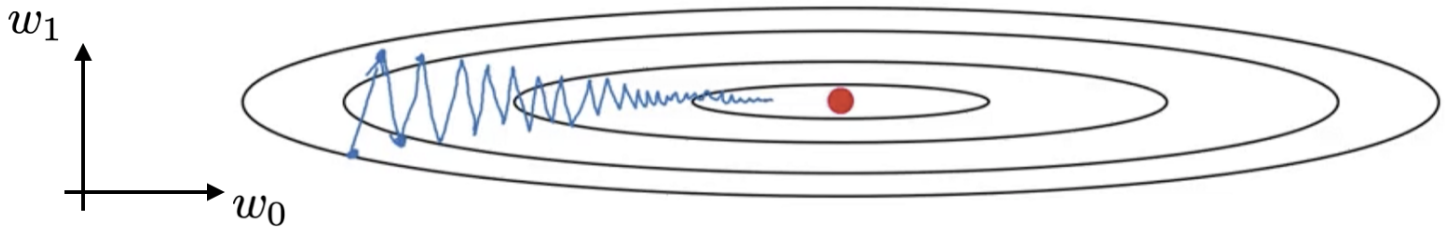
For ech epoch:
Gradient descent step is performed for all m training example

RMSprop



Standard gradient descent

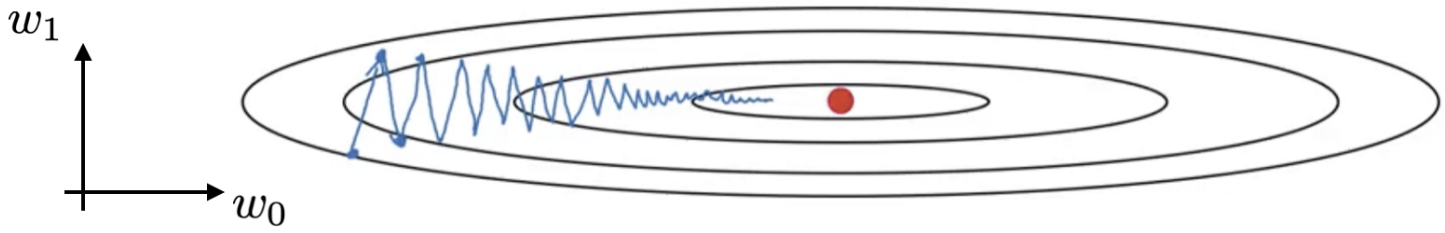
RMSprop



Standard gradient descent

Main idea: Speed up convergence in w_0 direction

RMSprop

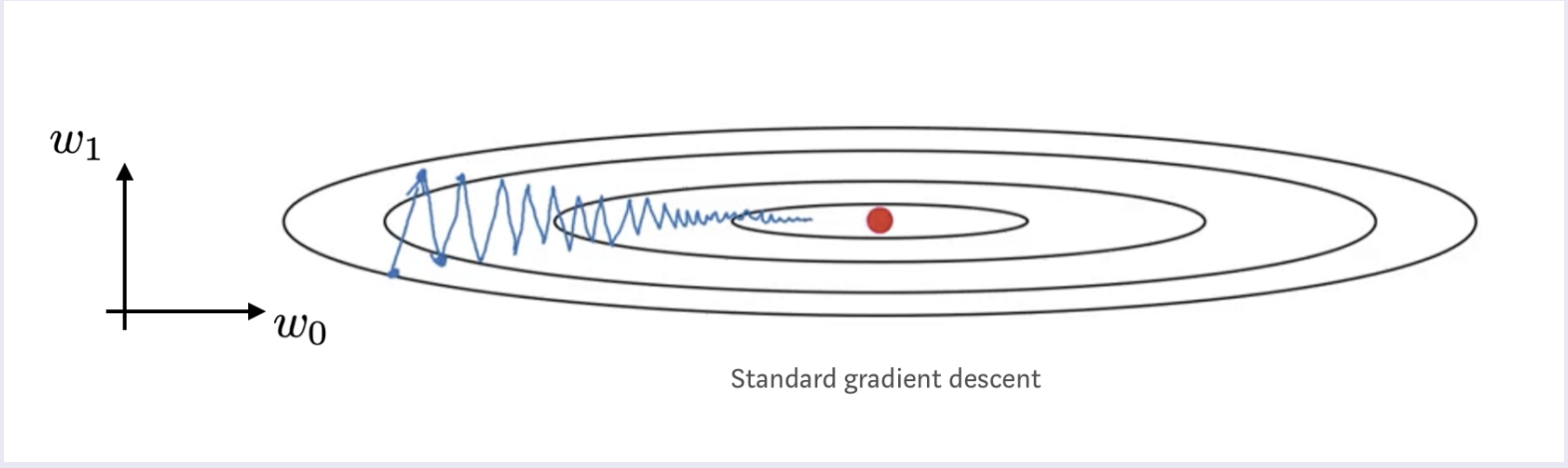


Standard gradient descent

Main idea: Speed up convergence in w_0 direction

$$
\begin{aligned}
S_{d\mathbf{W}} &= \beta S_{d\mathbf{W}} + (1-\beta) d\mathbf{W}^2 \\
\mathbf{W} &= \mathbf{W} - \alpha \frac{d\mathbf{W}}{\sqrt{S_{d\mathbf{W}}}}
\end{aligned}
$$

RMSprop



Main idea: Speed up convergence in w_0 direction

$$
\begin{aligned}
S_{d\mathbf{W}} &= \beta S_{d\mathbf{W}} + (1-\beta) d\mathbf{W}^2 \\
\mathbf{W} &= \mathbf{W} - \alpha \frac{d\mathbf{W}}{\sqrt{S_{d\mathbf{W}}}}
\end{aligned}
$$

Meaning of $S_{d\mathbf{W}}$: Exponentially weighted moving average of the squares of the weights dw^2.

K-fold Crossvalidation

IMDb Dataset - Binary Classification

Reviews of Movies (positiv/negativ).
Each review has to be cast into a One-Hot-Encoding representation.

Fashion MNIST - Multiclass Classification

Pictures of $K = 10$ classes of fashion items.
Rescale pictures from $[0, 255]$ to $[0, 1]$.

Boston Housing - Regression

Prices of Houses as a function of several features (average number of rooms, crime rate, access to highways ...).
Get read of mean and rescale features.
Very few examples $\rightarrow$ use K-fold crossvalidation

Boston Housing - Regression

Mean absolute error (MAE) of the Validation Set using K-fold validation with $K = 4$