XNDL - Lab
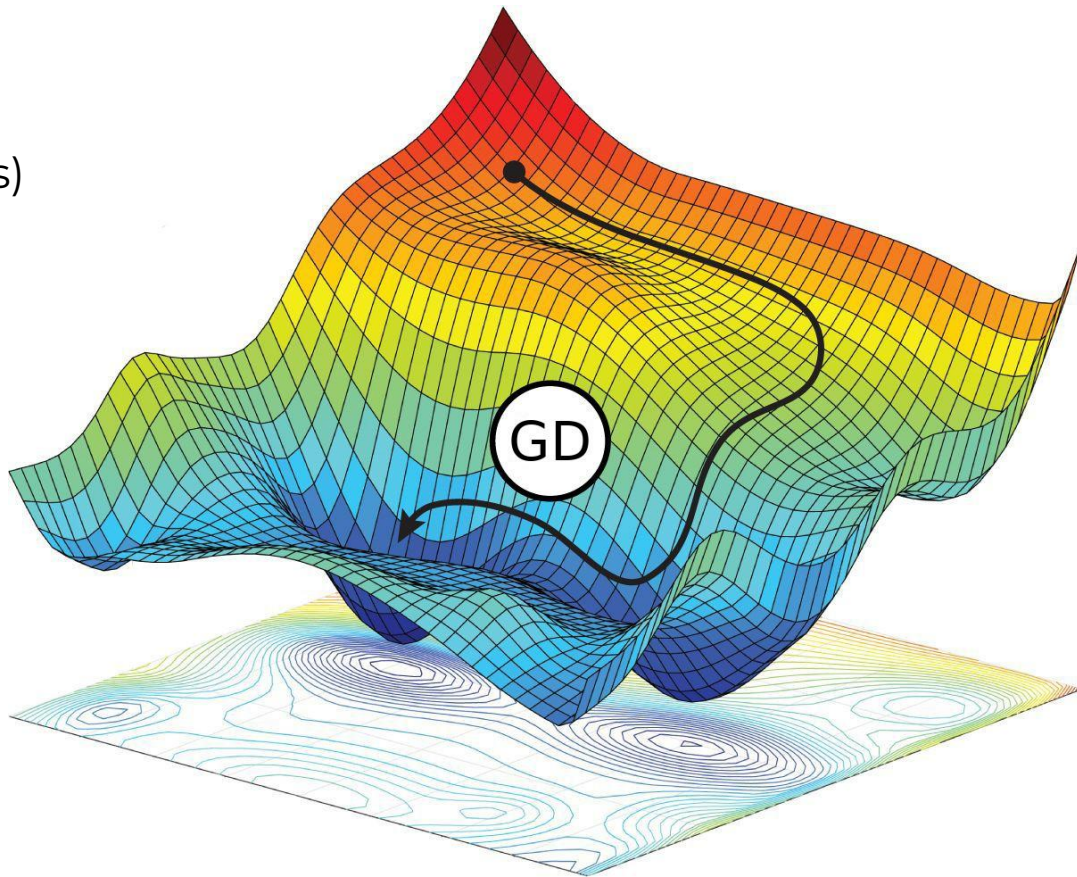
Delta rule, learning rate, momentum and early stopping

Delta rule

❖ Gradient Descent

❖ Optimization of a function (loss)

❖ Watch your speed

Loss Landscape


MODE CONNECTIVITY

OPTIMA OF COMPLEX LOSS FUNCTIONS CONNECTED BY SIMPLE CURVES OVER
WHICH TRAINING AND TEST ACCURACY ARE NEARLY CONSTANT

BASED ON THE PAPER BY TIMUR GARIPOV, PAVEL IZMAILOV, DMITRII PODOPRIKHIN, DMITRY VETROV, ANDREW GORDON WILSON
VISUALIZATION & ANALYSIS IS A COLLABORATION BETWEEN TIMUR GARIPOV, PAVEL IZMAILOV AND JAVIER IDEAMI@LOSSLANDSCAPE.COM

NeurIPS 2018, ARXIV:1802.10026 | LOSSLANDSCAPE.COM

LOSS (TRAIN MODE)

REAL DATA, RESNET-20 NO-SKIP,
CIFAR10, SGD-MOM, BS=128
WD=3e-4, MOM=0.9
BN, TRAIN MOD, 90K PTS
LOG SCALED (ORIG LOSS NUMS)

MINIMA

MINIMA

3.75

3.6

2.8

0.15

0.16

Delta rule

1. *Forward*:   out = $\sigma(w * x + b)$ x: vector of inputs σ: act.func.

2. *Error*:   e = - (gt * log(out) + (1 - gt) * log(1 - out))   binary cross entropy

3. *Gradient*: $\delta = \partial L / \partial z$ derivative of loss func wrt weighted input

4. *Weight update*: $\nabla w = \delta * x$ w' = w - $\alpha * \nabla w$ α: learning rate
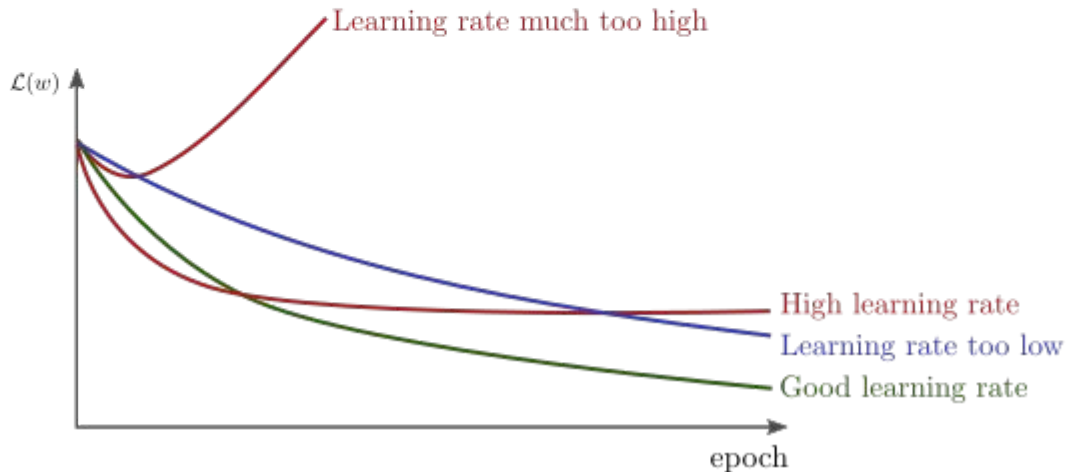
Learning rate

❖ Confidence

❖ Paralysis vs Bouncing

❖ Small > Big

❖ Decay

Learning rate

Learning rate too low   Good learning rate   High learning rate   Learning rate much too high
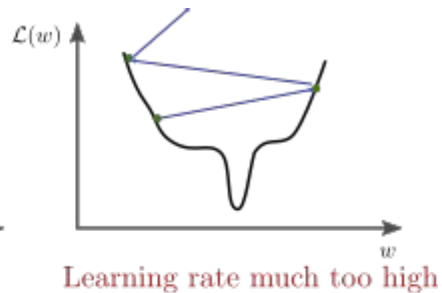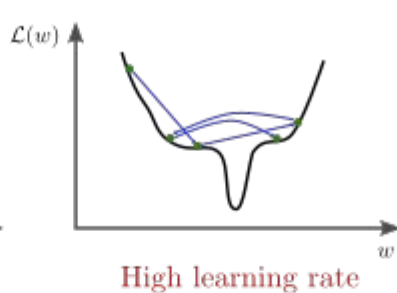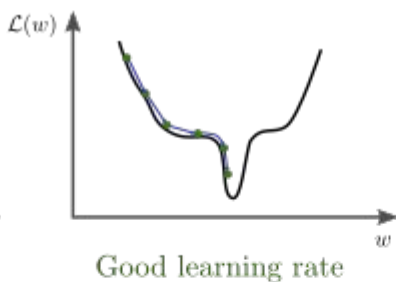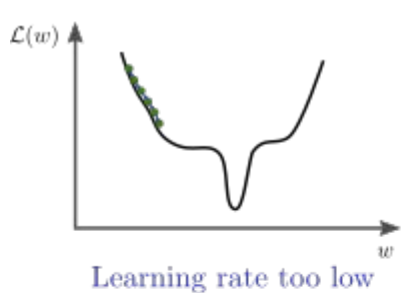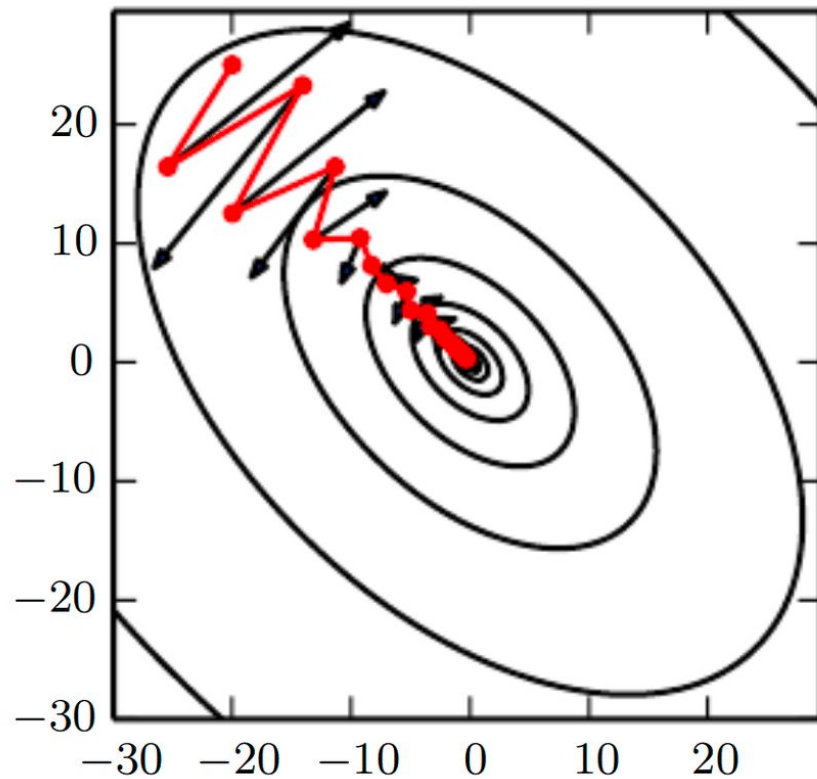
❖ Confidence

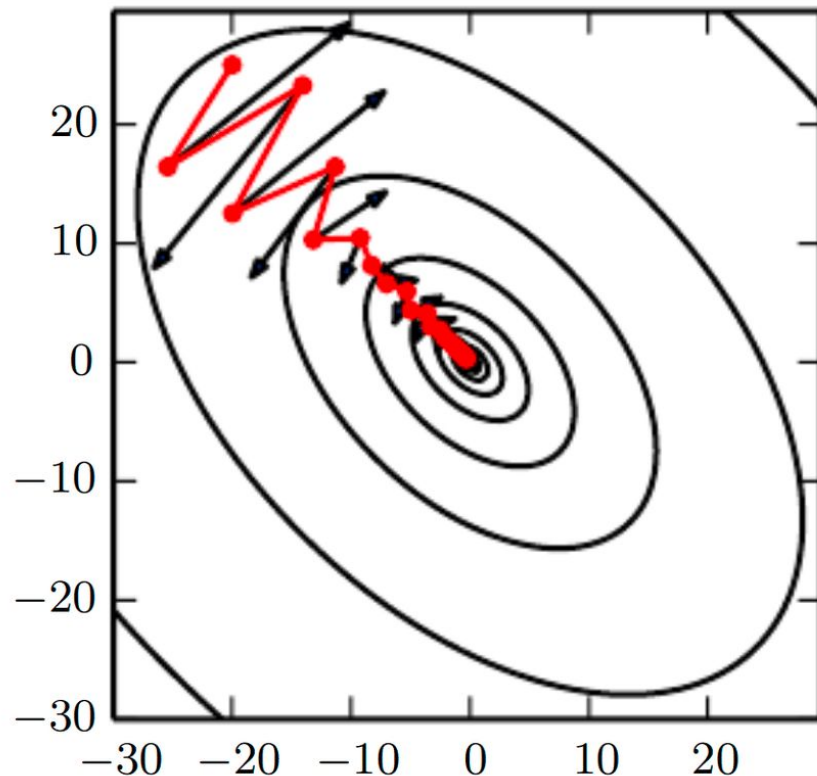❖ Paralysis vs Bouncing

❖ Small > Big

❖ Decay

Momentum

❖ Fraction of prev. update

Momentum

- ❖ Fraction of prev. update

- ❖ Inertia

- ❖ Reduce impact of noisy gradients

- ❖ Keep general direction

- ❖ Faster convergence

Early Stopping

❖ A policy

❖ val > train

❖ loss > acc

Breast Cancer dataset

- ❖ Using a perceptron

    - ■ Define & adjust your early stopping policy

    - ■ Tune learning rate

- ❖ Very small LR: Stuck

- ❖ Very large LR: Unstable

- ❖ Momentum on/off

- ❖ Look at the curves

    - ■ val vs train

    - ■ loss vs acc

- ❖ Conf. Matrix

Gates dataset

A very different domain

The end