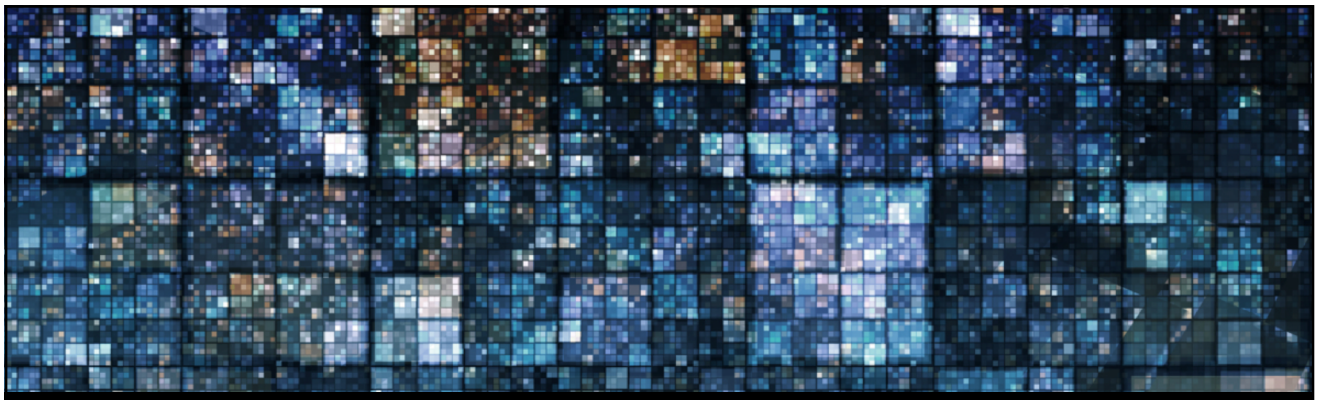
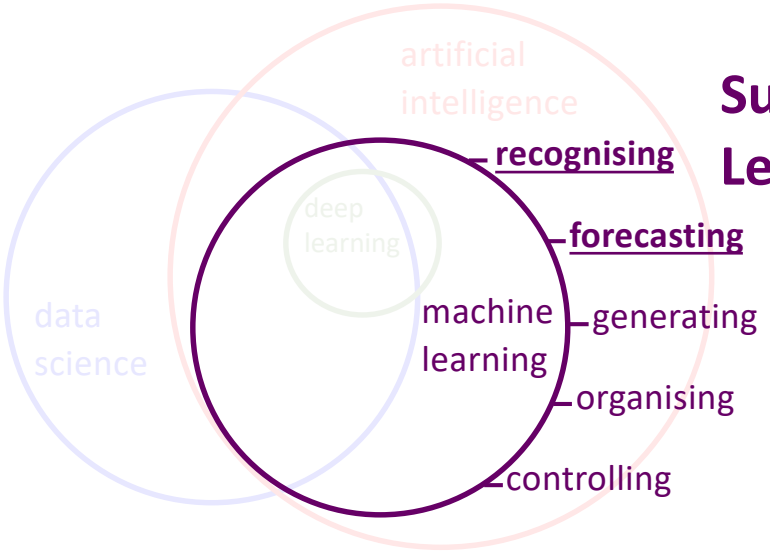


COMP47590
Advanced Machine Learning
Deep Learning - Optimisation

Dr. Brian Mac Namee

Information

Email:	Brian.MacNamee@ucd.ie
Course Materials:	All material posted on UCD CS moodlehttps://csmoodle.ucd.ie/moodle/course/view.php?id=663 Enrolment key **UCDAdvML2019**



Supervised Learning

- recognising
- forecasting
- generating
- organising
- controlling

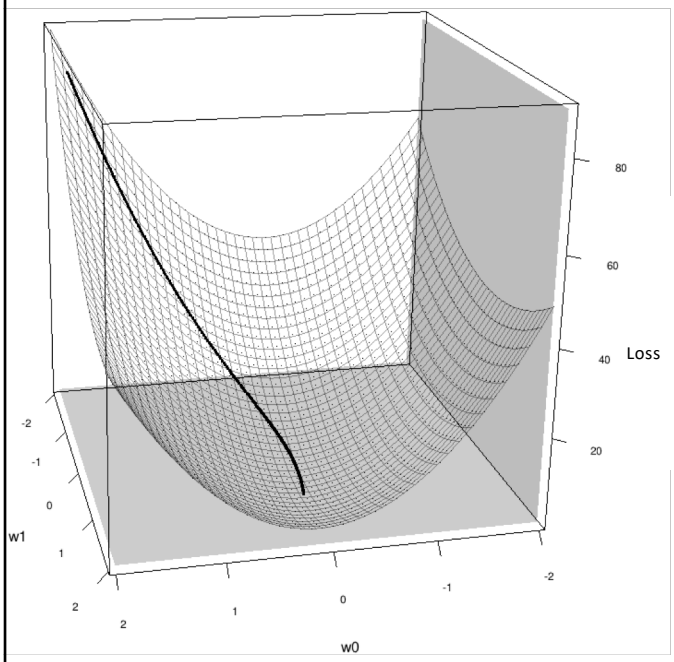
GRADIENT DESCENT & LEARNING RATES

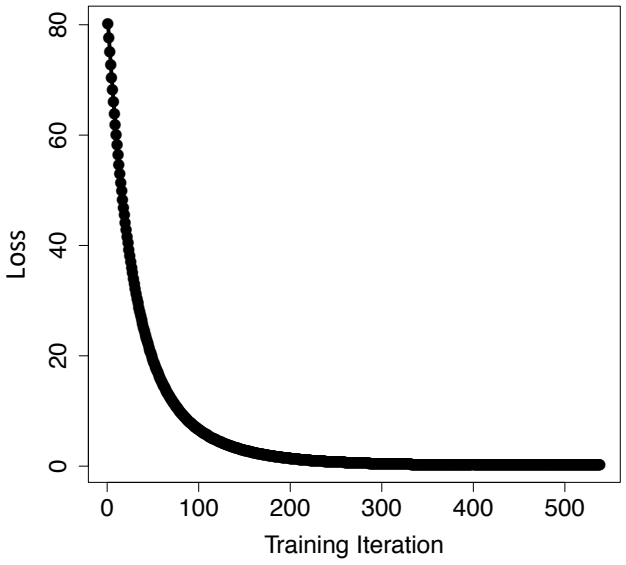
Gradient Descent

We introduced three variants of the gradient descent algorithm:

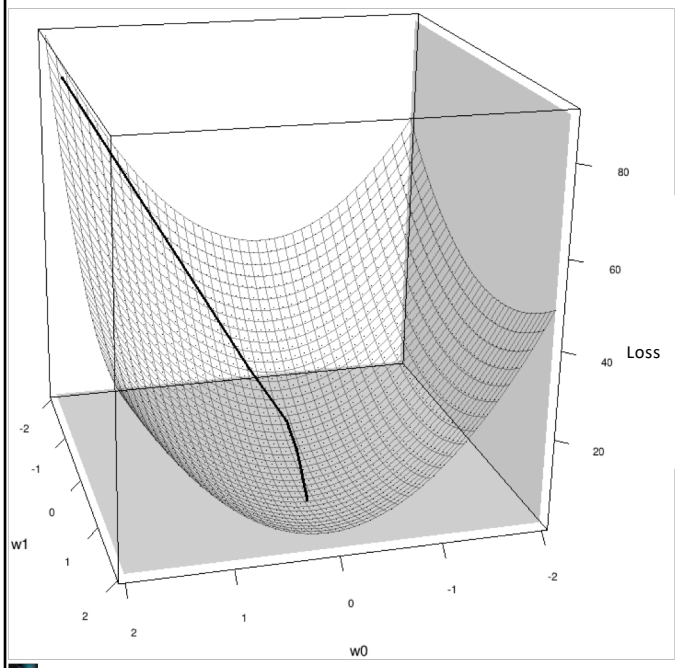
- Stochastic gradient descent
- Mini-batch gradient descent
- Batch gradient descent

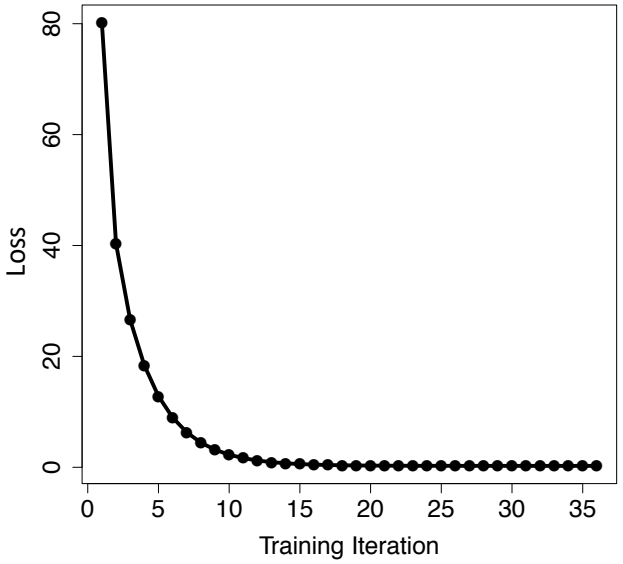
Learning rate is the key control we have over the gradient descent process

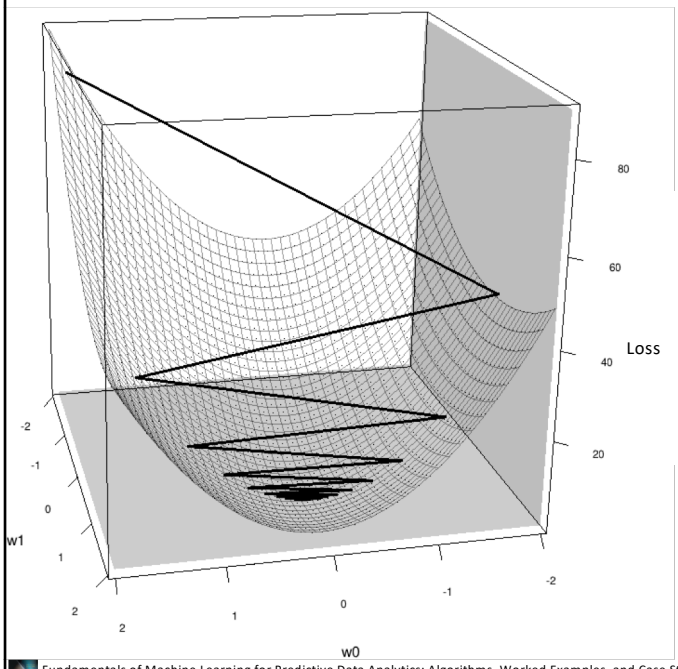


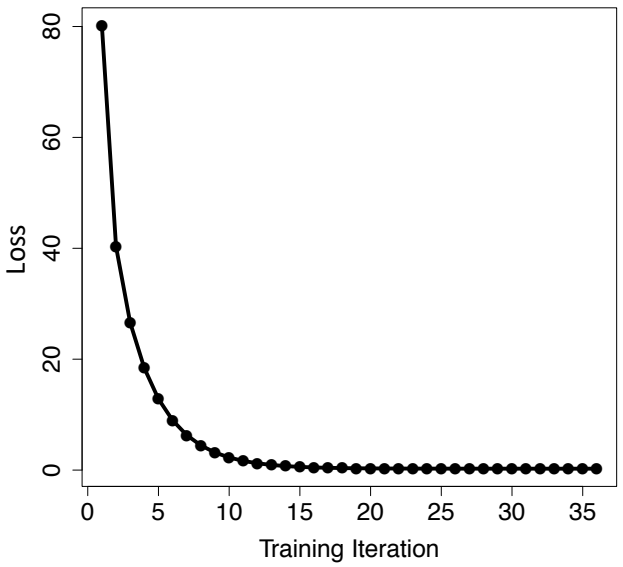


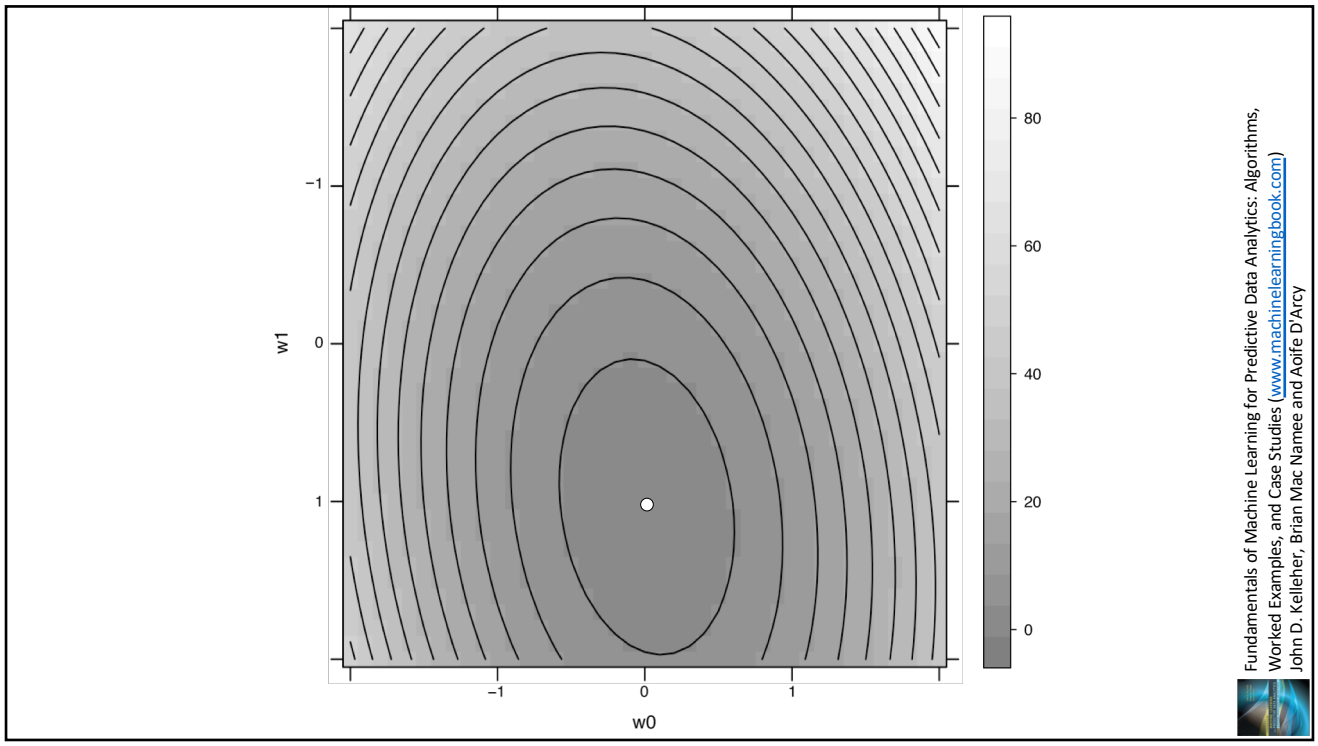

Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

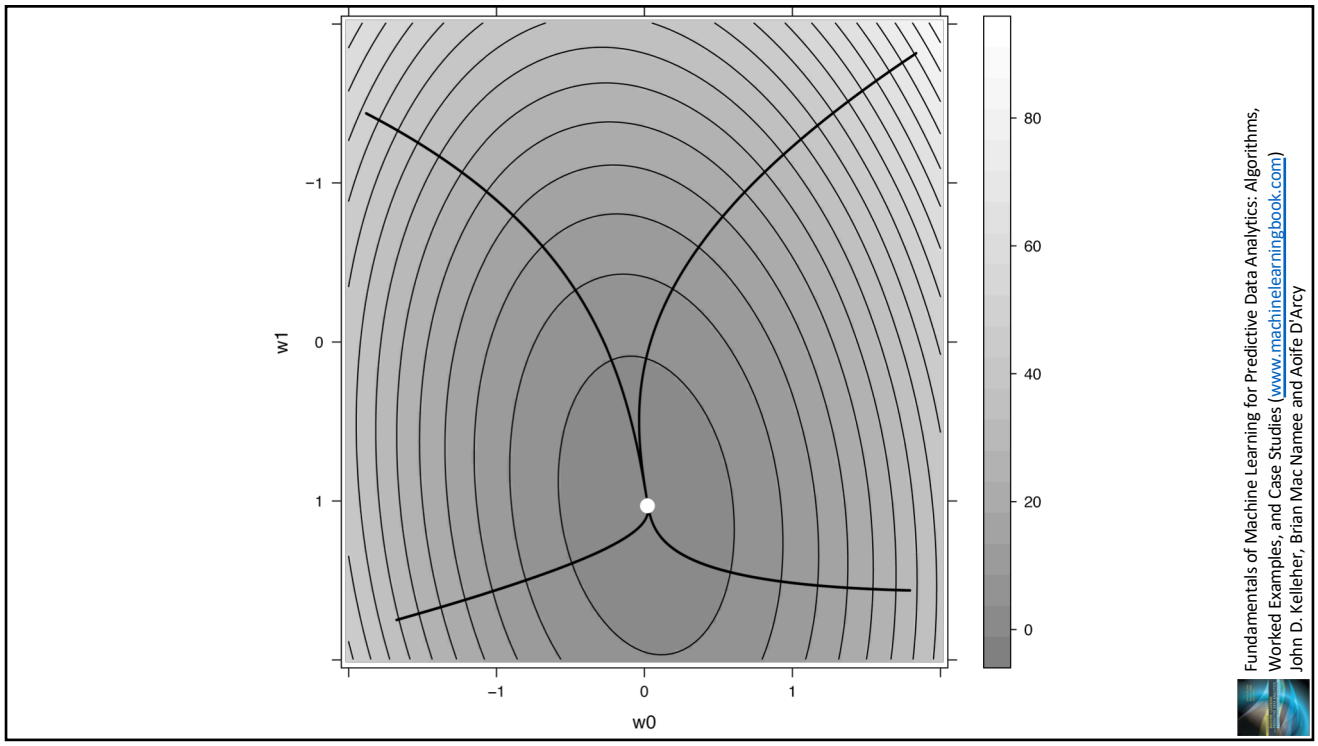

Loss
w1
w0
80
60
40
20
-2
-1
0
1
2
2
1
0
-1
-2

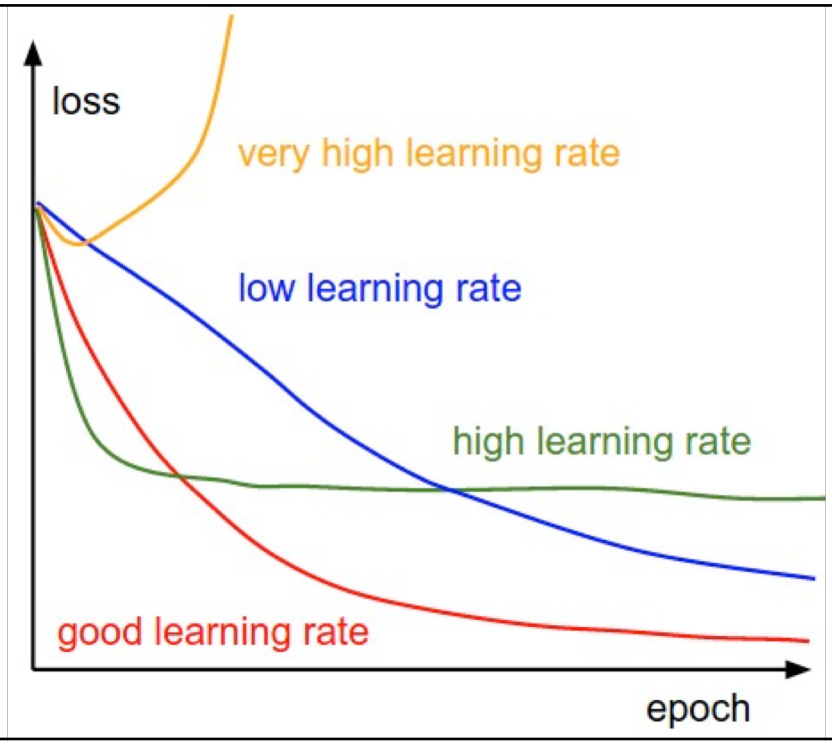

Loss
Training Iteration
0
20
40
60
80
0
5
10
15
20
25
30
35


Loss
w1
w0
80
60
40
20
-2
-1
0
1
2
2
1
0
-1
-2


Loss
Training Iteration
0
20
40
60
80
0
5
10
15
20
25
30
35


w1
w0
−1
0
1
80
60
40
20
0
Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy


w1
w0
−1
0
1
80
60
40
20
0
Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy



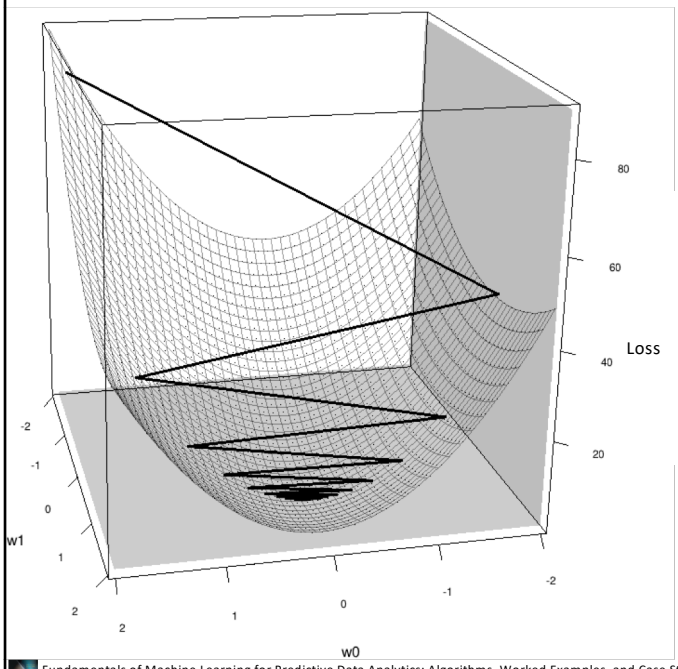
CS231n Convolutional Neural Networks for Visual Recognition
http://cs231n.github.io/neural-networks-3/

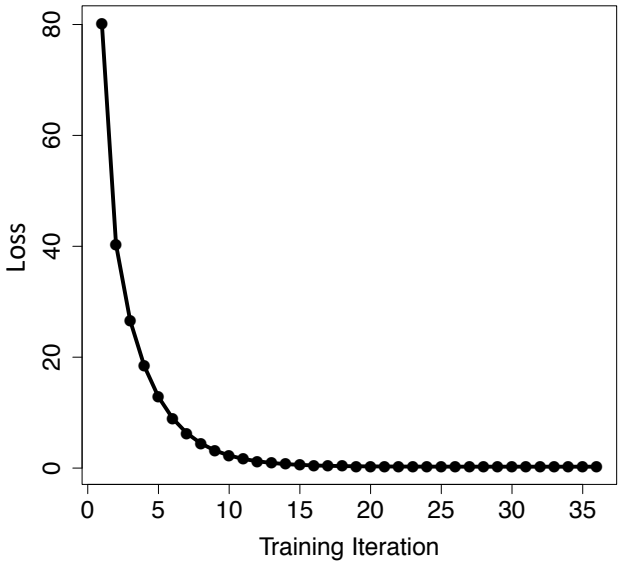
Gradient Descent

There are a set of common alternatives to basic gradient descent that can improve its performance

- Learning rate decay
- Gradient descent with momentum
- RMSPROP
- Adam

LEARNING RATE DECAY






Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

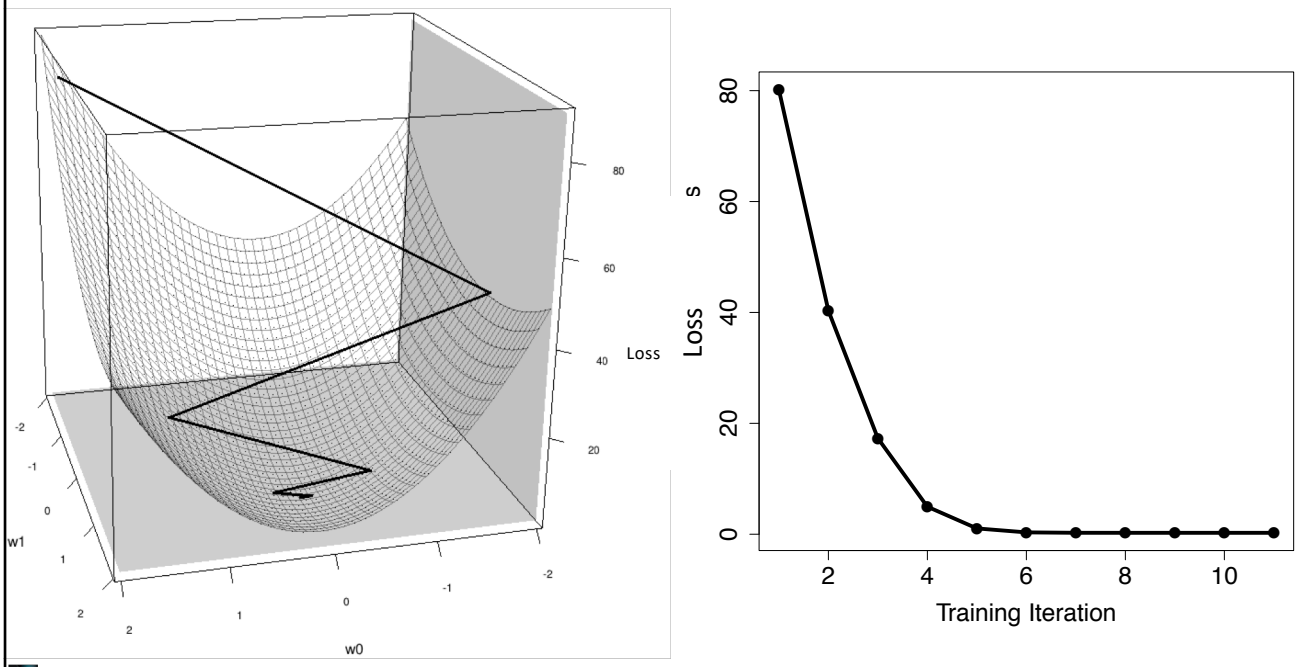
Learning Rate Decay

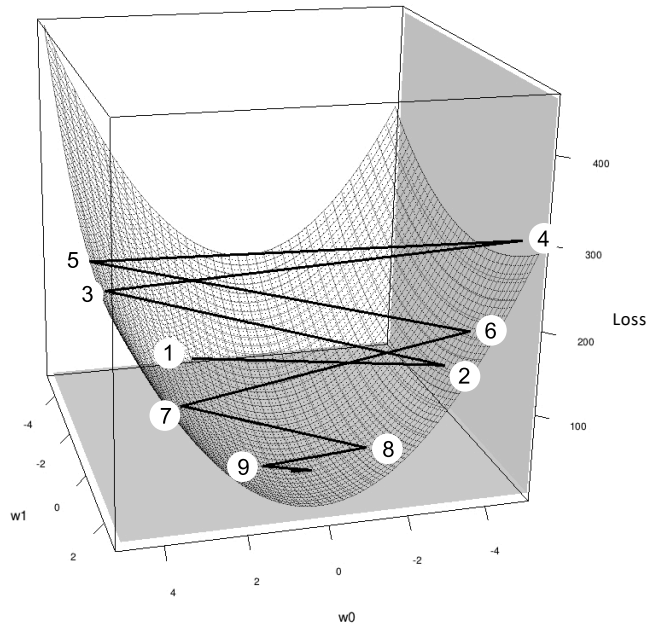
Optimises gradient descent by taking big steps at the start of the learning process and small steps later on

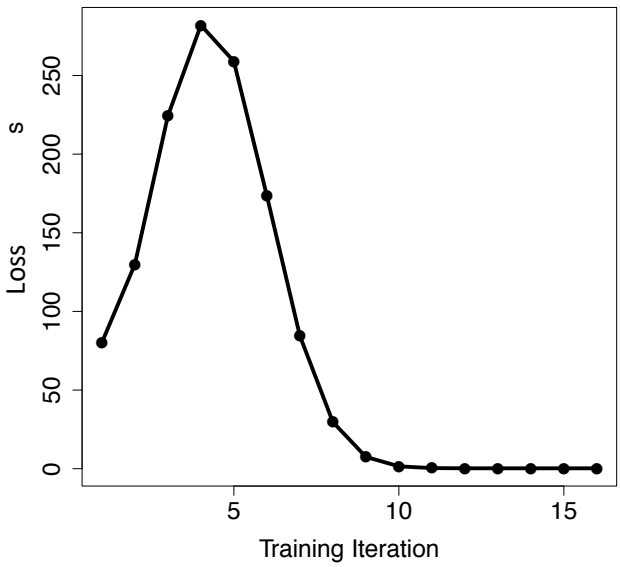
Learning rate decay can be applied as:

$$\alpha_\tau = \alpha_0 \frac{c}{c + \tau}$$

where α_0 is an initial learning rate, c is a constant, and τ is the current iteration

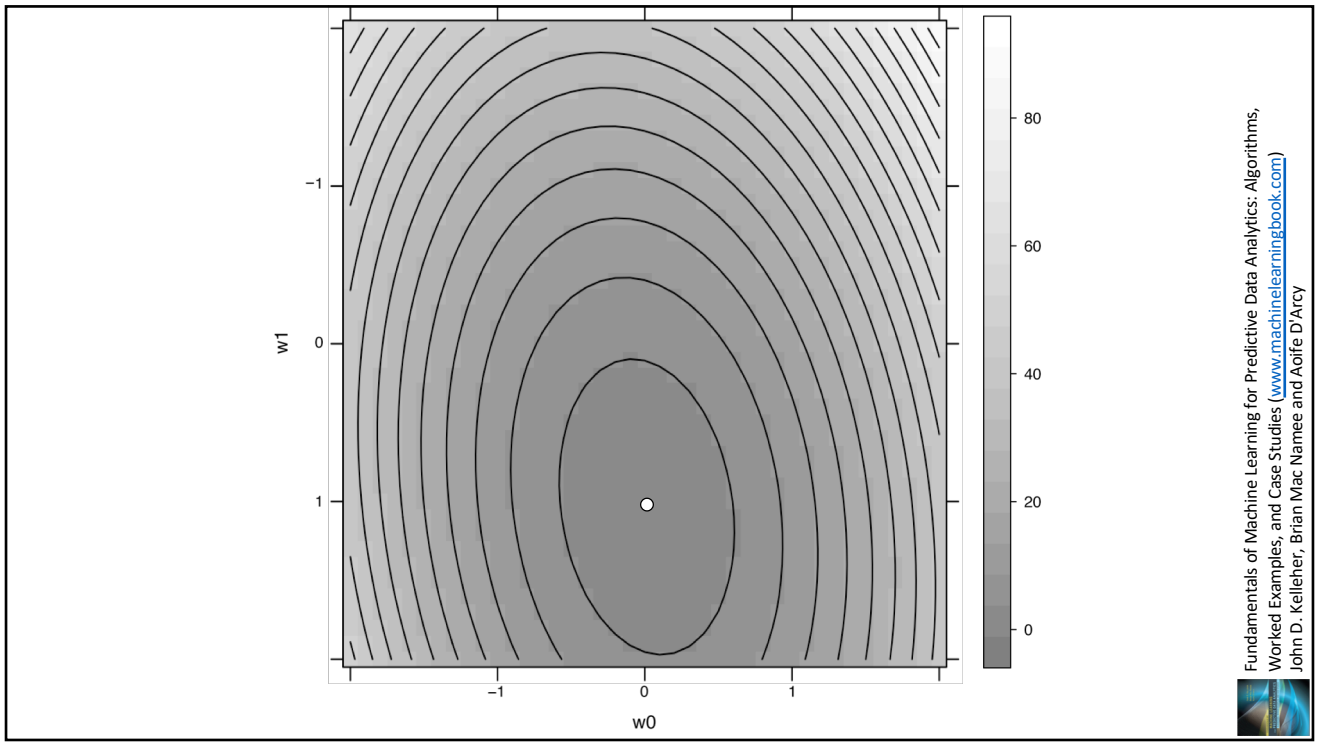


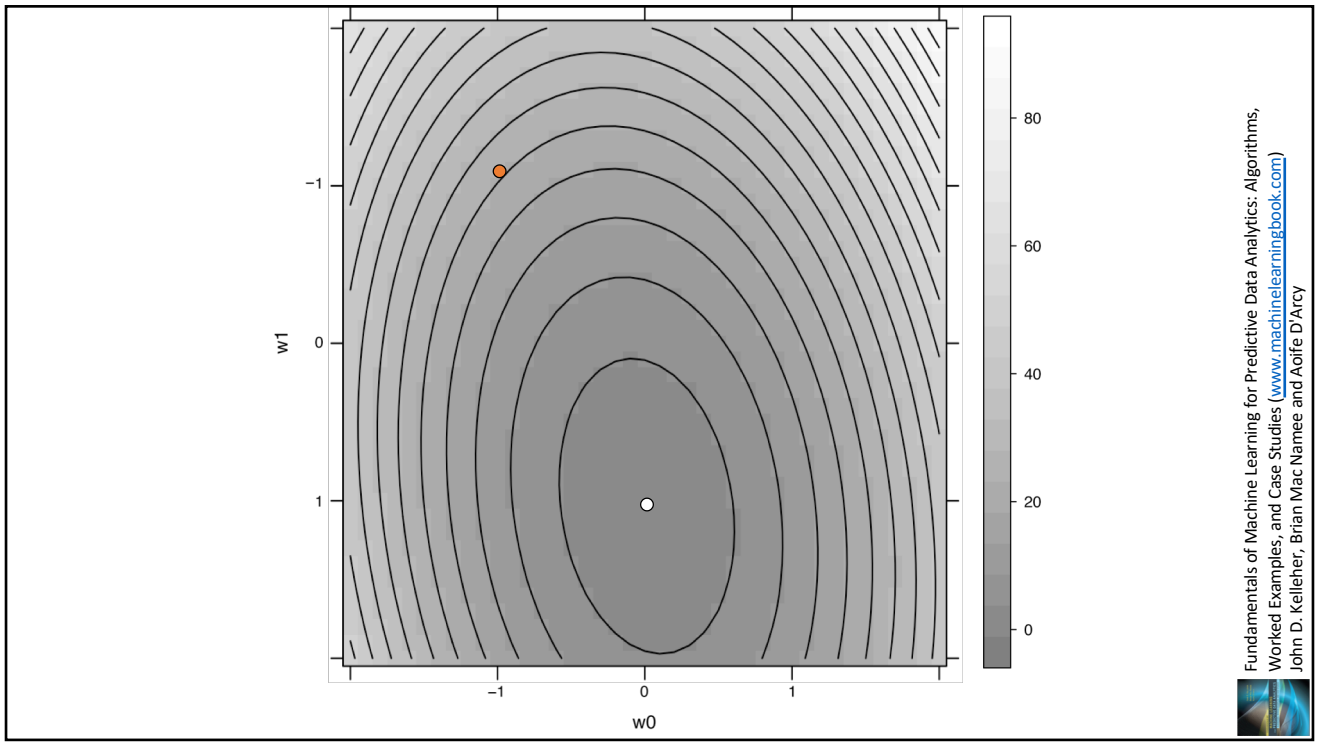


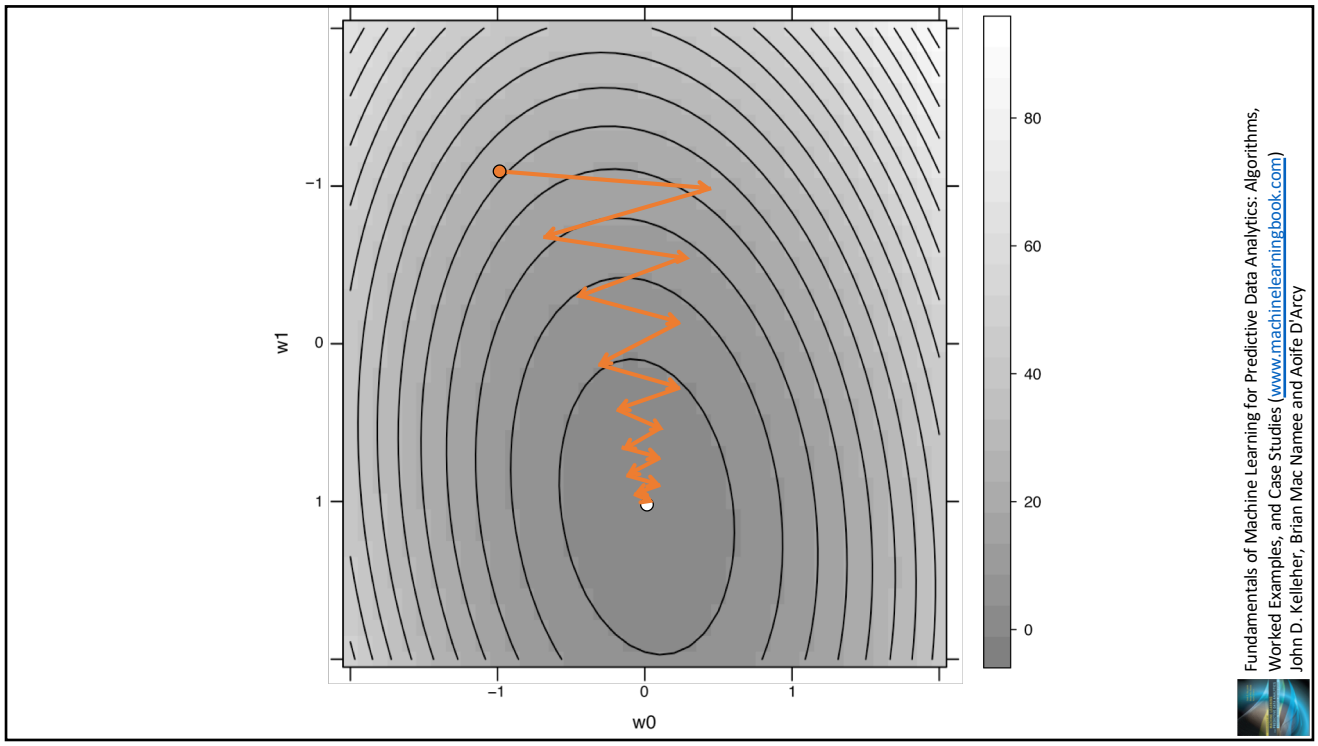


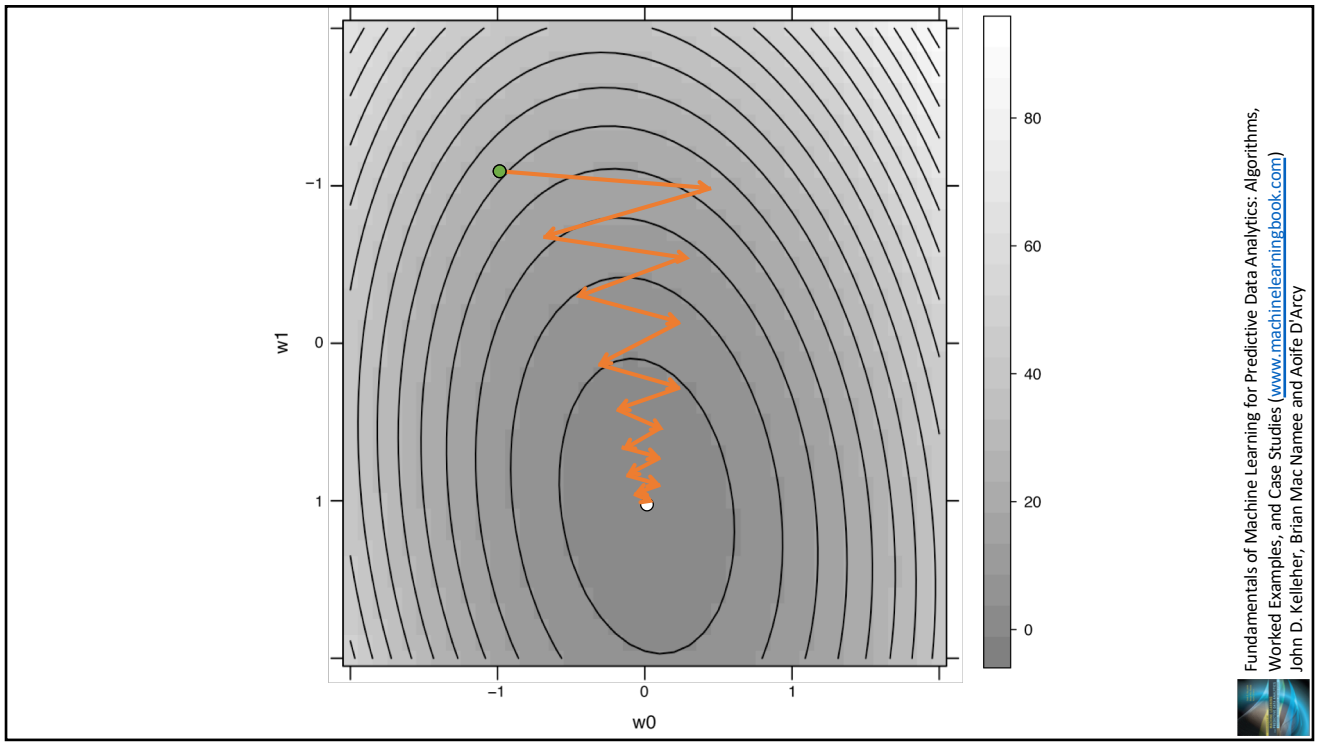

Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

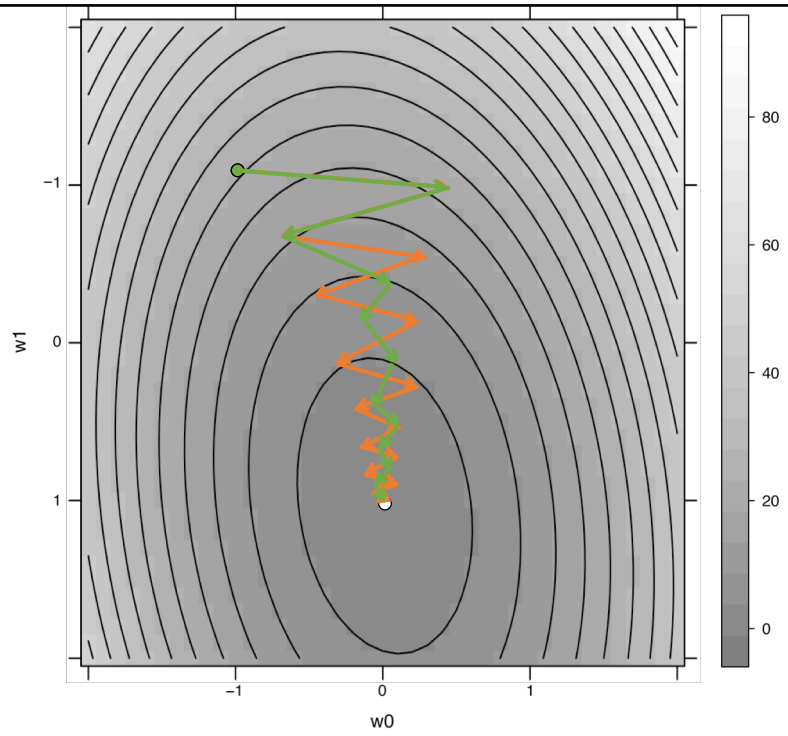
GRADIENT DESCENT WITH MOMENTUM


w1
w0
−1
0
1
80
60
40
20
0
Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy


w1
w0
−1
0
1
80
60
40
20
0
Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy


w1
w0
−1
0
1
80
60
40
20
0
Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy


w1
w0
−1
0
1
80
60
40
20
0
Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

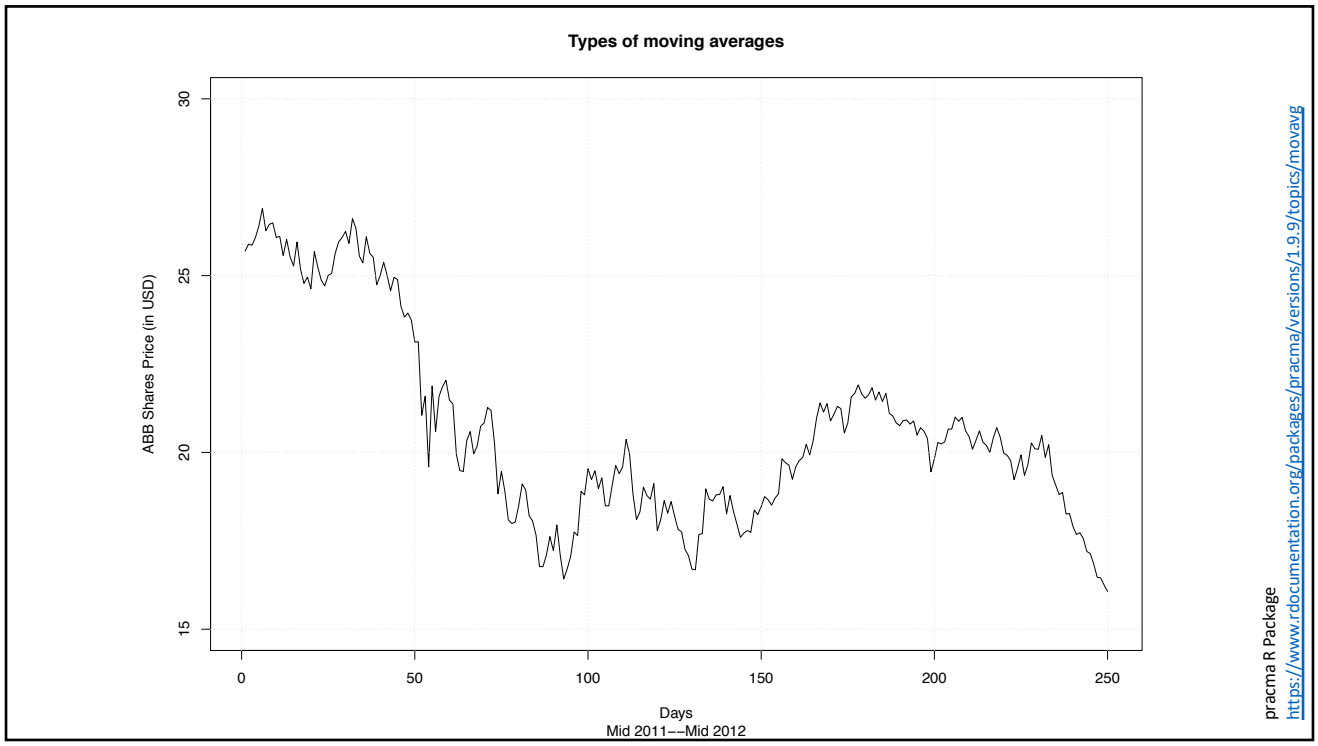


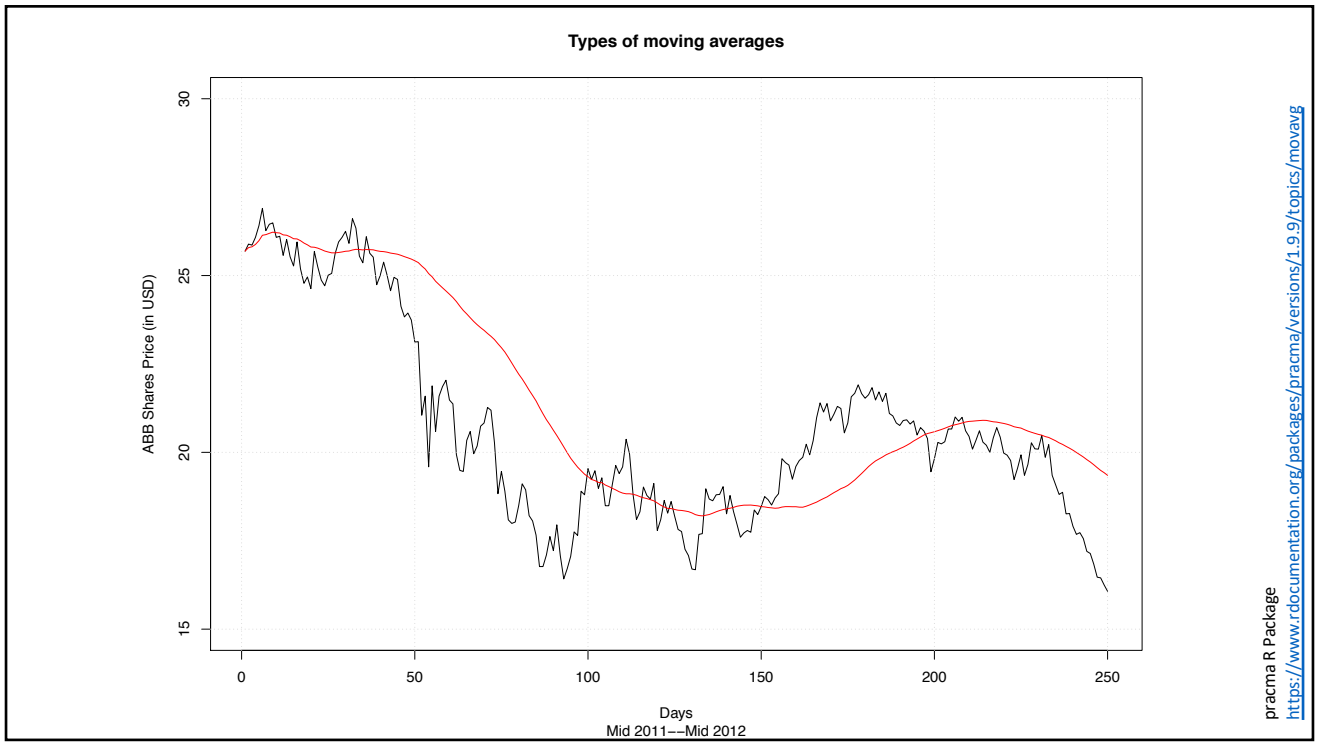
Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

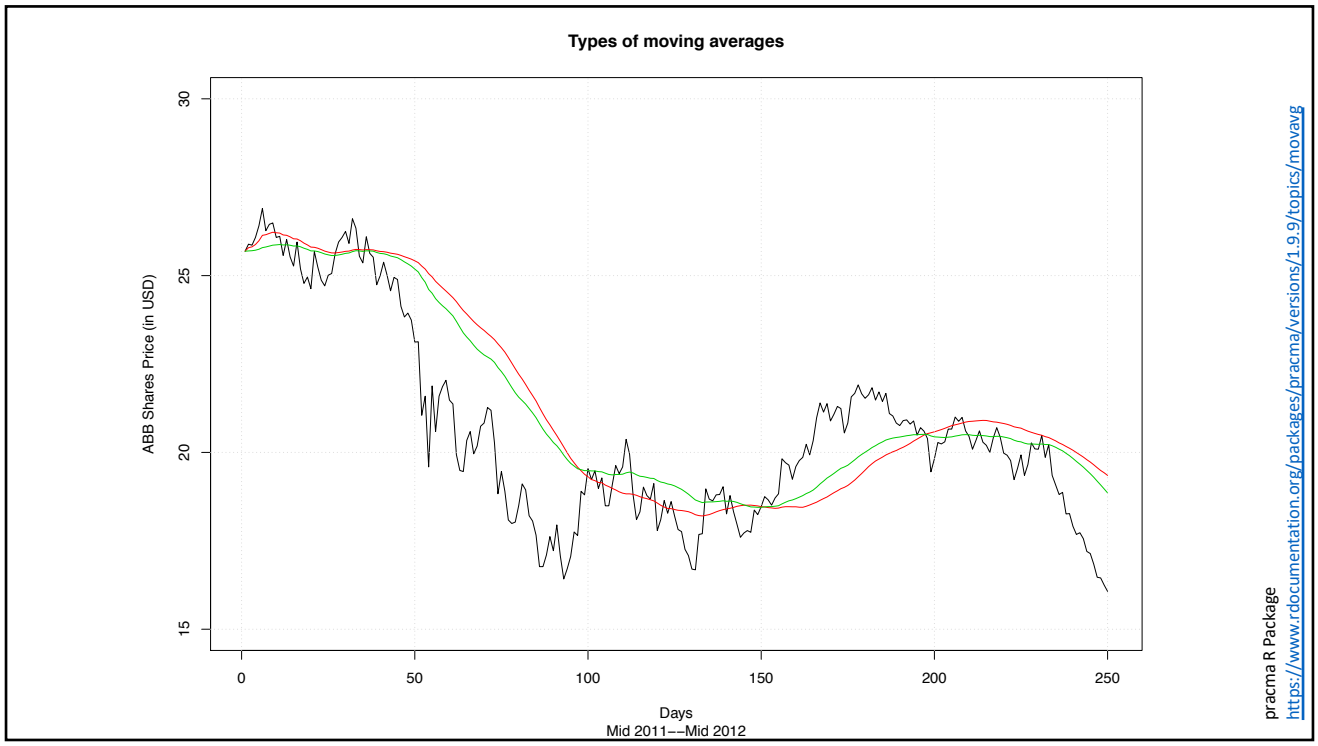
Exponentially Weighted Average

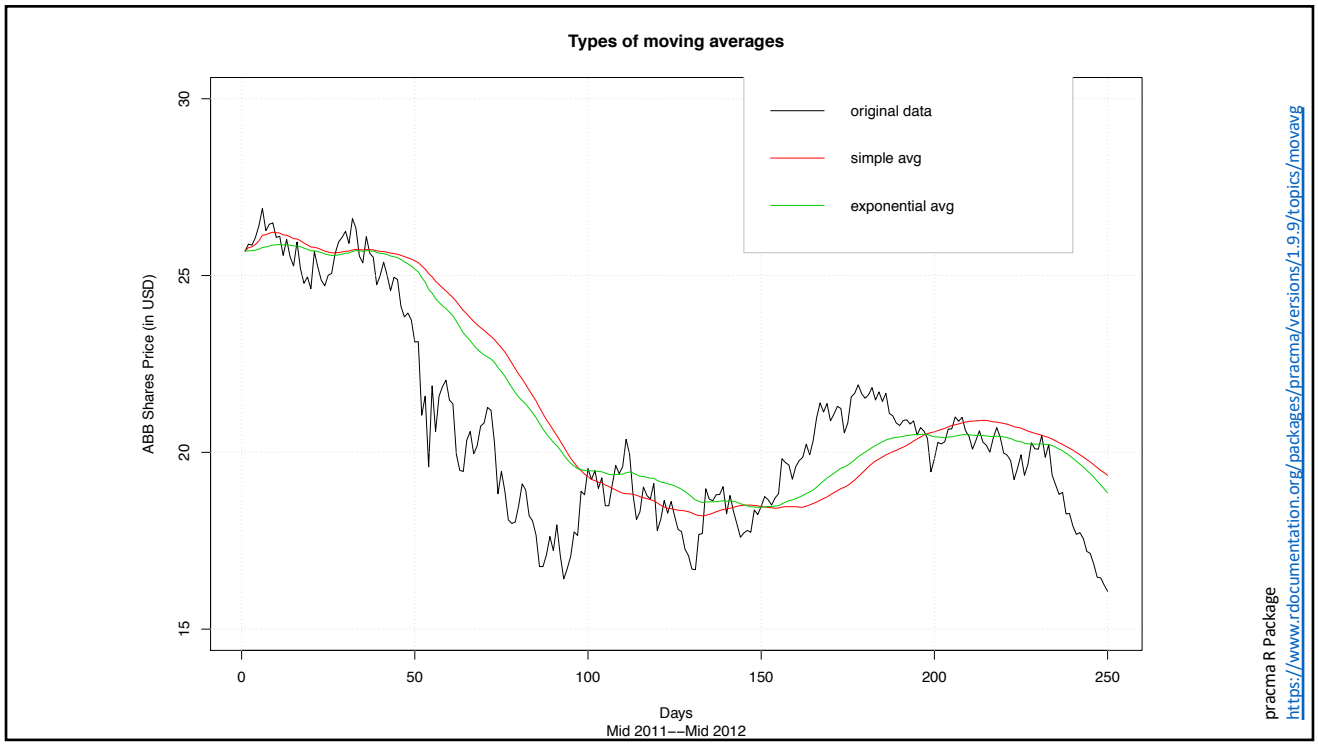
An exponentially weighted average is a simple way to smooth a signal

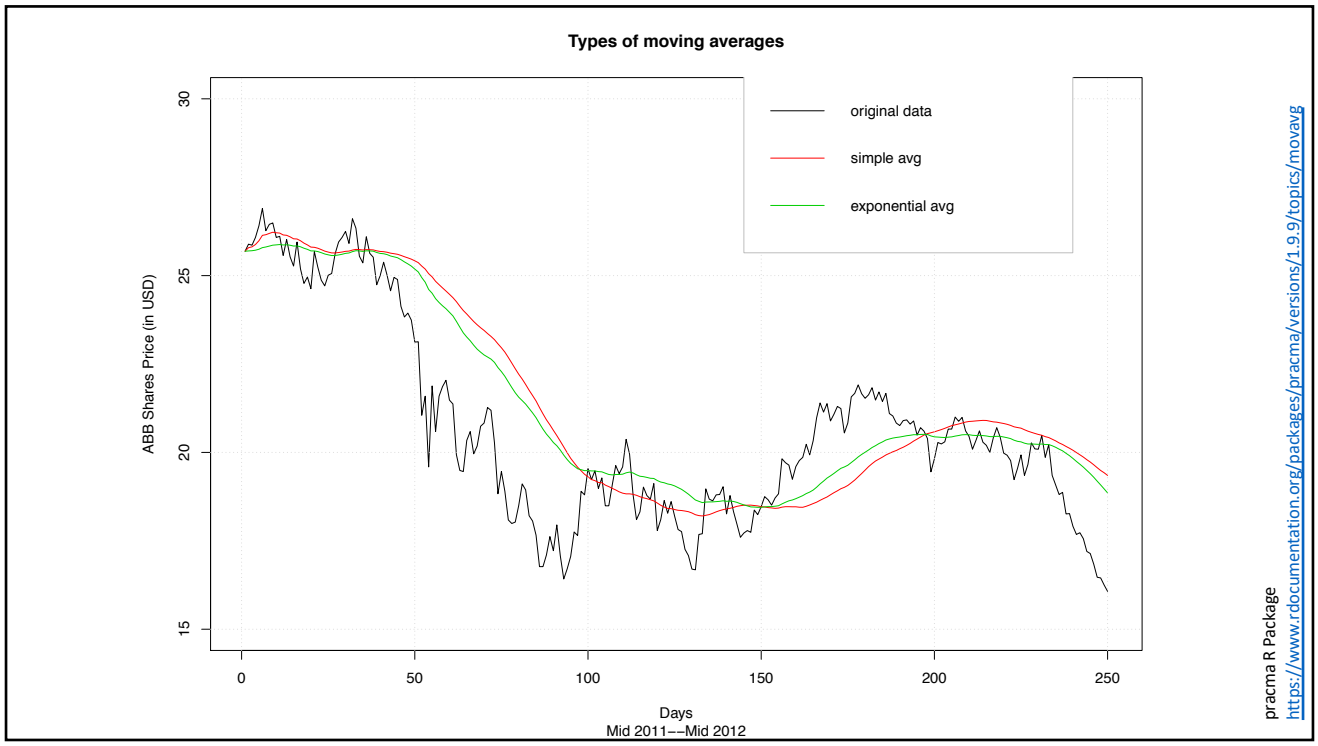
$$s = \beta\, s + (1 - \beta)\; x$$

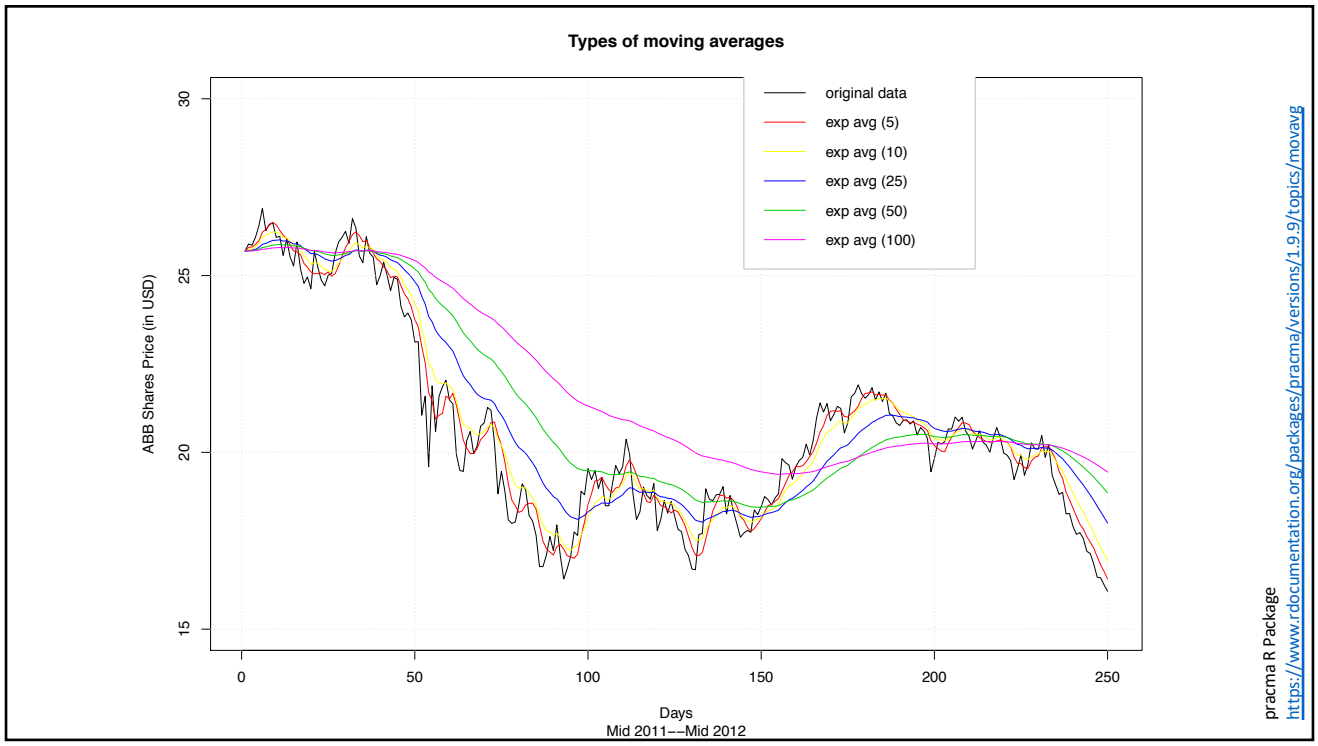

Types of moving averages
ABB Shares Price (in USD)
15
20
25
30
0
50
100
150
200
250
Days
Mid 2011––Mid 2012
pracma R Package
https://www.rdocumentation.org/packages/pracma/versions/1.9.9/topics/movavg


Types of moving averages
ABB Shares Price (in USD)
15
20
25
30
0
50
100
150
200
250
Days
Mid 2011––Mid 2012
pracma R Package
https://www.rdocumentation.org/packages/pracma/versions/1.9.9/topics/movavg


Types of moving averages
ABB Shares Price (in USD)
15
20
25
30
0
50
100
150
200
250
Days
Mid 2011––Mid 2012
pracma R Package
https://www.rdocumentation.org/packages/pracma/versions/1.9.9/topics/movavg


Types of moving averages
original data
simple avg
exponential avg
ABB Shares Price (in USD)
15
20
25
30
0
50
100
150
200
250
Days
Mid 2011––Mid 2012
pracma R Package
https://www.rdocumentation.org/packages/pracma/versions/1.9.9/topics/movavg


Types of moving averages
original data
simple avg
exponential avg
ABB Shares Price (in USD)
15
20
25
30
0
50
100
150
200
250
Days
Mid 2011––Mid 2012
pracma R Package
https://www.rdocumentation.org/packages/pracma/versions/1.9.9/topics/movavg


Types of moving averages
original data
exp avg (5)
exp avg (10)
exp avg (25)
exp avg (50)
exp avg (100)
ABB Shares Price (in USD)
15
20
25
30
0
50
100
150
200
250
Days
Mid 2011––Mid 2012
pracma R Package
https://www.rdocumentation.org/packages/pracma/versions/1.9.9/topics/movavg

Gradient Descent With Momentum

Apply an exponentially weighted moving average to gradient descent

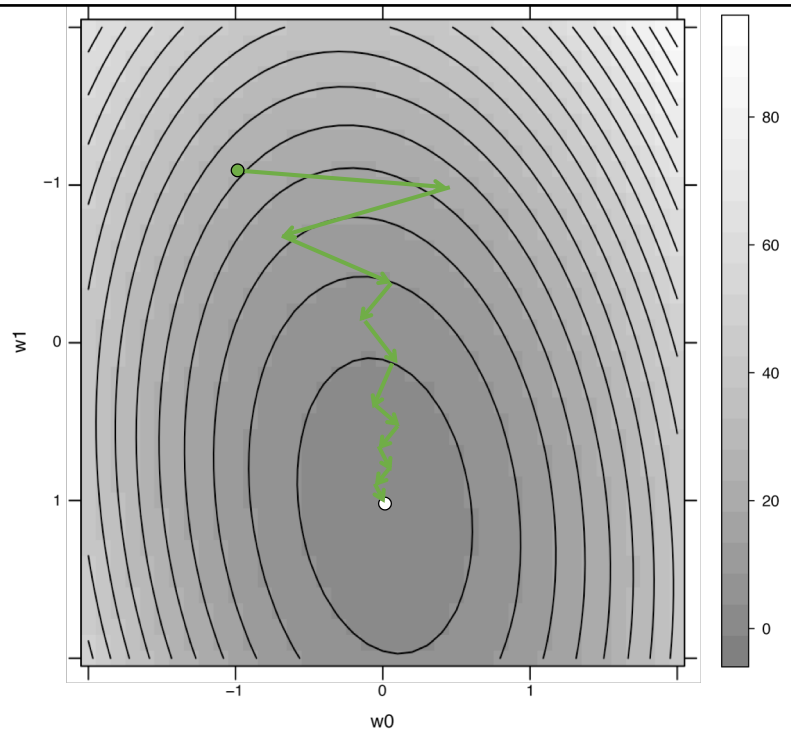
- Compute $d\mathbf{W}$ and $d\mathbf{b}$ as normal
- Exponentially weight gradients

$$v_{d\mathbf{W}} = \beta v_{d\mathbf{W}} + (1-\beta)\, d\mathbf{W}$$
$$v_{d\mathbf{b}} = \beta v_{d\mathbf{b}} + (1-\beta)\, d\mathbf{b}$$

Gradient Descent With Momentum

Weight update rule uses averaged gradients instead of raw gradients

$$\mathbf{W} = \mathbf{W} - \alpha v_{d\mathbf{W}}$$
$$\mathbf{b} = \mathbf{b} - \alpha v_{d\mathbf{b}}$$



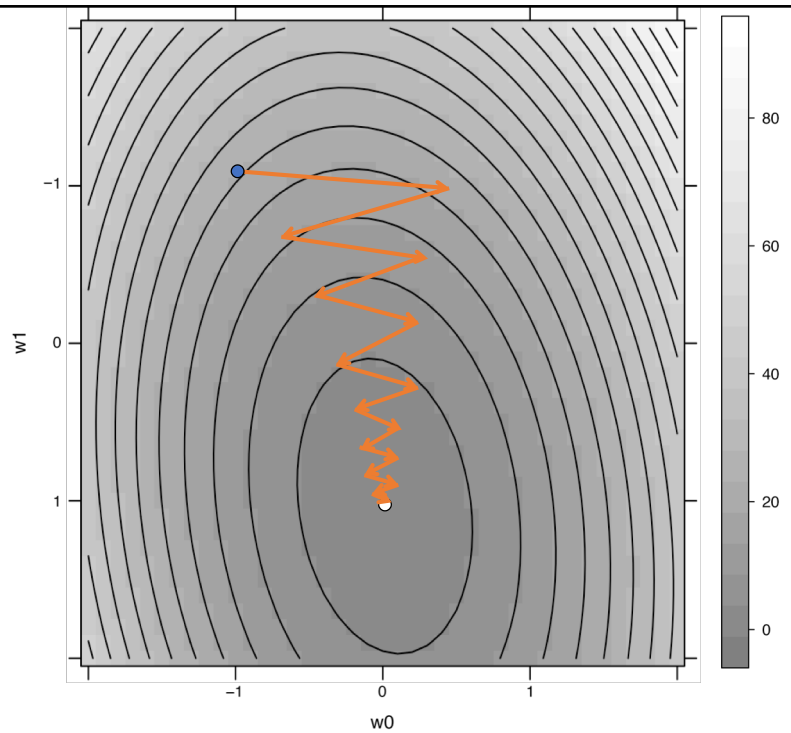
Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

Gradient Descent With Momentum

Decent values for hyper-parameters:

- α Learning rate, must be tuned
- β 0.9

RMSPROP



Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

RMSprop

Modifies gradients by scaling by exponential average of square of gradients

Modification term is calculated as follows

$$v_{d\mathbf{W}} = \beta\, v_{d\mathbf{W}} + (1-\beta)\, d\mathbf{W}^{*2}$$

$$v_{d\mathbf{b}} = \beta v_{d\mathbf{b}} + (1-\beta)\, d\mathbf{b}^{*2}$$

Neural Networks for Machine Learning, Lecture 6a, Overview of mini-batch gradient descent, Geoffrey Hinton et al http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf

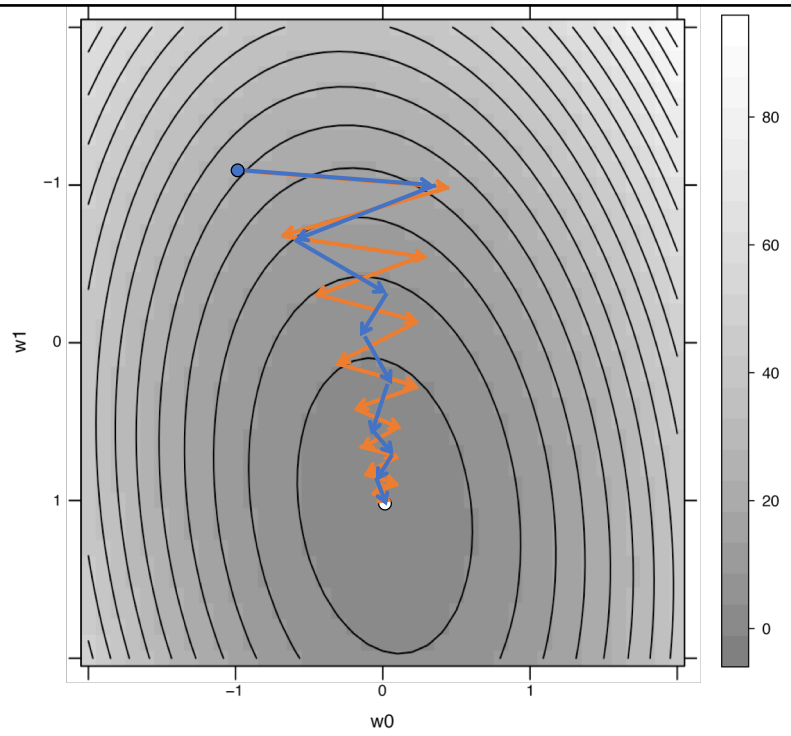
RMSprop

Weight update rule uses modified gradients

$$\mathbf{W} = \mathbf{W} - \alpha \frac{d\mathbf{W}}{\sqrt{v_{d\mathbf{W}}} + \epsilon}$$

$$\mathbf{b} = \mathbf{b} - \alpha \frac{d\mathbf{b}}{\sqrt{v_{d\mathbf{b}}} + \epsilon}$$

Neural Networks for Machine Learning, Lecture 6a, Overview of mini-batch gradient descent, Geoffrey Hinton et al http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf



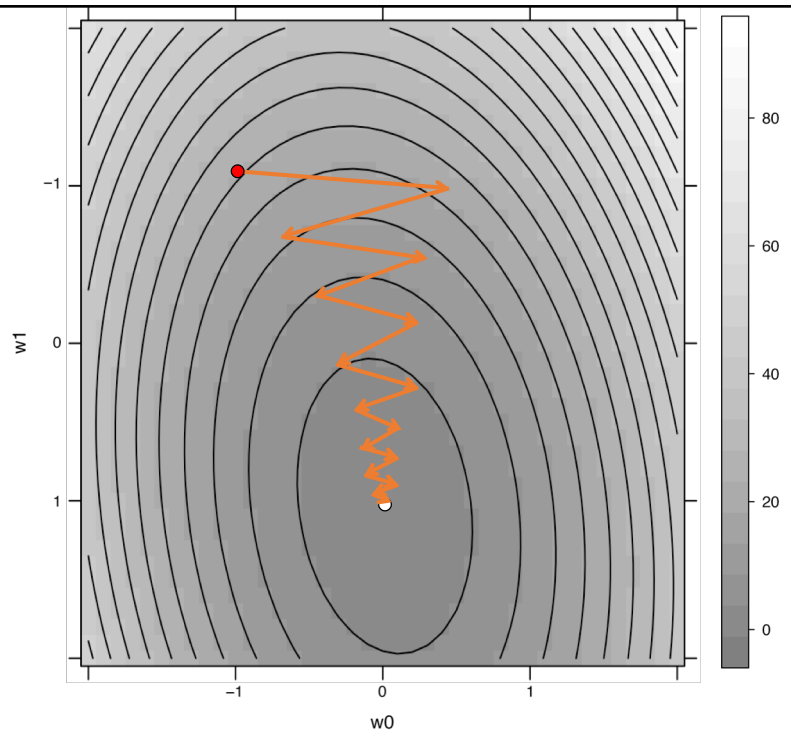
Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

RMSprop

Decent values for hyper-parameters:

- α Learning rate, must be tuned
- β 0.999
- ε 0.00000001

ADAM



Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

ADAM

Adaptive Moment Estimation method is a very popular optimiser and mixes aspects of Gradient Descent with Momentum and RMSprop

ADAM

Apply an exponentially weighted moving average to gradient descent

- Compute $d\mathbf{W}$ and $d\mathbf{b}$ as normal

$$v_{d\mathbf{W}} = \beta_1 v_{d\mathbf{W}} + (1-\beta_1)\, d\mathbf{W}$$

$$v_{d\mathbf{b}} = \beta_1 v_{d\mathbf{b}} + (1-\beta_1)\, d\mathbf{b}$$

$$s_{d\mathbf{W}} = \beta_2\, s_{d\mathbf{W}} + (1-\beta_2)\, d\mathbf{W}^{*2}$$

$$s_{d\mathbf{b}} = \beta_2 s_{d\mathbf{b}} + (1-\beta_2)\, d\mathbf{b}^{*2}$$

ADAM

Apply an exponentially weighted moving average to gradient descent

- Compute $d\mathbf{W}$ and $d\mathbf{b}$ as normal

$$v_{d\mathbf{W}} = \beta_1 v_{d\mathbf{W}} + (1-\beta_1)\, d\mathbf{W}$$
$$v_{d\mathbf{b}} = \beta_1 v_{d\mathbf{b}} + (1-\beta_1)\, d\mathbf{b}$$

Gradient Descent with Momentum

$$s_{d\mathbf{W}} = \beta_2\, s_{d\mathbf{W}} + (1-\beta_2)\, d\mathbf{W}^{*2}$$
$$s_{d\mathbf{b}} = \beta_2 s_{d\mathbf{b}} + (1-\beta_2)\, d\mathbf{b}^{*2}$$

RMSprop

ADAM

Weight update rule uses modified weight values

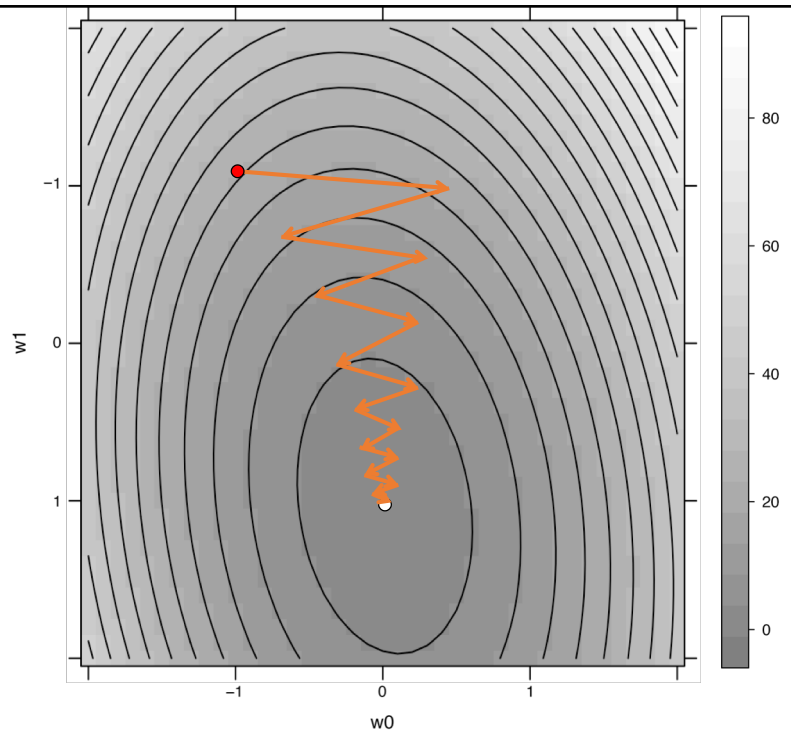
$$\mathbf{W} = \mathbf{W} - \alpha \frac{v_{d\mathbf{W}}}{\sqrt{s_{d\mathbf{W}}} + \epsilon}$$
$$\mathbf{b} = \mathbf{b} - \alpha \frac{v_{d\mathbf{b}}}{\sqrt{s_{d\mathbf{b}}} + \epsilon}$$

ADAM

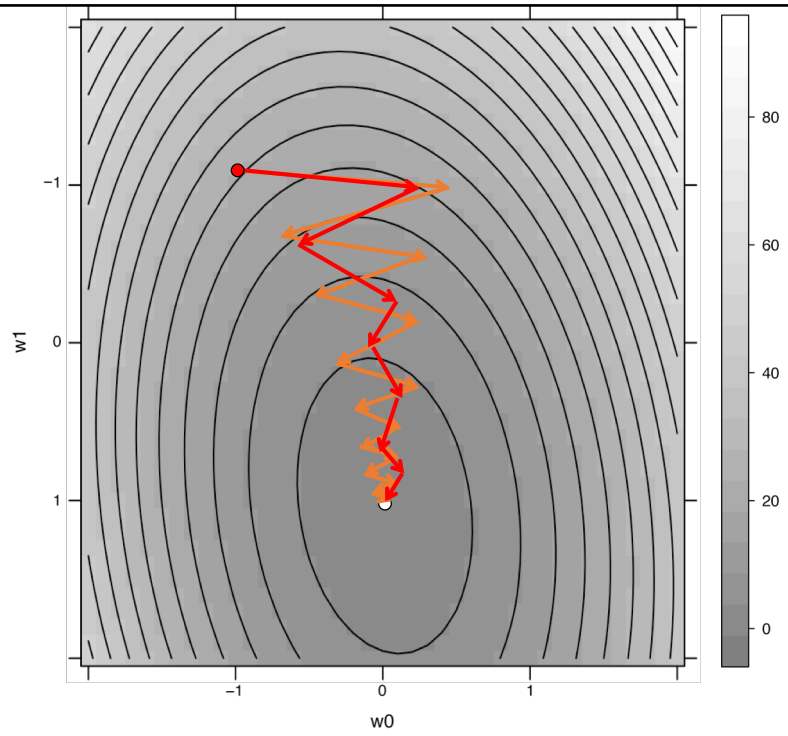
Some implementations apply bias correction - a exponential weighted averaging modification to adjust for the early points

$$v_{d\mathbf{W}} = \frac{v_{d\mathbf{W}}}{1-\beta_1^t} \qquad s_{d\mathbf{W}} = \frac{s_{d\mathbf{W}}}{1-\beta_2^t}$$

$$v_{d\mathbf{b}} = \frac{v_{d\mathbf{b}}}{1-\beta_1^t} \qquad s_{d\mathbf{b}} = \frac{s_{d\mathbf{b}}}{1-\beta_2^t}$$



Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy



Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

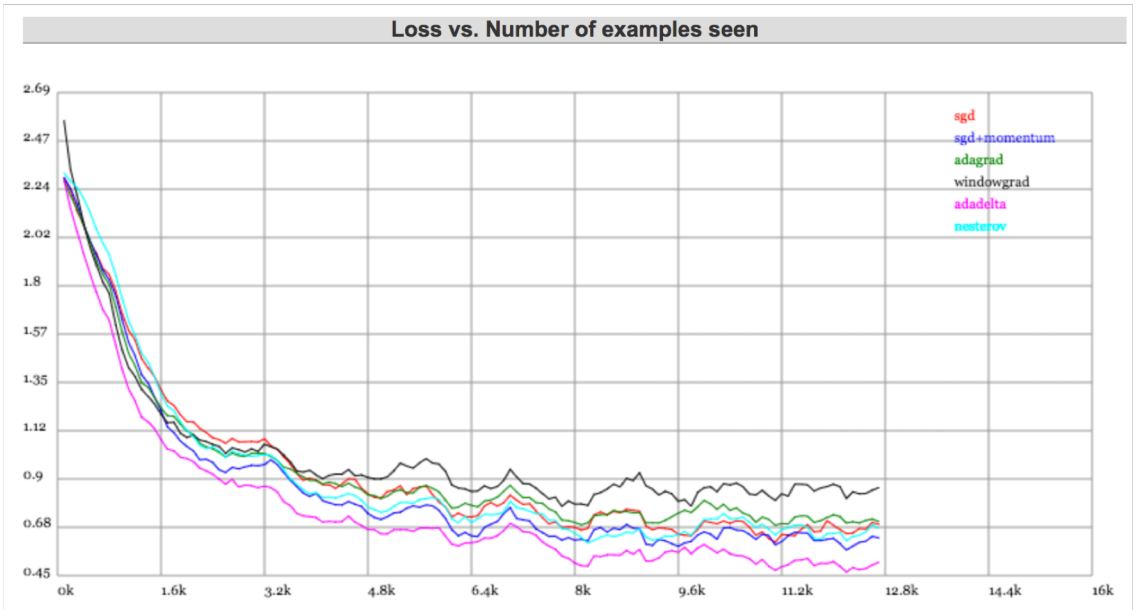
ADAM

Decent values for hyper-parameters:

α	Learning rate, must be tuned
β_1	0.9
β_2	0.999
ε	0.00000001

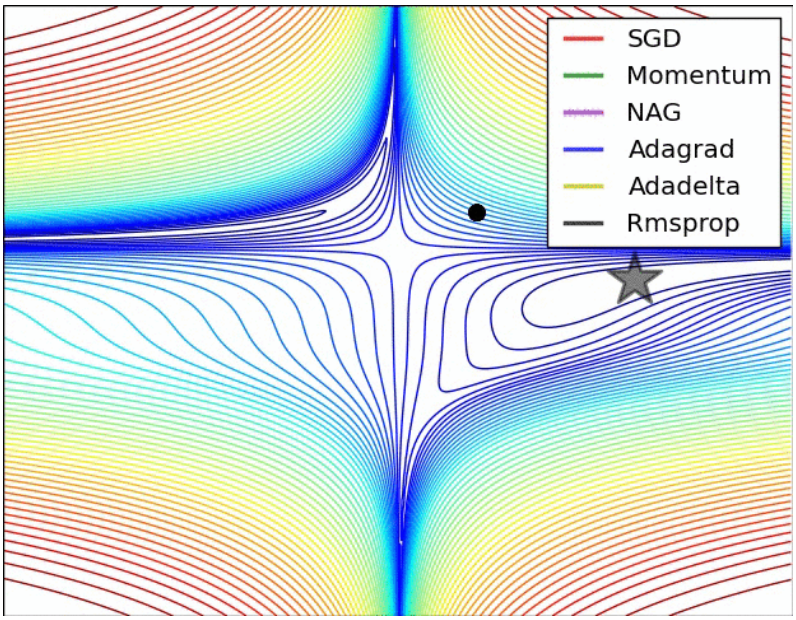
COMPARING OPTIMISERS

Nice Optimiser Demo



ConvNetJS Trainer demo on MNIST, Andrei Karpathy
https://cs.stanford.edu/people/karpathy/convnetjs/demo/trainers.html

Nice Optimiser Demo



Visualizing and Animating Optimization Algorithms with Matplotlib, Louis Tiao
http://tiao.io/notes/visualizing-and-animating-optimization-algorithms-with-matplotlib/

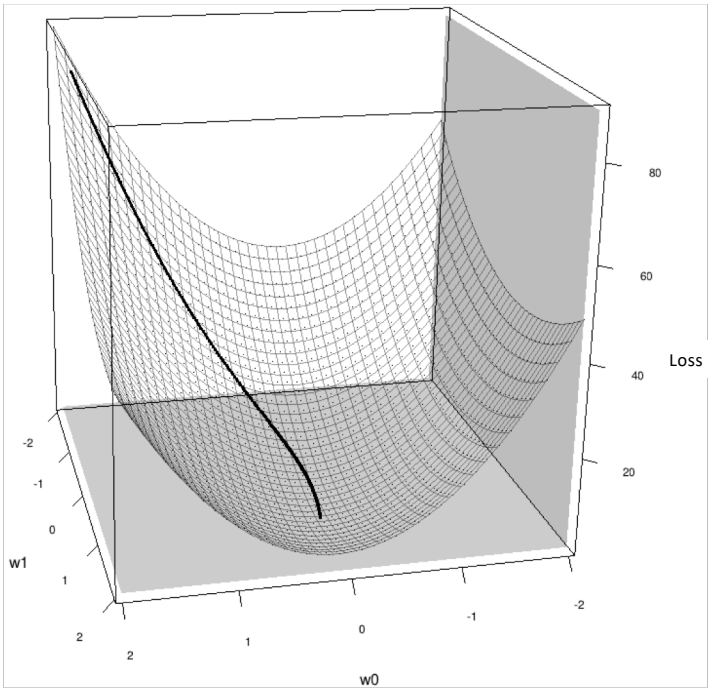
Comparing Optimisers

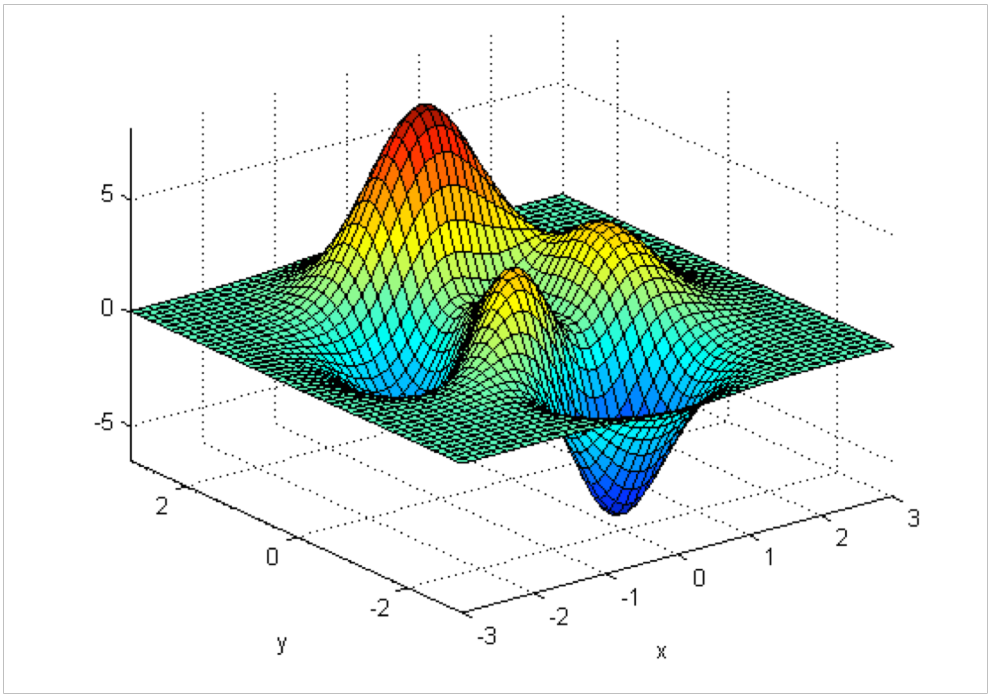
Comparing optimisers is a little like a hyper-hyper-parameter selection problem!

In general most of the optimisers will get you to the same place eventually, it just might take longer

Studies have shown that **adam** is generally a good choice

A WORD ON ERROR SURFACES

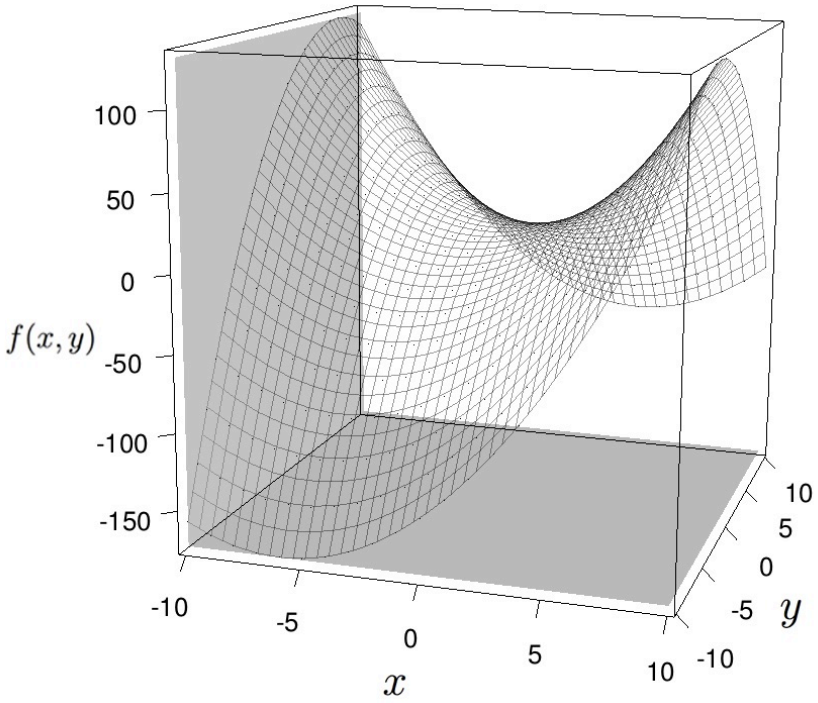




Local Optima

Local optima are actually really unlikely in high dimensional networks

- Saddle points are much more common, but not a particular problem not a problem

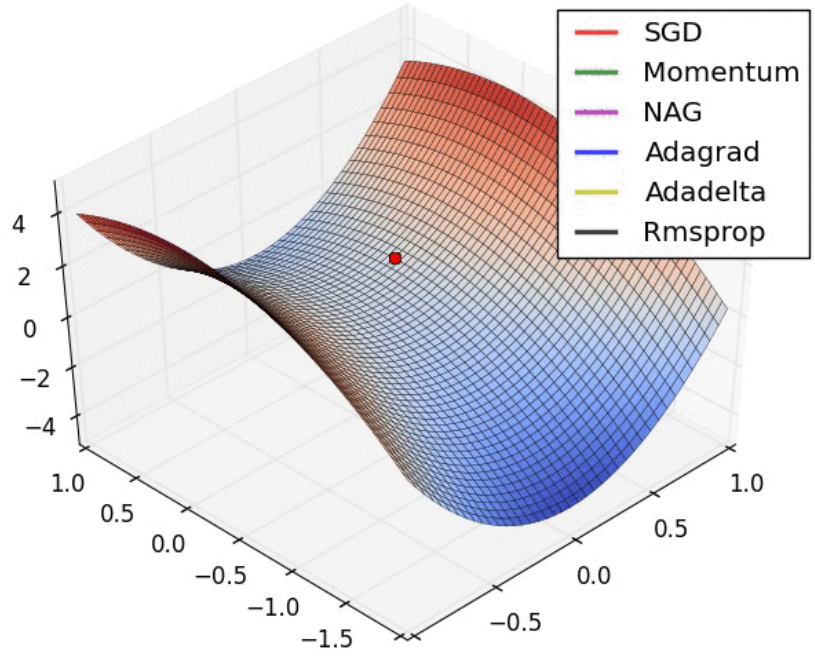


Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

Plateaus

Plateaus are a bigger problem than local minima

- Low gradient really slows down progress
- Algorithms like RMSprop, adam etc help with this



Visualizing and Animating Optimization Algorithms with Matplotlib, Louis Tiao
http://tiao.io/notes/visualizing-and-animating-optimization-algorithms-with-matplotlib/

SUMMARY

Summary

There are lots of modifications to the basic gradient descent algorithm that we can introduce to improve the learning process and make it more likely that we arrive at accurate models quickly

Optimisation algorithms are a useful lever to turn here

- If you don't have a better idea use **adam** with default parameters

Summary

There is no one answer, however, and new approaches are always being suggested

We can't avoid experimentation, however

Questions