The most common optimization algorithms

1.Gradient Descent

how much data we use to compute the gradient

accuracy **time**

1.1BGD

the entire training d taset

a local minimum for non-convex surfaces

very slow **do not fit in memory** **update our model online**

1.2SGD

(x(i),y(i))

new and potentially better local minima

faster can be used to learn online.

1.3 Mini-batch gradient descent

Challenges

2.Momentum(动量)

helps accelerate SGD

3. NAG(牛顿动量)

4.Adagrad

main weakness

5.Adadelta

restricts the wind
w of accumulated past gradients to some fixed size w.

RMSprop

Adam

How to Choose

input data is sparse **the adaptiv**
learning-rate methods

RMSprop **Adagrad** **its radically diminishing l**

arning rates

Adam adds bias-correction and momentum to R

Sprop

Adam

Insofar, Adam might be the best over

ll choice.

SGD usually achieves to find a minimum

much more reliant on

robust initialization and annealing schedule

fast convergence and train a deep or complex neural net

ork adaptive learning rate methods