Stochastic Gradient Descent from Convex Optimization Perspective

compiled by D.Gueorguiev, 3/29/2025

# Gradient Descent

Consider unconstrained smooth convex optimization

That is, is convex and differentiable with . Denote the optimal criterion value by and a solution by .

Gradient descent algorithm

choose initial point

repeat

stop when certain stopping criterion is satisfied

A grid with colored lines

Description automatically generated

A grid with colored lines and dots

Description automatically generated

At each iteration, consider the expansion

Quadratic approximation, replacing usual Hessian by

linear approximation to

proximity term to , with weight

Choose next point to minimize quadratic approximation:

A line with dots and a dotted line

Description automatically generated

Blue point is , red point is . The latter is given with

Question: How do we choose step sizes?

Simply taking for all can diverge if is too big.

A graph of a function

Description automatically generatedConsider , gradient descent after 8 steps:

Can be slow if is too small. Same example, gradient descent after 100 steps:

A graph of a graph of a stringed object

Description automatically generated with medium confidence

Convergence analysis later will give us a precise idea of “just right”

## Backtracking Line Search

One way to adaptively choose the step size is to use backtracking line search:

1 ) First fix parameters and

2 ) At each iteration, start with ,

while iter\_count < max\_iter\_count:

if :

shrink as

else:

perform gradient descent update

Simple and tends to work well in practice (further simplification: take )

A graph of a function

Description automatically generatedBacktracking interpretation

Here

Setting , backtracking picks up roughly the right step size (12 outer steps, 40 steps total).

A diagram of a dotted line

Description automatically generated with medium confidence

## Exact Line Search

We could also choose step to do the best we can along direction of negative gradient, called exact line search:

//TODO: finish the Gradient Descent section

# The Sub-Gradient Method

//TODO: finish the Sub-Gradient Method

# Proximal Gradient Descent

//TODO: finish the Proximal Gradient Descent

# Stochastic Gradient Descent

//TODO: finish the Stochastic Gradient Descent

# References

[1] [Gradient Descent, Convex Optimization 10-725, CMU, slides, Ryan Tibshirani](https://github.com/dimitarpg13/statistical_learning_and_kernel_methods/blob/main/literature/articles/gradient_descent/grad-descent_Ryan_Tibshirani_slides.pdf)

[2] [Sub-Gradient Method, Convex Optimization 10-725, CMU, slides, Ryan Tibshirani](https://github.com/dimitarpg13/statistical_learning_and_kernel_methods/blob/main/literature/articles/gradient_descent/sub-gradient-method-Ryan_Tibshirani_slides.pdf)

[3] [Proximal Gradient Descent, Convex Optimization 10-725, CMU, slides, Ryan Tibshirani](https://github.com/dimitarpg13/statistical_learning_and_kernel_methods/blob/main/literature/articles/gradient_descent/proximal-gradient-descent_Ryan_Tibshirani_slides.pdf)

[4] [Stochastic Gradient Descent, Convex Optimization 10-725, CMU, slides, Ryan Tibshirani](https://github.com/dimitarpg13/statistical_learning_and_kernel_methods/blob/main/literature/articles/gradient_descent/stochastic-gradient-descent_Ryan_Tibshirani_slides.pdf)