Machine Learning

Optimizers

Curtis Larsen

Utah Tech University—Computing

Spring 2025

Objectives

Outline

Objectives:

- Understand convergence
- Understand and implement optimizers

Convergence

Idea

- Gradient descent
- When does it stop?
- How long until it stops?
- Controlled through gradient descent and learning rate

Math

Gradient descent step.

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \boldsymbol{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \tag{1}$$

Momentum

Ideas

- Track sum of gradient values
- Use sum to update parameters
- Gradient is an acceleration now, instead of a speed

Math

$$\boldsymbol{m} \leftarrow \beta \boldsymbol{m} - \eta \boldsymbol{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \tag{2}$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \boldsymbol{m} \tag{3}$$

$$0 \leq \beta \leq 1$$

Implementation

```
keras.optimizers.SGD(momentum=0.9)
```

Nesterov

Ideas

- Look into future for gradient

Math

$$\boldsymbol{m} \leftarrow \beta \boldsymbol{m} - \eta \boldsymbol{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta} + \beta \boldsymbol{m}) \tag{4}$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \boldsymbol{m} \tag{5}$$

$$0 \leq \beta \leq 1$$

Implementation

```
keras.optimizers.SGD(momentum=0.9,
      nesterov=True)
```

AdaGrad

Ideas

- Scale updates in each direction based on historical gradient size.

Math

$$\boldsymbol{s} \leftarrow \boldsymbol{s} + \boldsymbol{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \otimes \boldsymbol{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \tag{6}$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \boldsymbol{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \div \sqrt{\boldsymbol{s} + \epsilon} \tag{7}$$

Implementation

```
keras.optimizers.Adagrad()
```

RMSProp

Ideas

- Scale updates in each direction based on recent history.

Math

$$\boldsymbol{s} \leftarrow \rho \boldsymbol{s} + (1 - \rho) \boldsymbol{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \otimes \boldsymbol{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \tag{8}$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \boldsymbol{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \div \sqrt{\boldsymbol{s} + \epsilon} \tag{9}$$

Implementation

```
keras.optimizers.RMSprop(rho=0.9)
```

Adam

Ideas

- Track mean historical gradient (first moment)
- Track variance in historical gradient (second moment)
- Scale steps based on them

Math

$$\boldsymbol{m} \leftarrow \beta_1 \boldsymbol{m} - (1 - \beta_1) \boldsymbol{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \tag{10}$$

$$\boldsymbol{s} \leftarrow \beta_2 \boldsymbol{s} + (1 - \beta_2) \boldsymbol{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \otimes \boldsymbol{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \tag{11}$$

$$\hat{\boldsymbol{m}} \leftarrow \frac{\boldsymbol{m}}{1 - \beta_1^t} \tag{12}$$

$$\hat{\boldsymbol{s}} \leftarrow \frac{\boldsymbol{s}}{1 - \beta_2^t} \tag{13}$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \eta \hat{\boldsymbol{m}} \div \sqrt{\hat{\boldsymbol{s}} + \epsilon} \tag{14}$$

Implementation

```
keras.optimizers.Adam(beta_1=0.9,
      beta_2=0.999)
```

Summary

Topics

- Momentum
- RMSProp
- Adam