Course > Unit 1 Linear Classifiers and Generalizations (2 weeks) > Lecture 4. Linear Classification and Generalization > 5. Stochastic Gradient Descent

**Audit Access Expires May 11, 2020**
You lose all access to this course, including your progress, on May 11, 2020.
Upgrade by Mar 25, 2020 to get unlimited access to the course as long as it exists on the site. **Upgrade now**

# 5. Stochastic Gradient Descent

## Stochastic Gradient Descent

0:00 / 0:00

1.50x

### Video

Download video file

### Transcripts

**Download SubRip (.srt) file**

**Download Text (.txt) file**

## SGD and Hinge Loss

1/1 point (graded)

As we saw in the lecture above,

$$J(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n} \text{Loss}_h\left(y^{(i)}\left(\theta \cdot x^{(i)} + \theta_0\right)\right) + \frac{\lambda}{2} \mid\mid \theta \mid\mid^2 = \frac{1}{n}\sum_{i=1}^{n} \left[\text{Loss}_h\left(y^{(i)}\left(\theta \cdot x^{(i)} + \theta_0\right)\right) + \frac{\lambda}{2} \mid\mid \theta \mid\mid^2\right]$$

With stochastic gradient descent, we choose $i \in \{1, \ldots, n\}$ at random and update $\theta$ such that

$$\theta \leftarrow \theta - \eta \nabla_\theta \left[ \text{Loss}_h \left( y^{(i)} \left( \theta \cdot x^{(i)} + \theta_0 \right) \right) + \frac{\lambda}{2} \mid\mid \theta \mid\mid^2 \right]$$

What is $\nabla_\theta \left[ \text{Loss}_h \left( y^{(i)} \left( \theta \cdot x^{(i)} + \theta_0 \right) \right) \right]$ if $\text{Loss}_h \left( y^{(i)} \left( \theta \cdot x^{(i)} + \theta_0 \right) \right) > 0$?

○ $y^{(i)} x^{(i)}$

● $-y^{(i)} x^{(i)}$

○ 0

○ $\lambda\theta$

○ $-\lambda\theta$

✓

**Solution:**

If $\text{Loss}_h \left( y^{(i)} \left( \theta \cdot x^{(i)} + \theta_0 \right) \right) > 0$,

$$\text{Loss}_h \left( y^{(i)} \left( \theta \cdot x^{(i)} + \theta_0 \right) \right) = 1 - y^{(i)} \left( \theta \cdot x^{(i)} + \theta_0 \right)$$

. Thus

$$\nabla_\theta \text{Loss}_h \left( y^{(i)} \left( \theta \cdot x^{(i)} + \theta_0 \right) \right) = -y^{(i)} x^{(i)}$$

.

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

## Comparison with Perceptron

1/1 point (graded)

Observing the update step of SGD,

$$\theta \leftarrow \theta - \eta \nabla_\theta \left[ \text{Loss}_h \left( y^{(i)} \left( \theta \cdot x^{(i)} + \theta_0 \right) \right) + \frac{\lambda}{2} \mid\mid \theta \mid\mid^2 \right]$$

Which of the following is true?

○ As in perceptron, $\theta$ is not updated when there is no mistake

● Differently from perceptron, $\theta$ is updated even when there is no mistake

✓

**Solution:**

We can see from

$$\theta \leftarrow \begin{cases} (1-\lambda\eta)\,\theta \text{ if Loss=0} \\ (1-\lambda\eta)\,\theta + \eta y^{(i)} x^{(i)} \text{ if Loss>0} \end{cases}$$

that $\theta$ is updated even when the sum of losses is $0$. This is different from perceptron.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

## Discussion

Hide Discussion

**Topic:** Unit 1 Linear Classifiers and Generalizations (2 weeks):Lecture 4. Linear Classification and Generalization / 5. Stochastic Gradient Descent

Add a Post

Show all posts by recent activity

Not clear the gradient of the function 2
Hello staff. I was wondering if you could point to us more material to understand the function from which the gradient was derived, i have ...

? The update of theta_0 would be just -y_i? 4

? Regularization Term Increase Margin 2

? random sampling 4
What is the idea behind random sampling? The outcome is better? We sample only a fraction of the set?

? How is $\theta_0$ updated? 2

☑ Mistake in the video? 8
Community TA

☑ Why do we use a decreasing learning rate. ? 11
Hi everyone, Can someone please explain clearly what professor means when he said we " use a decreasing learning rate due to stochastic...

? What is the gradient of the Regularization component of objective formula

In the example we take the gradient of the loss function only, what about the regularization piece. Is it simply Lambda*Theta ? Just intrigue... 2

Learn About Verified Certificates