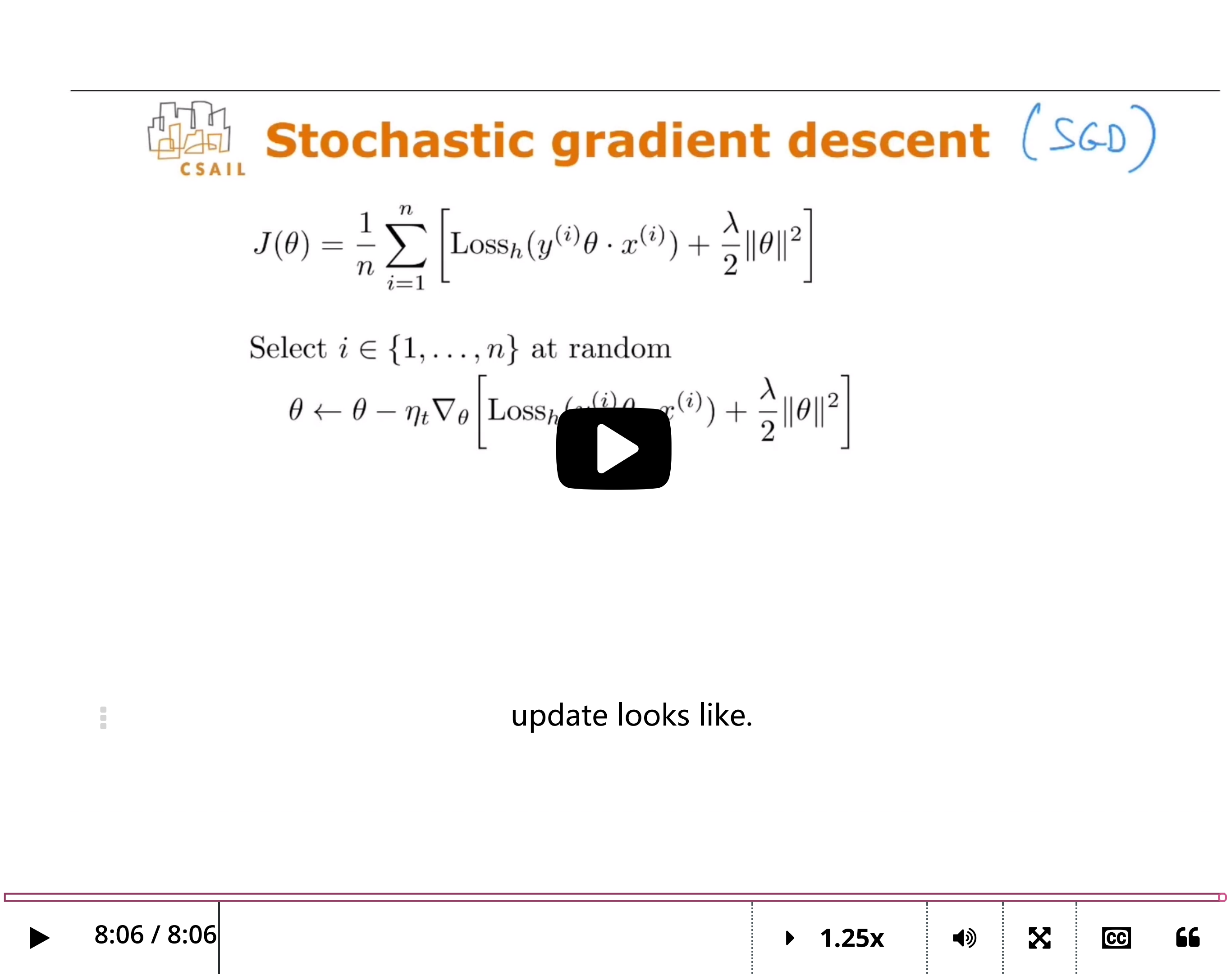
5. Stochastic Gradient Descent

Stochastic Gradient Descent



And that update looks like the perceptor update,

but it is actually made even if we correctly

classify the example.

If the example is within the margin boundaries,

you would get a non-zero loss.

So here, we have just a better way

of writing what that stochastic gradient descent update or SGD

update looks like.

End of transcript. Skip to the start.

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,

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

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

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

Which of the following is true?

- As in perceptron, θ is not updated when there is no mistake
- Differently from perceptron, θ 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 θ 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

Show Discussion

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