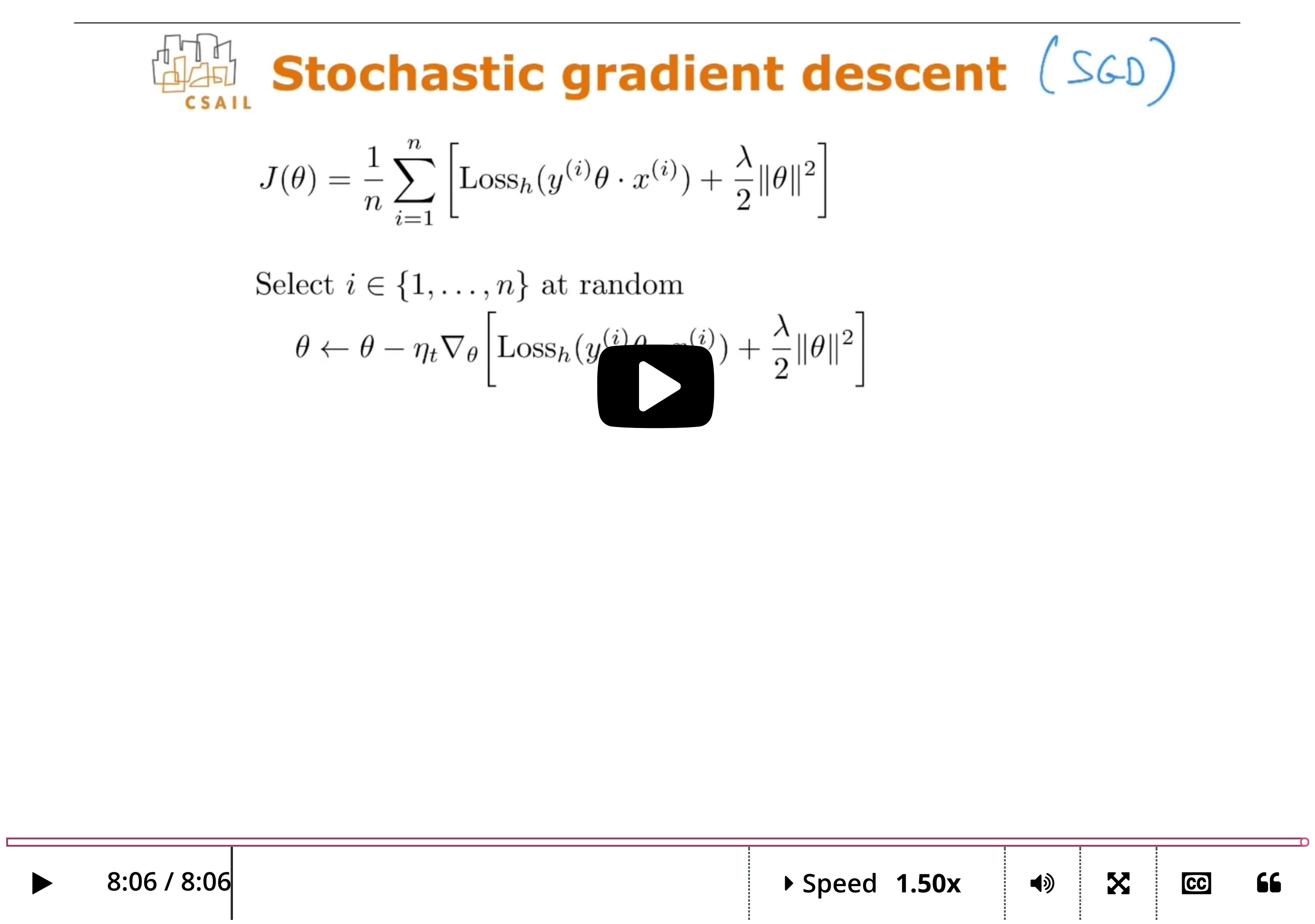
5. Stochastic Gradient Descent

Stochastic Gradient Descent



derivative

from the last terms, if the loss is non-zero.

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 $i\in\{1,\dots,n\}$ 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,

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

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

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 ▼

- [Staff] Formula in SGD and Hinge Loss — 3
- Constraints on Learning Parameter η (eta) — 2
  I am unable to understand the reasons and requirements behind constraints mentioned by Professor on the learning rate parameter η (starting at 3:18 in the video). What is ...
- Missing norm(...) in regularization term at 6:37? — 5
  Shouldn't there be the norm of theta instead of just theta when we took the derivative and outlined the case when the loss is not zero?
- How to check if the shape of objective function — 5
  Of perception, we want to minimize objective, but how to know the shape of it ? or all objective functions are assumed to have convex or concave shape ?
- Updating offset parameter — 4
- regularzation term when loss = 0 — 5
  In the lecture the prof. only note the regularzation term when loss >0, should we also include this term when loss = 0?
- difference between gradient descent and SGD — 3
  I am a bit confused with this lecture, can I say the difference between both of them is SGD converge faster than GD? because the step size of SGD is a function with time (e.g. ...
- Doubt about gradient of regularization parameter term. — 2
- Three differences with Perceptron — 7
  The professor mentioned three differences with perceptron. I could see two. Can anyone help me understand what the third difference is?
- x is missing superscript in several places on this page — 2
  .
- Comparison with Perceptron Confusion — 2
  [STAFF] I'm a bit confused on the Comparison with Perceptron question. It seems at odd with the lecture, and the updating algo in the solution is not what Prof wrote in his sli...

Learn About Verified Certificates