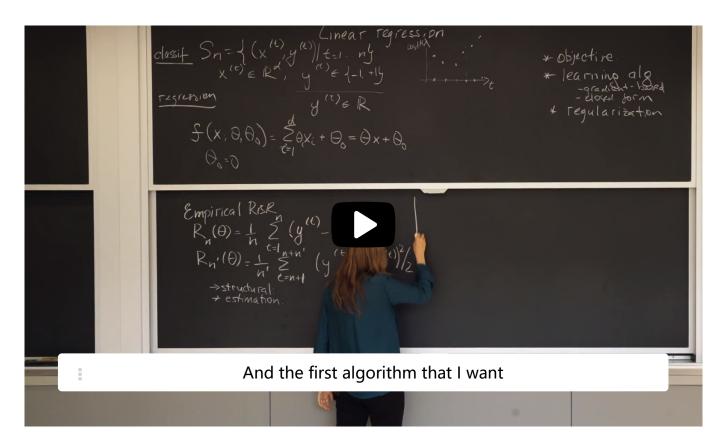


<u>Unit 2 Nonlinear Classification,</u> <u>Linear regression, Collaborative</u>

Course > Filtering (2 weeks)

> <u>Lecture 5. Linear Regression</u> > 5. Gradient Based Approach

5. Gradient Based Approach Learning Algorithm: Gradient Based Approach



o:02 / 7:03

Download video file

Video

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

▶ 1.0x 🔹 🛣 🚾

And the first algorithm that I want

to demonstrate-- so we'll just close where we are,

our objective.

We are done, and now we are talking about finding algorithm.

And as you can see here, we will start first with the gradient-based approach.

So the good news about this function, that it's actually differentiable everywhere, correct?

xuetangX.com 学堂在线

True or False

1/1 point (graded)

Let $R_{n}\left(heta
ight)$ be the least squares criterion defined by

$$R_{n}\left(heta
ight) = rac{1}{n}\sum_{t=1}^{n}\mathrm{Loss}\left(y^{(t)} - heta \cdot x^{(t)}
ight).$$

Which of the following is true? Choose all those apply.

- lacktriangledown The least squares criterion $R_n\left(heta
 ight)$ is a sum of functions, one per data point. lacktriangledown
- Stochastic gradient descent is slower than gradient descent.
- $m{artheta} \;
 abla_{ heta} R_n \left(heta
 ight)$ is a sum of functions, one per data point. $m{arphi}$



Solution:

For every point, the loss is a function of θ , so the least squares criterion $R_n\left(\theta\right)$ is a sum of functions, one per data point, and this is what makes stochastic gradient descent possible. We want to do stochastic gradient descent because it is faster than gradient descent. Finally, because $R_n\left(\theta\right)$ is sum of functions, one per data point, $\nabla_{\theta}R_n\left(\theta\right)$ is also a sum of functions one per data point.